UNIVERSITÉ DE NEUCHÂTEL

FACULTÉ DES SCIENCES ÉCONOMIQUES

Formalized Modeling of Qualitative Case Studies

PhD Thesis submitted to the Faculty of Economics and Business

Information Management Institute

University of Neuchâtel

For the degree of PhD in Computer Science

by

Dong Han

Accepted by the dissertation committee:

Prof. Kilian Stoffel, University of Neuchâtel, thesis director

Prof. Gerald Reiner, University of Neuchâtel, president of the comittee

Dr. Paul Cotofrei, Universitity of Neuchâtel

Prof. Yves Pigneur, University of Lausanne

Prof. Guangxing Tan, Guangxi University of Science and Technology, China

Defended on 19 December 2013

Faculté des Sciences Economiques
Avenue du 1er-Mars 26
CH-2000 Neuchâtel
www.unine.ch/seco

LETTRES ET SCIENCES HUMAINES
SCIENCES
DROIT
SCIENCES ECONOMIQUES
THEOLOGIE

UNIVERSITÉ DE NEUCHÂTEL

IMPRIMATUR POUR LA THÈSE

Formalized Modeling of Qualitative Case Studies

Dong HAN

UNIVERSITÉ DE NEUCHÂTEL

FACULTÉ DES SCIENCES ÉCONOMIQUES

La Faculté des sciences économiques,
sur le rapport des membres du jury

Prof. Kilian Stoffel (directeur de thèse, Université de Neuchâtel)
Prof. Gerald Reiner (président du jury, Université de Neuchâtel)
Dr. Paul Cotofrei (Université de Neuchâtel)
Prof. Yves Pigneur (Université de Lausanne)
Prof. Guangxing Tan (Guangxi University of Science & Technology, China)

Autorise l'impression de la présente thèse.

Neuchâtel, le 27 février 2014

Le doyen

Jean-Marie Grether

■ Téléphone : +41 32 718 15 00 ■ Fax : +41 32 718 15 01 ■ E-mail : secretariat.seco@unine.ch ■ www.unine.ch

Résumé

Cette thèse vise à résoudre les problèmes de modélisation et de traitement émergeant dans les études de cas qualitatives. Fondée sur la grounded theory, une méthode complète est proposée, initialement élaborée sous la forme d'une solution de workflow. Une combinaison d'ontologies est proposée servant à la représentation des connaissances, l'intégration et l'extraction. Sur la base de cette connaissance, l'analyse du sujet est menée afin de découvrir les informations latentes sur les documents originaux afin de représenter les thèmes implicites et mettre en place des structures hiérarchiques complexes de ces sujets. Avec les structures construites, les inférences basées sur l'ontologie est effectuée pour produire de nouveaux faits afin d'aider les tâches du domaine. Cette méthode est applicable dans la plage de multilinguisme y compris les langues non alphabétiques comme le chinois. Des expériences testées sur l'outil mis en place montrent que la méthode proposée donne des résultats satisfaisants par rapport aux méthodes existantes. La thèse générale apporte une solution nouvelle et complète pour étudier et analyser profondément des textes d'étude de cas avec une méthode qualitative. Ces résultats apportent une série d'avantages pour plusieurs domaines qui partagent la même essence du point de vue du traitement des données.

Summary

This thesis aims to solve the problems of data modelization and processing emerging in qualitative case studies. Established on grounded theory, a comprehensive method is proposed, initially elaborated in the form of a workflow solution. A suit of ontologies are proposed serving for knowledge representation, integration, and extraction. Based on this knowledge, topic analysis is conducted to discover the latent information out of the original documents in order to depict the implicit themes and set up sophisticated hierarchical structures of these topics. With the built structures, ontology inference is carried out to produce new facts to assist domain tasks. This method is applicable within the range of multiple lingualism including non-alphabetical languages such as Chinese. Experiments tested on the implemented tool demonstrate that the proposed method offers satisfactory results compared with existing methods. The overall thesis provides a novel and complete solution to profoundly study and analyze case study text with a qualitative method. It brings a series of benefits to a couple of domains which share similar essence from the point of view of data processing.

Mots clés études de cas qualitative, grounded theory, ontologie, entreposage de données, extraction-transformation-chargement, Latent Dirichlet allocation, apprentissage des ontologies, inférence des ontologies, la langue chinoise

Keywords qualitative case studies, grounded theory, ontology, data warehousing, extract-transformation-loading, Latent Dirichlet allocation, ontology learning, ontology inference, Chinese language

Acknowledgments

I would first of all like to thank my PhD supervisor - Prof. Kilian Stoffel. Under his supervision, I have learned a lot of precious knowledge for how to understand the subjects of computer science and information system, how to conduct research, and how to broaden my comprehension to a series of research topics from different points of view. My vision towards the research domains has been deepened to a more in-depth level during my entire PhD period by his direction. This has been playing a significant role in my research work and is going to be a valuable treasure for my whole future career. I have also learned a lot of instructive experience as the assistant of the courses he delivers. The way how he organizes the structures of knowledge, connects with multiple disciplines, and illustrates complex questions to intuitive ideas, gives me great inspiration. In the past years, whenever I need some help, for research, teaching, and daily life, Prof. Stoffel always encourages me with great motivation and concrete methods. I can always find very helpful suggestions from the discussions with him. This greatly helps me to cope with many difficulties and improve myself by the day-to-day progress.

I would like to express my gratitude to the jury members of my thesis and to thank Prof. Gerald Reiner and his research team for our collaboration during the past years. Meanwhile, many thanks belong to my colleagues. Since the very beginning of my PhD, I have received a lot of help from Dr. Paul Cotofrei. He gives me many useful advices for academic writing, research experiments, course delivery, and daily life. Dr. Abdelkader Belkoniene provides lot of detailed help for my research and daily life, which I sincerely appreciate. I would also like to express my thanks to Ms. Eugenia Cotofrei for her very considerate help in many aspects. This makes my daily work conducted in a more smooth way. Iulian Ciorascu, Eric Simon, and Christophe Kunzi give me frequently helpful advises during the regular collaboration and communication, which offer me great encouragement while coping with the questions I met. In addition, Hugo Marcelo, Tudor Calistru, and Fabrizio Alerbetti provides me many useful ideas for the research and daily life.

I am very grateful to my family. The constant support from them gives me great encouragement during my PhD period. I am deeply aware of their love to me. I am proud to be a member of this family. Besides, the three years' accommodation experience in *La maison de champreveyres* with so many friends is unforgettable to me. With them, I have profoundly learned European culture, languages, and the ways to reflect on many detailed questions. This implicitly gives me advantageous inspiration for my daily work and provides the opportunities to broaden my views in the international family.

Contents

List of Figures

List of Tables

1 Introduction

1.1 Overview

A momentous achievement of the modern society is its widespread involvement of technologies. Established on the shoulders of the prior works including mathematical theories and telecommunication infrastructures, information technology is probably one of the most prominent innovations over the past decades, unveiled through its pervasively used products from industrial domains, the research fruits of academic institutions, as well as the way how it has been changing people's life in many tiny aspects. From the super computing centres of space technology, the workstations for industrial usage, to the daily applications of PCs or even the interesting utilities on cellphones, we are profiting constantly of the contribution of information technologies. As our understanding to this technology is getting deeper, the demands for practical purposes are getting elevated correspondingly. The original idea to accelerate the capabilities of computing scientific questions is extended to a great many branches of advanced interest to profoundly comprehend and optimize the day-to-day performance for business and social purposes. It has a decisive role in data management, information sharing, and office automation for almost all the firms, no matter which domains they are engaged in and how large they are. As long as efficiency and optimal solutions are pursued, information technology is contributive regardless of the specific scenarios.

As an interdisciplinary field, the newly born "*information domain*" collects a large number of subjects from theoretical science, engineering, and even social sciences. Among all the subjects, *data* was from the very beginning a key element of this domain since all the core functions finally refer to data. The renowned equation "*Algorithms + Data Structures = Programs*" [1] indicates that the treatment of data in all computer programs and algorithms, whatever purpose they serve for, is essentially an art of data processing

to seek an optimal solution. At some moment, we used to believe that, thanks to the new-generation computers and their super computing capabilities, most problems of data processing have been solved or, in other words, the existing strategies are already satisfactory. However, as new problems emerge from time to time, the challenges of data processing are still increasing almost everywhere:

1. *Company A used to keep all the data in Excel spreadsheets. In order to manage the cross-referenced queries, they upgraded the system into a database. However, when new data arrives which does not follow the format of this database, plenty of time has to be spent to manually transform and load this data into the database.*

2. *Company B always finds that the exchange is not smooth between their internal organization and the experts from other domains due to inconsistency and misconceptions. This phenomenon occurs repeatedly and the concrete divergences vary from case to case. They expect information techniques to solve this problem for seamless integration from all the parties.*

3. *Company C is inclined to document all of their daily activities, but finally finds that the static data is not effectively usable for predicting the future even though a lot of efforts have been put into this type of data. Everything has ultimately to be done by their staff.*

The examples presented above are only a few of the typical phenomenons existing in industry today. They are happening quite frequently to companies regardless of their size and domains of activity. These challenges show up not because our computing capabilities are not sufficient in terms of efficiency or speed, or because the algorithms are not reliably efficient. On the contrary, they come from the manners how data is understood and treated. The same problems arise regularly in organizations, institutions, and even in our daily life. Some generic methodologies are hence called for to solve these problems to improve the existing technologies.

For the data *per se*, inspiring questions occur as computer systems are being utilized more often. People are using statistical software packages, such as *SPSS*, *STATA*, *R*, or even *Excel* to help them with their daily work. Nevertheless, if the original data is not easy to quantify, then these software programs are not able, or at least in an intuitive way, to achieve what is desired. A framework which helps to investigate this type of data is thereby in demand. In other words, if the data itself is found to be qualitative, i.e. hard to be represented with digits and formulas, an innovative methodology is hence needed to improve the activities for personal and enterprise purposes.

Besides that, the computational functionalities of machines are always appreciated. Thus we need, for the cases described above in which qualitative data is engaged, improved analytical capabilities. The investigation and semantic construction of the concepts captured from the original data is the principal focus. This aspect is very important since it helps human beings to understand the data at the first glance and profoundly depicts the latent patterns in the data. Some advanced algorithms, if well designed, will be constructive to set up novel knowledge beyond the analysis of human beings and, at the

same time, meaningful enough to assist domain experts for their consequential work. The "*learning*" aspect of these systems is thus accentuated based on which useful information is elicited.

Furthermore, the results learned from machined-based systems will have to be verified for their effectiveness. This analysis is supposed to be carry out on a theoretical basis as well from the human perception. Domain experts will certainly play an essential role as these methods ultimately serve for them. A generic framework is thus of great necessity to validate the results of the proposed methodology in this domain by conducting user experiments to simulate the practical scenarios and then to ameliorate this methodology. A validation process not only provides advantages to improve the methodology *per se*, but deepens the comprehension of the domains from a computer science standpoint.

1.2 Motivation and Problem Definition

Over the past years, we have been working on an interdisciplinary research project named "*Formal modeling of Qualitative Case Studies: An Application in Environmental Management*"[1]. The project aims to set up a conceptual framework for analyzing investments as well as strategies in regard to environmental issues, energy, and resource consumption. The proposed methodology is abstracted from real case studies in this ongoing project. Selected from the *Global Reporting Initiative (GRI)* database [2] and other database sources, the textual data contains annual reports of banks and other financial institutions regarding their regular performance and behaviour. Researchers participate in this work conducting case studies and contributing their insights based on their domain knowledge. Basic information techniques such as scripts for file processing have been developed to assist the experts. However, they are repeatedly confronted with a series of challenges. These arising challenges are summarized as the problem set mainly targeted in this thesis, based on which the overall research work is expanded to find solutions at the theoretical and applied levels:

- **Domain integration** Prior knowledge and newly created knowledge cannot be integrated effectively due to its heterogeneity. As plenty of information was previously developed in the domain of finance and environmental sciences, seamless integration of information from different sources becomes a necessity to improve the productivity of case studies. Otherwise, costly work of domain experts and IT developers has to be carried out and even information loss can occasionally be the consequence. New methods are required to incorporate the external knowledge in order to make it available for the projects in a smooth and systematic way.

- **Data maintenance** The data, including the annotations of the experts and the intermediate information gathered during the transformation of the data, is very often inadequately maintained. As the work load on large-scale data is fairly expensive in terms of time and cost, the demand for a suit of reasonable data structures is

[1]This work is supported by Swiss National Science Foundation (SNSF) - project No. CR21I2_132089/1.

very high. A serviceable data model is of great significance to the main framework as the data has to obey this model while being converted towards the final targets. Once this model is set up, all the processes are supposed to follow the same model to produce, convert, and maintain data, resulting in an increased consistency.

- **Knowledge acquisition** The capability of learning and reasoning about new knowledge is so far not well supported by qualitative research software. Products such as *ATLAS.ti* or *NVivo* offer interfaces for the users to annotate the documents, but they do not provide sophisticated utilities for subsequent treatment of the knowledge acquired during annotation. Once the information is retrieved from original data, experts have to do almost all the consequential operations on this information based on their own experience with ad hoc tools. Here machine based assistance is in obvious need to supplement the human inspection and to guarantee uniformity during the investigation. An inference engine, as an example, is potentially able to produce novel facts beyond the study of the experts. In addition, ontology learning techniques can help to discover new patterns in the data.

- **Globalization and multilingualism** Most existing tools for qualitative study are mainly oriented on alphabetical languages such as English and French. In our project, however, documents written in other languages such as Chinese show up regularly, leading to the need to extend the traditional text processing techniques. This demand-led problem comes essentially from the widespread of the applications because of the globalization. A solution addressing the multi-linguistic issues is therefore desirable to be integrated together with a thorough methodology.

- **Output producing** In quite a few cases, different types of reports have eventually to be produced. These reports vary greatly in their objectives, structures, involved data, as well as the methodologies of study. There exists a great potential to avoid repetitive work since many of these tasks share similar elements. An approach of data extraction and transformation can promote the smoothness of this process. A general framework will therefore be established to help automatically generate the reports in different forms with tools that can be personalized.

The issues are particularly important for enterprises associated with finance, banking, logistics, etc. The companies devoted to these domains have quite a few common characteristics. In most cases, they have to deal with large amount of data during their daily operations stemming from the customers, collaborators, as well as from internal reports related to their strategy execution and product lines. This data is to a large extend represented in form of text, or in other words, as qualitative data. Significant knowledge is included in this data as a latent benefit for these institutions if it can be effectively captured. Intensive processes evaluation in a variety of forms is needed. This creates a heavy work load on the participants, especially to guarantee the consistency of the results. Besides, from time to time a large number of domain experts have to coordinate their work towards specific goals. Even if the companies are aware of their specific demands, communication with IT specialists is not straightforward since domain knowledge is involved

and has to be integrated into the proposed methodologies. Furthermore, some existing data and IT infrastructures have already been deployed and it is necessary to consider the possibility of integration with these existing resources to reduce cost.

1.3 Contribution of the Thesis

Inspired by the above-stated question, the research target of this thesis is to provide a data modeling and processing methodology to extract, structure, and investigate the data which is presented in qualitative forms. The data involved is large, fast evolving with complex structures, and contains multidisciplinary facets. Mainstream software programs utilized for qualitative study are taken as the starting point and our approach will be compared with existing systems to improve on some of the key deficiencies of the state-of-the-art software, especially for the interoperability, data transformation, and learning aspects. Our system should succeed in defining a common procedure for similar projects, as a sequential process represented in form of a workflow, allowing to a great number of homogeneous cases to be serialized for automatic treatment.

The proposed methodology is based on *ontologies* as a tool for knowledge representation and for bridging the gap between systems and domain experts. In compliance with the *OWL* standard, a paradigm suit of ontological elements is formulated as the theoretical prototype, before being put into practice. This characterization at an algebraic level promotes the comprehension of the elementary constituents regularly used in OWL. Following the defined paradigm, ontological data structures are proposed as the guideline for annotation and system processing. These structures are to be extended for the purpose of serving a wide variety of projects at the level of data storage. The proposed methodology employs existing ontology-utilities to realize the process of *extraction-transformation-loading (ETL)*. This process allows to automatize the conversion of the annotated data, even in intricate cases, with constant algorithmic complexity. This consistent way of extracting the data allows for the establishment of one single data warehouse. With this idea, data from different sources converges at a relational model, considered as a virtually standardized layer to supply intermediary data for complex queries and semantic learning.

The data archived in the data warehouse, normalized as linked entities, contains rich information to be further investigated. One of the most important tasks is to capture the thematic topics latently distributed in the data with valuable meaningfulness to the domain evolution. A *Latent Dirichlet Allocation (LDA)* based approach is proposed which successfully yields semantic groups of topics from the extracted data. These topics reveal the insightful semantics of the original documentation. A newly-designed ontology learning technique is applied to construct hierarchies upon the acquired topics. The resulting hierarchies unveil the semantic structures of these topics instead of maintaining them only in a flat, unstructured way. It offers the possibility to extensively develop a series of practical applications with a hierarchical representation, such as recommendations and strategy validation. The obtained hierarchies will finally be archived as new ontologies to be incorporated with the prior domain knowledge. The integrated ontologies are able to

produce new facts with their embodied concepts and relationships via an inference process. The facts, as the output of the process, give insights to domain experts to iteratively enhance their analytical work. Moreover, many concrete problems are considered, for example how to handle Chinese documents. A mechanism is set up to retrieve knowledge from text written in Chinese as an indispensable part of our entire methodology.

As the core element of this thesis, *data* plays a significant role for a thorough methodology. The problem of modelization of the qualitative data is approached, and the corresponding functionalities are provided to make the data more digested for the users. An application tool has meanwhile been developed as the test bed and prototype of the theoretical design. It has a couple of advantages compared to its counterparts such as ATLAS.ti and NVivo. Experiments and user tests have been conducted to validate the proposed methodology. The results indicate that this methodology as well as the implemented system achieves very positive results in terms of efficiency and algorithmic effectiveness. The proposed approach eventually acts as a guideline for how data is modelized and analyzed for a series of similar domains.

1.4 Structure of the Thesis

Enlightened by the previous statements, a comprehensive methodology will be elucidated in detail as the contribution to prior theories and applications. Following the structure depicted in figure 1.1, the remainder of the content of this thesis is organized as follows:

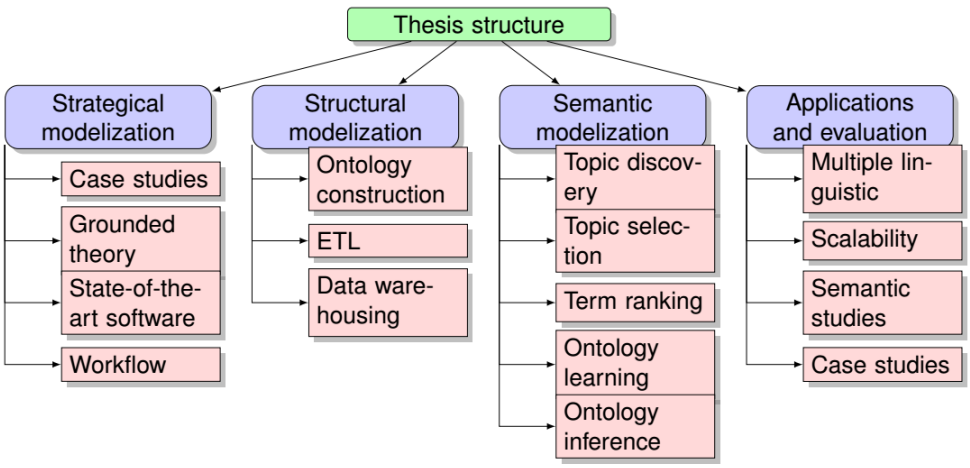

Figure 1.1: Thesis Structure

The "**Strategical modelization**" section, comprising chapter 2, reveals the background scenarios and outline of this thesis. It discusses the case study approach and presents an approach named grounded theory for qualitative research. State-of-the-art

software programs are studied as the starting point of our approach. A workflow is developed to depict the sequential elements of the proposed methodology.

The "**Structural Knowledge modelization**" section, including chapter 3, presents the ontology archetype and modelling processes for qualitative data. It establishes a set of ontologies that serve as the standardized and formalized schema. An ETL process is employed to transform the involved information into a suit of normalized data warehousing models, serving for cross-referenced queries.

The "**Semantic modelization**" section, made up of chapter 4, chapter 5, and chapter 6, describes the process of semantic learning which turns the modelized data into novel knowledge. It proposes a series of approaches to acquire the latent topics from the data stored in the established models and then builds up the hierarchical relations of these topics. Inferences are additionally carried out to produce derived facts as the indicative results.

The "**Applications and evaluation**" section, consisting of chapter 7 and chapter 8, manifests some practical issues and the evaluation of the proposed methodology. Globalization factors are considered, mainly for processing documentation in the Chinese language. A prototypical tool is implemented upon which experiments are conducted to assess the performance of the implementation and the effectiveness of the learning results from the presented approach.

2

Strategical Modelization
- From Grounded Theory to Workflow

This thesis is motivated by the issues related to the business domains and the possibility of using qualitative case studies where massive amounts of data are involved. The basic idea is to use *Grounded theory* as the guideline of our proposed methodology. Furthermore, a workflow solution is presented in this chapter as the sequential outline built upon the analysis of the existing approaches.

2.1 Case Studies

Information techniques, with their unique features, are able to play a vital role in practical domains via innovative ideas [3]. At present, a key issue for enterprises, seen from the IT point of view , is the *big data* problem as analyzed in appendix A. The examination of the collected data often falls into the category of the big data scenario particularly because it is essentially an interdisciplinary subject with various branches involved [4]. For example, the assessment framework of a company is usually associated with its production, cycling of the materials, sales, energy consumption, and even human resources [5]. Each sector has plenty of textual reports in different forms and these reports are usually updated regularly. They can be seen as *qualitative data* contained in the original documents. Valuable information is dispersedly distributed in these documents [6]. As a consequence, an approach to investigate the textual data established upon data modelling, and processing techniques producing structures with learning capabilities, is necessary for the research and the applications in the practical domains.

In economics and management, there are already a number of existing methodologies obtaining satisfactory results [7]. Most of these methodologies follow a case study approach. Case studies have been widely used in a variety of fields, especially the ones

lacking formal ways to evaluate the maturity of enterprise perception [8].

As many questions which are not straightforward to answer exist in the sustainability domain [9], the development of a solid methodology to assess and manage these factors is thus challenging [10]. In other words, a system is in demand to assist the case study processes for data investigating in many similar domains, especially emphasizing qualitative elements. There already exist some systems for this purpose, for example, *Quality function deployment (QFD)* [11]. It is apparent that more and more functions will be added incrementally on top of these systems as the project evolves. Working in this method brings the advantage that for the projects controlled within a certain range the results can be delivered conveniently [12]. If the projects become larger, however, bottlenecks will emerge [12].

In different cases, many questions posed by domain experts stem from the theories associated with their expertise. The way in which these theories are organized is normally different from the ones typically used in information domain [13]. Plenty of time and efforts have to be invested to convert these questions into a format implementable in computer systems [14]. In addition, validation approaches are needed to spot this kind of requirements. Furthermore, re-usability should be a significant issue to consider during the implementation of these systems in order to avoid repetitive development [15]. The techniques from the information domain are supposed to regularize external requirements which do not have clear schemes when they are originally initiated.

2.2 Grounded Theory

Based on the statement above, it is obvious that some methodology is needed to be the guideline of the conducted research. In this thesis, the *grounded theory* approach (GT) [16] will be a reasonable solution. We are particularly focusing on the construction of a theory starting from data [17].

2.2.1 Analytical Steps

The elementary steps of grounded theory are discussed as follows: At the beginning, the original data is labelled with a number of *codes* [18]. These codes are keywords to describe, abstract, and generalize the useful information or key features of the primary data. They are considered as the basis of potential studies in the next steps. The development of codes, as a task incorporated into the entire processes, is considered by domain experts as a significant step [19]. Basically two types of coding are employed in our methodology - open coding and limited coding. Open coding [20] gives to users full freedom to add any code considered reasonable for the content. Limited coding [21], as the name implies, restricts the codes to a list of words that the users have to follow. Usually the choice between the two types depends on the level of the users' domain knowledge. For limited coding, it is easier to conduct data aggregation and interoperation, whilst open coding offers more possibilities to discover new ideas. A hybrid way is to propose a list of codes at the beginning, with the possibility for the users to freely add their own codes

if necessary. For advanced applications, the codes are not necessarily organized in a flat list. Instead, some codes contain more general keywords than others, making the structures hierarchical [22]. A tree-based or even graph-based coding structure is employed in many scenarios. Furthermore, values such as weights and ratings can be attributed to the codes. These weights are visible to the users during the annotation phase and, therefore, they can make use of the weights during the coding activities.

The next layer of grounded theory is named *concepts*. The coding phase offers to users the possibility to work directly with the original data whilst the concepts are built up upon these codes. Usually these concepts are abstract terms generalized from the codes. For example, the codes "*long term*", "*resource*", and "*energy*" can refer to the concept "*sustainability*". A concept can either be an ordinary code or a completely new word that never shows up in the original text. The step of conceptualization is indispensable especially for grounded theory [23], since in this type of studies the main themes are usually implicit. It is thus difficult to discover the concepts or topics in the original data. In this thesis an approach is proposed to handle this problem. However, the method remains to be verified in order to unveil its advantages. Once the concepts are formed, they are grouped into different *categories*. These categories depict the different conceptual clusters at the same semantic level.

Existing research shows that grounded theory is an effective way to produce theories for innovative organizational processes and creativity requirements [24]. The organizational processes are expected to be led by the needs of documents and should lead to cost optimization and strategy promotion [25]. In this context, the ultimate goal of this discipline can simply be summarized as trying to capture the implicit rules in the documents and then provide actionable knowledge to the enterprises. There are a number of ways to establish practicable knowledge such as data mining and ontology inference. In this thesis our own methodology is set up on top of the existing approaches and is then applied in order to infer *theories*. These theories should help the domain experts for the purpose of information investigation, hypothesis creation, and conclusion establishment. The forms of the acquired theories vary from case to case, but it is advantageous to develop a generic approach to support theory generation.

2.2.2 Big Data Issues

Backtracking to the characteristics of big data, it is not difficult to conclude that grounded theory is an appropriate approach to deal with the problems arising. For this particular issue, there always exist large amounts of data in the first place without substantial attainments. Even though we are aware that valuable information is encapsulated in this data, it is demanding to discover the hidden principles at the first glance. This conforms with the grounded theory, which starts the investigation from the collection of data instead of proposing hypothesis. With the coding schemes, a more profound comprehension of the original data is achieved. Through these codes, the problem of the large *volume* of data is alleviated since basically our ensuing work will only depend on the codes instead of the original data. And the number of codes is limited and therefore more manageable.

For the problem of the complex-structure of the data, the coding step is also an en-

couraging solution because it essentially sets up to a virtual layer above the data, making it transparent for the subsequent studies. Even if the formats of the data are different, the codes employed to express their content are the same at the semantic level, which facilitates the overall development. As the data is categorized, when new data emerges, it will be attributed to the existing categories, making the integration more efficient and standard. Based on these categories, the methodology already set up can be applied directly to newly arriving data. This is a promising solution for the *velocity* problem. Furthermore, with the conceptualization step, the *value* of the data is identified easily and precisely.

Based on these arguments, it can be concluded that grounded theory is a capable approach, especially for big data arising in case studies, and it provides a promising solution to the requirements of case studies. In other words, it is a methodological bridge to connect the three major issues focused on in this chapter as shown in figure 2.1 in which the intersection of the three circles represents the grounded theory.

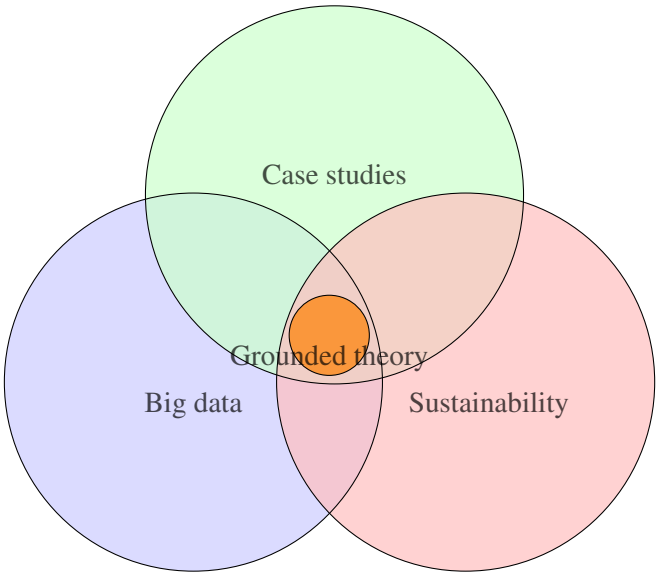

Figure 2.1: Grounded Theory as the Intersection

2.3 Qualitative Software as the State of the Art

Later in this thesis, a methodological implementation will be developed firmly based on the grounded theory approach. Derived from the idea of grounded theory, a number of commercial software programs have been released for qualitative applications. *ATLAS.ti* and *NVivo* are the most prominent ones. They emphasize on the analysis of elementary concepts captured from the text and their implicit relations. The main objective of these products is to annotate original documents by experts. With the analytical framework supported by these software programs, domain specialists are able to develop further

insights and to reveal novel characterization of their findings from the cases.

2.3.1 System Features

ATLAS.ti has been designed on the basis of a group of concepts [26]: Beyond all, *primary documents* are the original data source. Textual, graphical and multimedia data are supported as primary documents. A marked section or piece of these documents, namely a *quotation*, represents the significant or valuable parts of primary documents. These quotations can be exploited by adding *codes* with their names invented by the users or with words existing in the text. Another concept, *memo*, is used to comment the concepts with annotation. In ATLAS.ti, all these elementary concepts are defined as nodes and the system provides the mechanism for creating a relational model among these nodes. The concept of a *family* is used to cluster or classify nodes into groups to indicate their similarities. *Relation*, a self-defined attribute, aims to link nodes to explicate their mutual connection. Furthermore, a *network* is designed as a diagram to illustrate the global structures of nodes and their relations. ATLAS.ti reflects the implementation of grounded theory from an industrial point of view.

As the counterpart of ATLAS.ti, another company, *QSR International* [27], offers products including *NVivo 8* and *XSight*, two software programs that help the users to model their source data, assign it to specific nodes, and analyze the data from the perspectives of the relations or links among concepts. NVivo 8 owns similar functionalities as ATLAS.ti but with richer technical modules. In addition, XSight permits the perception of data in the form of an analytical framework leading to an easier way for the researchers to structure the data obtained from the surveys and investigations. As their user interface adopts a more widely-accepted style than ATLAS.ti, comparable to *MS Outlook* [28], the operations of NVivo 8 and XSight are more intuitive for the end users to get accustomed to, especially during the initial phases.

2.3.2 Potential Improvement

In order to provide a more satisfactory solution, it is necessary to find out the exact functionalities that our proposed methodology can add compared with the state-of-the-art software. Despite the many interesting functionalities provided by the classical programs, there are a number of issues that should be addressed:

- **Platforms and languages** The existing qualitative software, for example ATLAS.ti, works only on the *Windows platform*. This causes an obvious problem as a certain proportion of users are making use of other operating systems including *Mac OS* and *Linux*. It is not realistic to require them to move to a new system. In this case, a cross-platform solution would clearly be more convenient for the users. Furthermore multiple languages would have to be considered because globalization has a high impact on practical domains. With our proposed solution, users with different technical and language background are able to cooperate simultaneously and coordinate the tasks collectively and smoothly.

- **Graphical interface** Almost all software tools and on-line services offer graphical user-friendly interfaces. This is a critical factor to gain the confidence from clients and obtain their preference. For document navigation and annotation, most users are accustomed to *Adobe Acrobat* [29] for navigating PDF files. ATLAS.ti, however, has its own interface, not straightforward for the users to transit to from their regular habit. This might be challenging for people with different background and specialisation domains. A better solution is thus to provide the users with manageable interactions from the graphical point of view. Ideally, this solution is supposed to inherit from standard document readers but offer a more interactive interface to operate on the files.

- **Systematic analysis** One of the chief tasks of these systems is to aid domain experts to produce insightful output, derived from the approaches such as data modelling, data mining, etc. A system is supposed to supply perceptive functions to facilitate the users' operations and, in addition, to give them more intelligent feedback based on its computational capabilities. This feedback is the reason why an end user relies on systems instead of working with paper and pencils. However, applications such as ATLAS.ti and NVivo do not deliver sufficient services for analytical examination, which means that a large amount of work has ultimately to be completed by the experts without aid of the tools. In this thesis, a system is to be designed to support the users in all aspects of investigating the original data. These functions cover most of the phases a user works on throughout a project from the initialization to the theories that are produced. In this sense, the *learning* abilities of ontologies will be employed to establish novel knowledge from collected and annotated information.

- **Interoperability and output** Another significant concern is related to the interoperability with other mainstream software. Once all the data from the primary information to the users input has been archived, it is required to be able to export this information. Moreover, the results from the basic systems are supposed to be delivered to other systems; otherwise, the data is restricted to the scope of an isolated program without sufficient communication and interchange. This is actually one of the drawbacks of the existing systems even though some export functions are offered. The output of ATLAS.ti and NVivo is very rigid and does not offer the possibility for personalization. Furthermore, this output does not reflect the complete information of a project, which can lead to information loss. Consequently, in our proposed system the design has to be integratable with other widely-accepted programs via standard data exchange formats.

- **Automatic data transformation** The data involved in qualitative studies is not always static. Therefore a rigid integration process for the raw data is a major drawback. ATLAS.ti and its counterparts have their obvious deficiencies in this sense. Our proposed methodology will thus take this aspect into account by designing a suit of automatic ways for data transformation and integration. It is designed to

be capable of converting data from different sources into its derived formats. Consequently the combination of heterogeneous data is facilitated.

- **Validation** When the users are exploring the data, a framework should be able to validate their behaviour, the produced knowledge, and to suggest strategies, so that errors and misconceptions can be avoided. For example, when there are contradictions or conflicts between different segments of the data, the system should be able to notify the users about the inconsistencies arising. According to the feedback collected from domain experts, this functionality is in great necessity due to its practical impact. This function has a particularly vital role in the phase of data preparation, because this phase may take more than half of the time of the entire projects [30]. However, considering the applications discussed so far there is no support for this type of validation. Therefore a framework will be formulated to show the feasibility of this type of validation.

2.4 Workflow Design

Considering the potential improvement mentioned above, a complete solution is in demand and will be addressed by the methodology proposed in this thesis. The previous remarks make it evident that a well defined workflow is of great importance. The concrete workflow will be described in the following.

Before elaborating the proposed workflow in detail, it is necessary to briefly introduce the business scenario. The description of the business scenario accelerates the understanding of the proposed methodology and sets up a generic model for similar projects and cases. In this project domain experts work in parallel in several groups to analyse the performance of financial firms regarding their specific roles. They have different educational backgrounds, language skills, operating systems, and user habits. All the original data the experts are exploring is in form of textual documents provided by the companies under consideration. The domain experts are working with the proposed system in an interactive way. The workflow comprises several key steps: first, the users analyse the original data - the PDF documents. The interaction between the users and the data is supported as analytical annotations (the data elements are formalized in appendix B). All the annotation actions are stored in ontological files. Next a process of *extraction-transformation-loading (ETL)* is conducted to retrieve useful information from the produced ontologies into data warehousing models. These models guarantee for the uniformity of the intermediary data. Then topic discovery will be performed using a LDA-based approach. Once the topics are acquired, ontological learning techniques are put into practice to establish the hierarchical relations between these topics. With these newly-built hierarchies, a variety of applications are available to produce feedback to the users. As soon as the users receive the feedback, they will take it into consideration and enhance their prior analytical insights to produce new contribution to the system. The designed workflow is presented in figure 2.2.

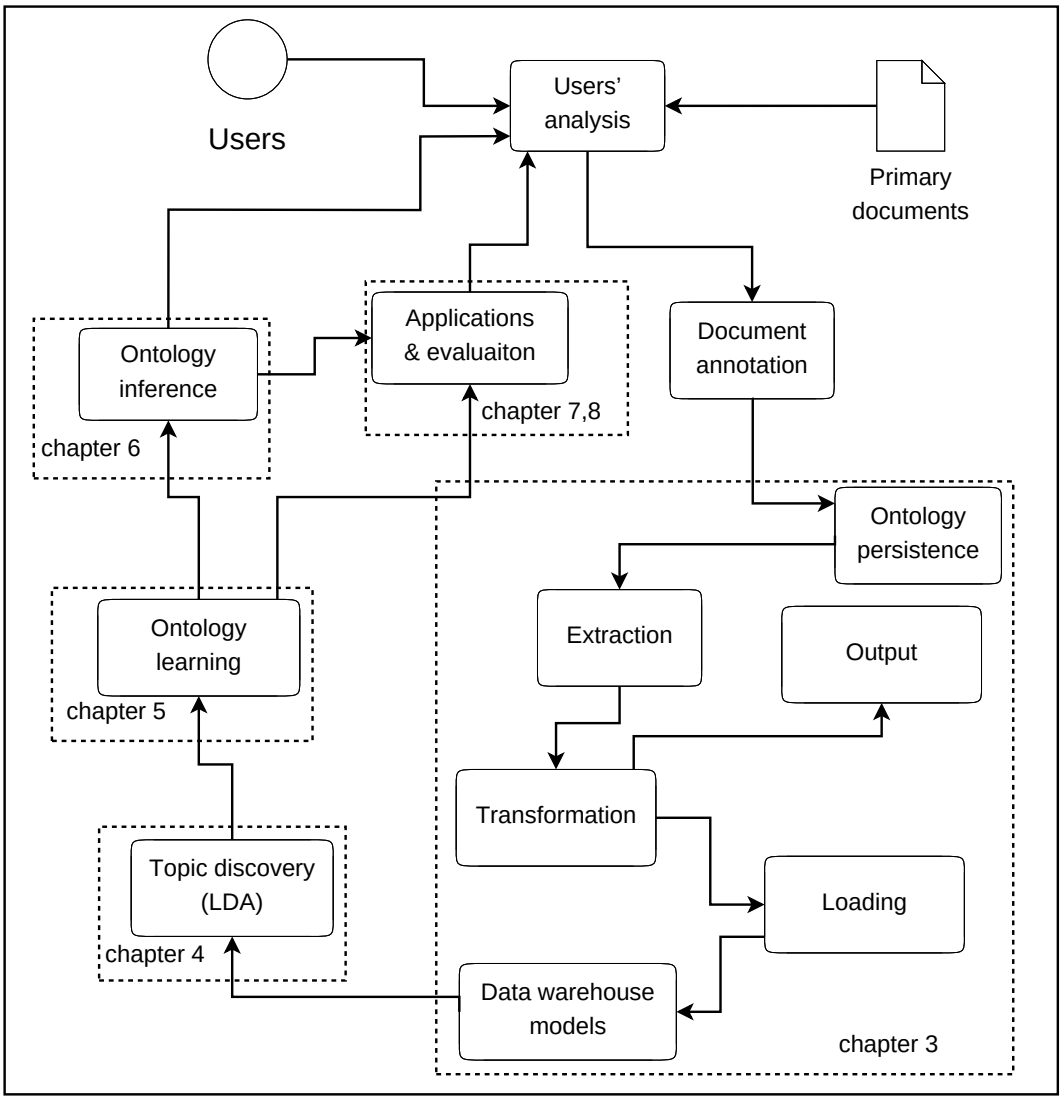

Figure 2.2: Workflow of the Business Scenario

(Source: [31] (improved))

2.5 Conclusion

Interdisciplinary subjects require efficient data processing methodologies as a fundamental element. Grounded theory is an effective solution for this basic challenge. The detailed properties of the state-of-the-art software, implementing partially this approach and widely used in industrial applications were characterized in this chapter. In order to improve the existing tools, a workflow solution was proposed based on the formalization of the original data and annotations. The outline of this thesis will also follow the pro-

posed workflow. In the next chapter, a tool for knowledge representation - *ontology* - is going to be presented. Then its capabilities for knowledge representation and the process of extraction will be illustrated.

3

Structural Knowledge Modelization[0]

- Ontologies and Data Warehousing

In order to transfer grounded theory into machine based methodologies, a suit of ontologies are designed to guide the further development.

3.1 Upper Level Ontologies

Before designing the concrete ontologies, a suit of upper level ontologies are needed to guide the overall projects. As in a project, different data, methods, and roles of participants involved with each other, a reasonable framework is thus significant.

When the upper level ontology is being outlined, the particular features of case studies are underlined as they are the focus of our research. In general, it is complicated to use a single set of ontologies to guide a large number of cases, because they vary in regard to data sets, methodological structures, and other details. For this reason, the major focus of our task is simplified to target only two of these aspects - data and methods. Input data has a vital role in case studies. For qualitative data, it contains rich information and the principles which, if well discovered and utilized, reflect directly the problems the case study is focusing on. As a result, the upper level ontology is supposed to include data as a primitive element. Regarding the methods, as grounded theory has been specified to be the guideline of our methodology, the upper layer ontology will take this theory, especially the coding process, as the backbone of all methods.

Figure 3.1 elucidates the upper level ontology which aims to describe the general framework of this thesis. Each derived project will then be implemented as an inherit-

[0]The work of Section 3.1 and 3.2 has been briefly described in [32] and the basic work of 3.3 and 3.4 has been presented in [31]. More advanced details are revealed in this chapter.

ing set of lower level ontologies following the constraints imposed from the upper level ontology.

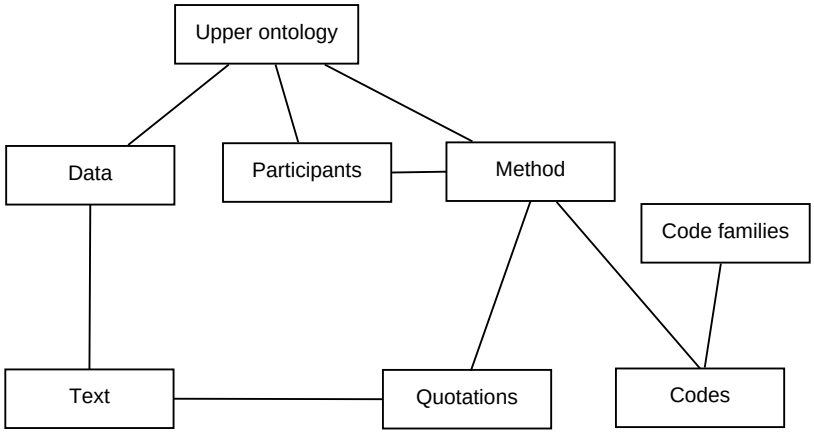

Figure 3.1: Upper Layer Ontology

3.2 Lower Level Ontologies

Following the upper layer ontology illustrated above, a suit of ontologies will be defined and implemented to support the proposed ontological methodology.

3.2.1 Project Ontologies

A first set of ontologies aims to implement the process of grounded theory with its structural elements. It formulates how projects are conducted with a series of constraints. These ontologies are called *project ontologies*. They also enable reasoning engines to build up novel knowledge based on the ontological data. The main purpose of these ontologies is to ensure that the whole project is in compliance with the upper level ontology and thus the coherence is guaranteed among all the participants.

With the project ontologies in place, the scope of a project is effortlessly disclosed as a *template*. It outlines the path which a project will follow in the next stages, independently of the participants, data, and methods used. This template is automatically attributed to all the involved data. Also, the newly-discovered knowledge will be derived in accordance with this template. The advantage of applying project ontologies is that, conducted in this way, only general rules are necessary when applying them for generating new facts. Hence, the following ontological elements are defined:

- A set of *Classes (CL)* is defined to outline the project requirement in respect to the data and annotation elements. While the project is expanded, the subsequent ontologies will have to be in compliance with these elements.

$$CL_{document} = \{_{owl} C_{Class}, _{rdf}\ V_{id} = "document"\} \tag{3.1}$$

$$CL_{quotation} = \{_{owl} C_{Class}, _{rdf}\ V_{id} = "quotation"\} \tag{3.2}$$
$$CL_{code} = \{_{owl} C_{Class}, _{rdf}\ V_{id} = "code"\} \tag{3.3}$$
$$CL_{rating} = \{_{owl} C_{Class}, _{rdf}\ V_{id} = "rating"\} \tag{3.4}$$

- *Data type properties (DP)* are established upon the classes. The objective of these properties is essentially to define some primary values to describe the properties of the data and annotation elements. For example, a *String* is used to record the users' names in order to identify the users who extract quotations from the original text.

$$DP_{quser} = \{_{rdfs} C_{DatatypeProperty}, _{rdf}\ V_{id} = "quser", page, String\} \tag{3.5}$$
$$DP_{qname} = \{_{rdfs} C_{DatatypeProperty}, _{rdf}\ V_{id} = "qname", page, String\} \tag{3.6}$$
$$DP_{cuser} = \{_{rdfs} C_{DatatypeProperty}, _{rdf}\ V_{id} = "cuser", code, String\} \tag{3.7}$$
$$DP_{cname} = \{_{rdfs} C_{DatatypeProperty}, _{rdf}\ V_{id} = "cname", code, String\} \tag{3.8}$$

- *Object type properties (OP)* are in addition used to connect two elements. The relations among different elements are supported by these properties. Taking the example of a quotation, a "*hascode*" property is set up to express the fact that this quotation has been assigned to several codes.

$$OP_{hasquotation} = \{_{rdfs} C_{ObjecttypeProperty}, _{rdf}\ V_{id} = "hasquotation", document, quotation\} \tag{3.9}$$
$$OP_{hascode}\{_{rdfs} C_{ObjecttypeProperty}, _{rdf}\ V_{id} = "hascode", quotation, code\} \tag{3.10}$$
$$OP_{hasrating}\{_{rdfs} C_{ObjecttypeProperty}, _{rdf}\ V_{id} = "hasrating", quotation, rating\} \tag{3.11}$$

3.2.2 Reference Ontologies

Sometimes, besides the previously defined project norms, domain experts may want to integrate some external knowledge into the projects as supplementary information. *Reference ontologies* are thus employed to represent the views from their subjective opinions as well as existing insights from specialized analysis. The incorporation of reference ontologies enhances the smoothness of the external data integration.

An example of reference ontologies is elucidated in figure 3.2. They will then be integrated into the entire framework by adding them to other ontological knowledge to enhance the learning and the inference processes.

A decisive question at this step is how to accommodate heterogeneous information. In a lot of domains there exist published knowledge sources. For example, *eXtensible Business Reporting Language (XBRL)* [33] is at present one of the most accepted standards to express financial information [34]. Since this language is preserved in the form of ontologies, it is thus manageable to conduct an integration into our system. Some other domain information, however, has to be transformed in order to be supported by our system. For instance, *Environmental Data Explorer* [35], an authoritative data source for environmental variables, has its own database format. In order to solve its integration problems, a couple of functional libraries following the *facade pattern* [36] were implemented in order to convert this valuable information and to integrate it into our system.

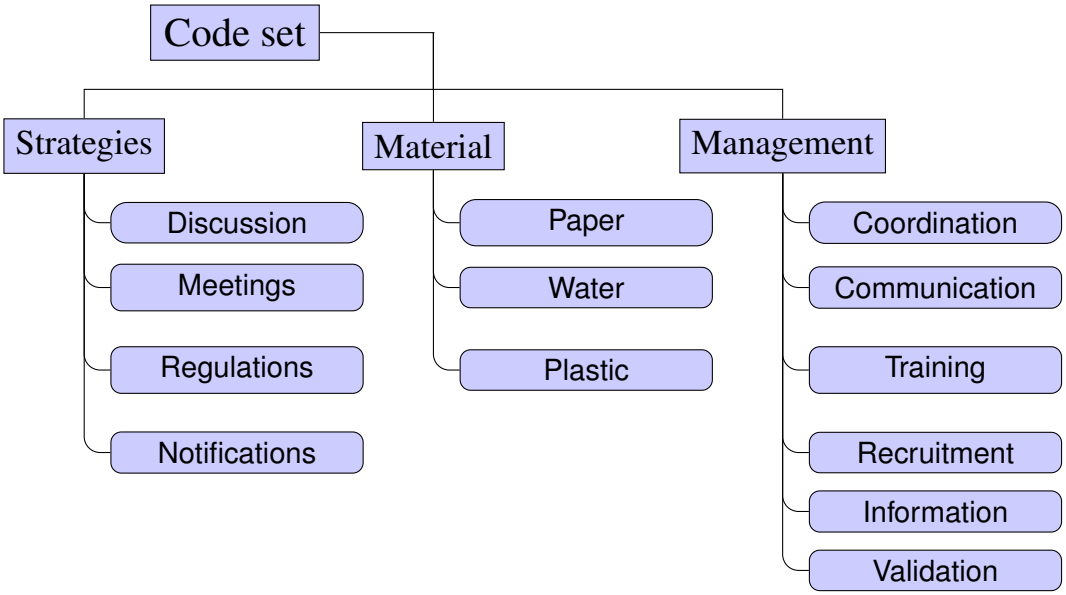

Figure 3.2: Reference Ontology

3.2.3 Coding Ontologies

Another set of ontologies, named coding ontologies, serve for recording the raw data as well as the users' coding information. While a user starts the coding operations, these ontologies are created automatically to register their interaction with the primary documents. Project ontologies and reference ontologies are inherited to complete the coding ontologies. As a result, all the constraints imposed by the previous two types of ontologies are respected while the concrete information is recorded. A simple fragment selected from an instance of coding ontologies is depicted in figure 3.3. Quotation and code are two classes and they are linked by the relation of *hascode* (an object type property). This example can be expressed with our established definitions as follows:

- A quotation instance was created with a sentence from the primary document extracted. The content of this sentence is: "*Last year, the paper consumption of this bank increased 20%*". A six-digit ID (308524) was randomly generated to identify this quotation.

$$\exists i_{308524} \in Individual, i_{308524}.V_{Class.id} = "CL_{Quotation}" \bigwedge i_{308524 \cdot rdf} V_{content} = "Lastyear..." \tag{3.12}$$

- A code instance was created named "*Recycle*".

$$\exists i_{recycle} \in Individual, i_{recycle}.V_{Class.id} = "CL_{Code}" \bigwedge i_{308524 \cdot rdf} V_{id} = "Recycle" \tag{3.13}$$

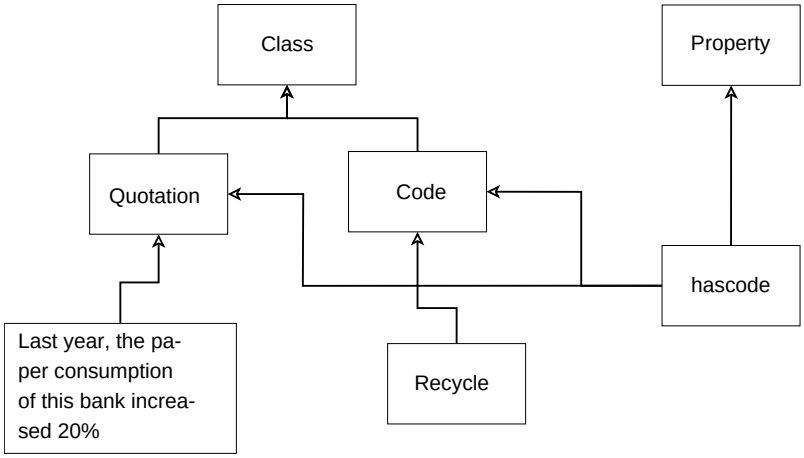

Figure 3.3: Code Ontologies

(Source: [31])

- Moreover, the code "*Recycle*" was assigned to the quotation i_{308524} as the experts believe that the theme of this quotation can be represented by that code. Based on equation 3.10, 3.12, and 3.13, the following equation can be constructed:

$$i_{308524}.children("OP_{hascode}").{}_{rdf}V_{resource} = "Recycle" \tag{3.14}$$

3.3 Extraction, Transformation, Loading

Extraction, Transformation, and Loading (ETL) defines a series of processes implicitly in order to build the framework of the data warehouses [37]. These processes include elementary steps assembled with a predefined template [38]. In many cases, the major objective of ETL is to establish an approach to automatize complex business scenarios. Taking into account the problems for complex-structured information management, it is very reasonable to apply a multi-faceted ETL plan for different steps involved [39]. The techniques of ETL are thus incorporated into our research as a pivotal method.

3.3.1 Extraction

As all the newly-produced data is archived in the form of ontologies, while being annotated, the data will accordingly be transformed into appropriate forms in order to build the data models. The main techniques applied to achieve this goal are described in this section. For extraction, hence, two issues are relatively vital - the data sources and the target formats of extraction. In our project, the former issue is relatively elementary by taking advantage of the formalization of the original data (see appendix B). Hence the major challenge is the latter one.

Table 3.1: Relational Table

Code	Company	Year	Number
Applying standards	BKA	2006	14
Community programs	BKA	2006	6
Continuous improvement	BKA	2006	16
Customer relationships	BKA	2006	13
Customers policies	BKA	2006	1
Developed Service	BKA	2006	23

The second issue is tricky in the sense that at this step the concrete representation of the data is not yet known. For different applications, the format of the data may vary. The best solution is to find some generic approach which is still implementable. Initially, it is necessary to define a group of dimensions to describe the data. It will then become apparent which properties to extract from the original data and what formats of structures to load into the data models. For each ontology file, two tuples, external and internal respectively, are set up by default [31]:

$$Q_{external} = \{group, user, project, company, year, file\} \tag{3.15}$$

$$Q_{internal} = \{file, quotation, coordinates, codes\} \tag{3.16}$$

The ETL process is supposed to be in compliance with these two tuples. Correspondingly, two kinds of operations are designed - global extraction and local transformation. *Global extraction* aims to elicit descriptive information of the submissions delivered by the experts without navigating the content of the files. The purpose of this step is to reorganize all the data in a structured way and give each ontology file a unique identifier. Once this step is completed, all the knowledge is stored in a uniform format, each file labelled with the attributes in tuples 3.15. This step is also called *external extraction* as it only deals with the external properties of the data. For the purpose of this process, a *Data extraction algorithm* has been developed as described in algorithm 1.

3.3.2 Transformation

After all the ontologies have been regularized, *local transformation* is conducted next. Local transformation aims at extracting internal information from the ontologies following the definition of tuples 3.16. The results of this step are divided into two categories with respect to their data structures:

1. ***Relational tables***. As shown in table 3.1, in a relational table each attribute represents one dimension, for example, *company*, *year*, *code*, *code frequency*. Relational

Algorithm 1 Data extraction algorithm ($F_e(D)$)

Require: File path D_o as the input, path D_f as the output
Ensure:

```
1:  temp ← D_o
2:  if temp ∈ File then
3:      fName ← temp.getName()
4:      fNames ← fName.split("_")      {retrieve each part of the file's absolute name}
5:      group ← fNames[0]      {assign values from the file path to different variables}
6:      user ← fNames[1]
7:      company ← fName[2]
8:      year ← fName[3]
9:      if fName.contain("owl")||fName.contain("owlx") then
10:         content ← temp.getOWLFile()      {get the content of the OWL file}
11:         fNewName ← concate(group, user, company, year, coding)
12:         sendOWL(D_f, fNewName, content)      {send the original content to the new
            path}
13:     end if
14:     if fName.contain("csv") then
15:         content ← temp.getCSVFile()
16:         fNewName ← concate(group, user, company, year, performance)
17:         sendCSV(D_f, fNewName, content)
18:     end if
19: end if
20: if temp ∈ Folder then
21:     j ← 0      {assign an initial value to j}
22:     while temp.hasNextChild do
23:         F_e(temp.children(j))      {invoke the F_e(D) function recursively}
24:         j++
25:     end while
26: end if
```

tables are a format that fits well the schema of relational databases [40]. It is thus straightforward to import these tables into a relational database management system (RDBMS) and to make them available for advanced queries.

2. ***Pivot tables***. A second type of tables, as shown in table 3.2, is called pivot tables. A pivot table is commonly used for storing resulting data in a form useful for graphical export and basic statistical operations, such as aggregations [41].

A critical decision to make at this step is to decide whether to use relational tables or pivot tables. As an intuitive principle, if a table is designed to be presented to the users as a tabular report with little interaction with other tables, a pivot table is preferred. Relational tables are more commonly used as an intermediary format for transformations and to be

load *join queries* into database systems. Since the data structures of these two types of tables are determined, transformation tools have been developed to convert them in a convenient way. For example, algorithm 2 describes a process to transform the data into a relational table.

Another task of the transformation step is to aggregate the primary information. For example, after all the submissions from a group of experts are collected, the average values of their submissions can be calculated. An *aggregation* function will handle this task after transforming the original information. In this case, programming languages or SQL commands (see listing 3.1) can be employed to obtain the results of these aggregations.

Listing 3.1: Data Aggregation in SQL

```
select code, company, expertName,avg(y2011), avg(y2012)
from tbl_code inner join tbl_expert
on tbl_code.expertId=tbl_expert.expertId
where userType="expert"
group by code,company;
```

3.3.3 Loading

The last step of the ETL process is loading the transformed data into a relational database or a data warehouse. This step is relative simple compared to the previous two steps. A typical example of a loading assignment is shown in listing 3.2. Its major concern is the interaction between data warehouse entities. In other words, the principles of databases and data warehouses have to be respected, such as primary keys, foreign keys, cardinalities, etc. When the loading is conducted, there are many technical questions worthwhile to answer. For example, a CSV file may have *comma* or *semi-colon* as its separator, but some DBMS takes only one type among them as its separator. For this purpose, the data is prepared at the loading phase by comparing it with the specific rules of the target systems.

Table 3.2: Pivot Table

Code	Company	2006	2007	2008
Applying standards	BKA	13	14	8
Community programs	BKA	7	6	7
Continuous improvement	BKA	10	16	6
Customer relationships	BKA	16	13	17
Customers policies	BKA	3	1	5
Developed Service	BKA	11	23	16

Algorithm 2 Data transformation algorithm

Require: File path D_f as the input, file f_{target} as the output
Ensure:

```
companyHashMap ← ∅
f_target ← code.csv
while D_f.hasNextChild do
    temp ← D_f.child()
    codeHashMap ← ∅
    if temp ∈ OWL then
        company ← temp.getName()
        while temp.hasNextChild do
            content ← temp.getFileContent()
            {count the number of codes in an ontology file}
            codeHashMap.addCount(content.selectNodes("//code[contains(@rdf :
            ID")])
            j++
        end while
        companyHashMap.add(company, codeHashMap)
    end if
end while
while companyHashMap.hasNextChild do
    company ← companyHashMap.getKey(k)
    codeHashMap ← companyHashMap.getValue(k)
    while codeHashMap.hasNextChild do
        code ← codeHashMap.getKey(k)
        {output companyHashMap into a new CSV file}
        fNewContent ← concate(code, company, codeHashMap.size())
        f_target.append(fNewContent)
    end while
    k++
end while
```

Listing 3.2: Data Loading in SQL

```
create table tbl_code
(id varchar(8), code varchar(255), groupp varchar(255), name varchar(255),
    company varchar(255), expertId varchar(8), y2011 decimal, y2012 decimal,
PRIMARY KEY (id),
FOREIGN KEY (expertId) REFERENCES tbl_expert(expertId));
.separator ,
.import .\\codecompany_year.csv tbl_code
```

3.4 Data Warehousing

One of the objectives of ETL is to build a data warehouse from the original data. In order to set up the models representing precisely and efficiently the project data, we will describe in this section the principles of the models produced as the results of the ETL steps. A data warehouse is designed to support data interpretation, the decisions of management, and information combination with different perspectives over the computational resources [42]. It emphasizes the organization of certain datasets, with features to incorporate temporal factors or to provide solutions for practical scenarios [43]. Leveraging on its explicit advantages, a data warehousing approach is exploited to build the data models for the sake of semantic studies.

3.4.1 Subject Area Model and Business Data Model

Based on the principles proposed by [44], a series of procedures are conducted as described in this section. After applying the steps described, a couple of subjects are abstracted from the business scenario for the purpose of the creation of a *Subject Area Model*. These subjects are presented in a generalized form and they should cover all the data with the corresponding relations.

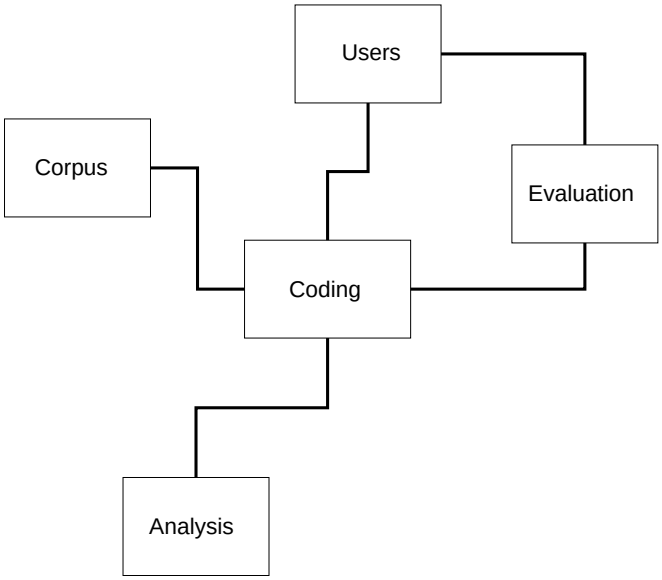

Figure 3.4: Subject Area Model

As shown in figure 3.4, the designed subjects include:

- **Corpus** [45] All the original documents are maintained as the corpus and they are partitioned into document clusters. For example, the reports from the same company for different years are assigned to one cluster. Each document is composed of

pages on which the elements are words and characters. These concepts are formalized in appendix B to clarify their exact content. The subject "*corpus*" contains all these concepts as general subjects for the original data. The corpus set for our project is listed as an illustration in table 3.3.

Table 3.3: Corpus

Company	Start year	End year
BANCO SANTANDER SA	2003	2011
CREDIT SUISSE	2004	2011
DEUTSCHE BANK	2000	2011
RABOBANK GROUP	2002	2011
UBS AG	1999	2011
ZURCHER KANTONALBANK	2000	2010

- **Users** Users involved in the project are covered by the subject "*users*". The users are divided into two types - functional users and domain experts. Functional users are those who only participate in the system for simple functional operations, such as navigating original documents and visualizing the data, whilst domain experts cooperate in the project at a more profound level. Domain experts use the system to carry out annotation and to get analytical feedback. Experts in the same group may work on the same documents and finally their data will be aggregated.

- **Coding** The annotated information is extracted from the original documents in form of quotations and codes, so this action is named "*coding*" [32][46][26]. Coding is the core element in this subject area model and it connects to all other subjects. The data covered by this subject includes static data and event data. Static data comes from the knowledge defined beforehand, such as domain specifications and terminologies. Event data records the evolving history of the entire project.

- **Evaluation** Experts give ratings for the performance of institutions based on their actions, efforts, and results. This means expertise information in form of scores has to be imported into the system for strategy examination. This subject furthermore provides the possibilities for the users' annotations to be validated. With the validation, the historical data submitted from the users can be refined to make sure that only the valuable data remains in the system. It is easier to observe the implicit patterns from the derived data after it is reduced.

- **Analysis** Algorithms are applied to the existing data and reveal new results as feedback to domain experts. For instance, latent topics in the corpus can be discovered and associations between coding and evaluation may be found. Later in this thesis,

topic discovery, ontology learning, and inference will be designed to support semantic studies.

The Subject Area Model illustrated above depicts the general view of the project, whilst a derived *Business Data Model* elaborates the entities in each subject area. This model is described in detail in figure 3.5.

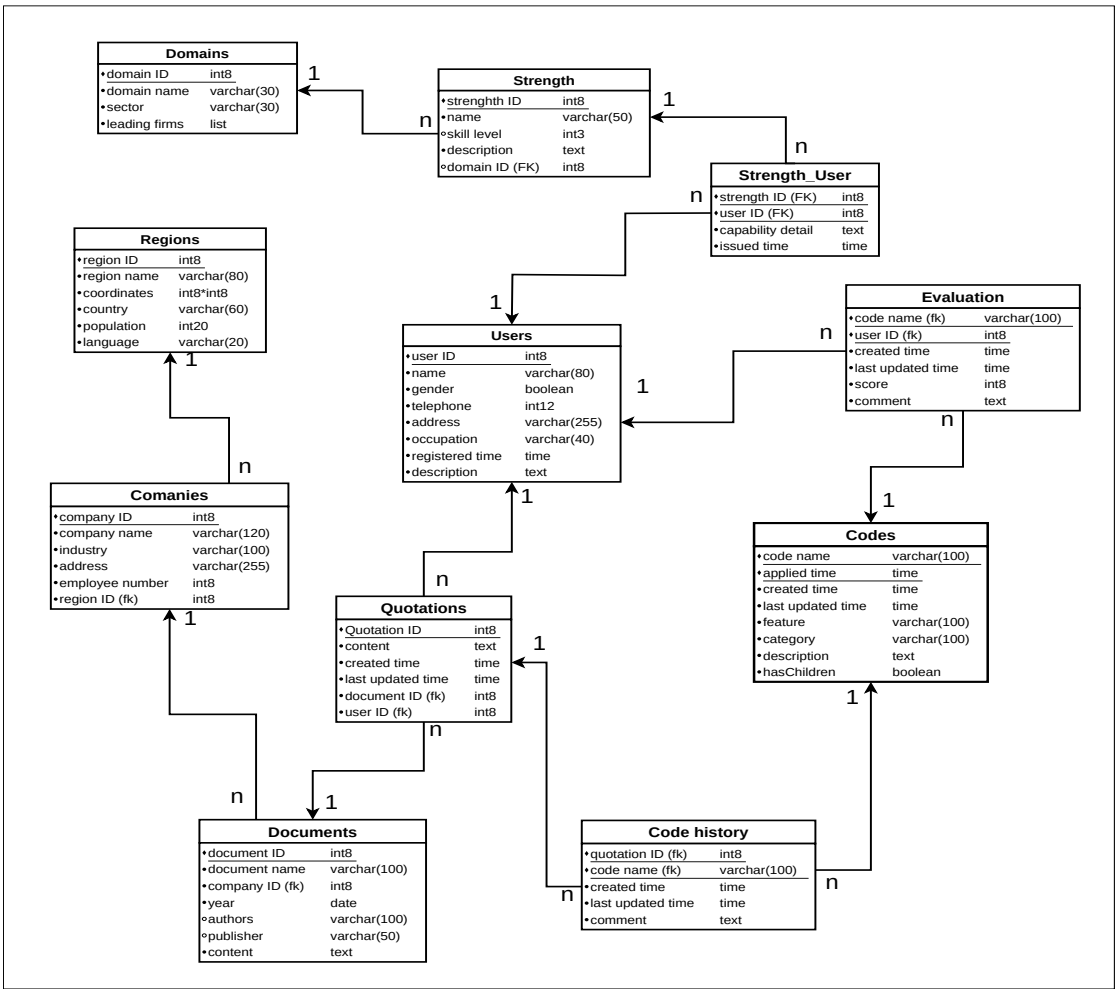

Figure 3.5: Business Data Model

(Source: [31] (improved))

3.4.2 Features of the Models

The previously established data warehouse models promote the entire methodology by offering the following capabilities:

- **System development** As the data is already archived in the form of linked entities, it is then easy to develop derived systems in order to represent or investigate this data. Existing techniques such as SQL and programming libraries are able to provide agile development solutions following the proposed data warehouse structures.
- **Function reuse** Occasionally the users' demands may give the impression to be very different but as a matter of fact they are very often similar from a technical point of view. By converting these demands into data warehousing requests, existing functions can be reused to a large extend to avoid redundant development.
- **Feasibility validation** With the assistance of theories in databases and data warehousing, e.g. *normalization*, *calendar*, and *hierarchies*, new requests can easily be validated in regard to their feasibility or evaluated for their computational complexity. This helps domain experts to assess their strategies with information technologies.

3.5 Conclusion

A suit of ontologies were designed to be the tool for bridging the gap between theoretical methods and system implementations. Based on this tool, the ETL process, a restricted set of ontological data structures and systematic tools were specified. The purpose of this process was to build a series of data warehousing models to achieve the goal of cleaning high quality data in some normalized forms. Based on the data stored following these models, it becomes much easier to fulfil meaningful tasks such as acquiring semantic patterns from the data. Approaches for the topic discovery functions will accordingly be presented in the next chapter.

4

Thematic Modelization
- LDA based Topic Analysis

In the previous chapters, ontologies were introduced to represent and formalize the annotated knowledge. Furthermore, data warehousing models were established upon the ontologies via an ETL approach. These steps provide models for the data for the purpose of offering convenience for a wide variety of applications such as data marts or data mining. One of the main outcomes of this approach is supposed to be the possibility of discovering implicit topics in the documentation, of revealing and then refining the thematic concepts and their semantic links. In this chapter, a novel methodology, namely *cLDA*, is presented. It serves to further analyze the topics inferred from the acquired data based on *Latent Dirichlet Allocation (LDA)*. The core model of *Dirichlet distribution* will therefore be presented as the basis of LDA. The Dirichlet distribution is an advanced idea coming from a series of fundamental statistical concepts and theorems. It has many promising advantages for text investigation and natural language processing. Later in this chapter we will present an improved algorithm which is able to improve the results of the basic LDA.

4.1 Dirichlet Distribution

The Dirichlet distribution, denoted as $Dir(\alpha)$, is a distribution class induced by a small set of basic distributions. The discussion in this section starts with the introduction of Bernoulli trials and of binomial distribution as some of the fundamental concepts.

4.1.1 Binomial distribution

Definition *Bernoulli trials* is a series of independent experiments with binary random outcomes - 1 and 0 [47]. Assume that the probability of the outcome 1 is p. The variable modelling the output of a single trial (denoted $X(p)$) is called a *Bernoulli variable* of which the probability density function is given by [48]:

$$f(k;p) = P(X(p) = k) = p^k(1-p)^{1-k} = \begin{cases} p & \text{if } k = 1 \\ 1-p & \text{if } k = 0 \end{cases} \tag{4.1}$$

Definition The variable modelling the number of outcomes 1 in a series of n Bernoulli trials (denoted $B(n,p)$) follows a *binomial distribution* whose probability density function is given by [49]:

$$f(k;n,p) = P(B(n,p) = k) = \binom{n}{k} p^k(1-p)^{n-k}, \text{ for } k = 0,1,..,n \tag{4.2}$$

In the case when $n = 1$, binomial distribution turns to Bernoulli distribution [49].

Definition *Multinomial trials* is a series of random independent experiments with multiple possible outcomes, namely $O_1, O_2 ... O_k$, so in this sense Bernoulli trials are a special case of multinomial trials (for $k = 2$) [50]. The variable modelling the output of a single multinomial trial is denoted $X(p_1, p_2, .., p_k)$, where

$$p_i = P(X(p_1, .., p_m) = O_i), \text{ for } i = 1..k \tag{4.3}$$

Hence, the vector $\mathbf{X} = (X_1, X_2, ..., X_k)$ modelling the number of each possible outcomes in a series of n multinomial trials follows a *multinomial distribution* whose probability density function is given by [50]:

$$f(x_1, .., x_k; n, p_1, .., p_k) = Pr(X_1 = x_1, .., X_k = x_k) = \begin{cases} \frac{n!}{x_1!...x_k!} {p_1}^{x_1} ... {p_k}^{x_k} & \text{if } \sum_{i=1}^{k} x_i = n \\ 0 & \text{otherwise} \end{cases} \tag{4.4}$$

where x_i are positive integers.

Another important distribution class for our purposes is the *beta distribution*.

Definition The *beta distribution* with parameters α and β is defined by the density function [51] :

$$f(y;\alpha,\beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} * y^{\alpha-1}(1-y)^{\beta-1} \tag{4.5}$$

In this equation, $y \in [0,1]$ is the variable of the probability density distribution, whereas α and β are two dynamic real positive parameters subject to adaptation. The beta distribution (denoted $B(\alpha,\beta)$) relies on the *gamma function* [52]:

$$\Gamma(z) = \int_0^\infty t^{z-1} e^{-t} dt \tag{4.6}$$

4.1.2 Dirichlet Distribution

Based on the definitions and theorem given above, the most significant distribution for our proposed method - the *Dirichlet distribution* - is going to be discussed in more details in this section. This distribution is the theoretical foundation of the *Latent Dirichlet Allocation* - an approach for topic discovery.

Definition *Dirichlet distribution* is a distribution for the parameters in multinomial distribution. The probability density function of a Dirichlet distribution, defined on the simplex $\sum_{i=1}^{k} x_i = 1$, is formulated as follows [53] :

$$f(x_1, ..., x_{k-1}; n, \alpha_1, ..., \alpha_k) = \frac{1}{\beta(\alpha)} \prod_{a}^{b} x^{a_i - 1} \tag{4.7}$$

where

$$\beta(\alpha) = \frac{\prod_i^k \Gamma(\alpha_i)}{\Gamma(\Sigma \alpha_i)} \tag{4.8}$$

4.2 LDA Model

In the previous section, several probability distributions were discussed. They are considered to be the theoretical foundation of the topic models we are going to use since at moment most of these models are based on probabilities. On this basis the idea of topic model is elaborated in this section. A discussion of these models leads to the selection of the Dirichlet model as the candidate for implemented and improved solution of topic discovery and refinement [54][55].

Based on the discussion given above we decided to used *Latent Dirichlet allocation (LDA)* to discover the latent topics implicitly distributed in the documents in order to overcome the drawbacks of the models presented previously. The classical Latent Dirichlet allocation, which is regarded as a generative probabilistic model, is established on the basis that topics are combined to represent a document. These topics are distributed with probabilities over terms [54]. LDA establishes the following generative process for each document w in a corpus D [54]:

1. Generate $N \sim Poisson(\lambda)$.
   LDA model assumes that the number of words N follows a Poisson distribution. In most cases of our project N is a large number, and the Poisson distribution has its particular advantage that, for very large N, it is a good model to estimate events [56].
2. Generate $\theta \sim Dir(\alpha)$.
   α is the parameter of the Dirichlet distribution in equation 4.7. Two other parameters θ and β will be used to generate the topics' distribution.
3. For each of the N words w_n:

- Generate a topic $z_n \sim Multinomial(\theta)$.
  Notice that θ - the parameter used for topic generation - comes from the Dirichlet distribution in step 2. That is the reason why this process is called a Dirichlet-based model.
- Choose a word w_n from $p(w_n|z_n, \beta)$.
  This step, a multinomial probability conditioned on the topic z_n, essentially assigns a group of words to the previously selected topic.

LDA is a simple but effective to capture some patterns represented by latent topics in primary documents [57]. The overall process of LDA is not very complicated, but it provides good performance for capturing the themes in the text.

4.3 cLDA - an Innovative LDA Model

Based on the LDA model discussed above, we will present in this section our own approach, namely *cLDA*. This approach takes advantage of the features of data produced by case studies as well as of the works conducted by domain experts. Experiments, presented in section 8.4, reveal that it outperforms the classical LDA in respect of effectiveness of topic discovery. The whole process is described in the following.

4.3.1 Stop-word Removal and Normalization

Once the primary document has been annotated by domain experts, in a subsequent process the topics which have significance for the text have to be extracted. At the beginning of the topic extraction some non-informative words are supposed to be removed. Here three steps are necessary:

1. The list of *stop-words* has to be created. The removal of stop-words is a vital step for the systems dealing with natural language processing.

2. Furthermore, the *Porter Stemming Algorithm (PSA)* [58] is adopted to to stem the vocabulary. It is mainly designed for the English language, but the general idea can be applied to quite a few other alphabetical languages. The simpleness of PSA is underlined by the fact that it utilizes only 60 rules which are easily understandable to human beings and furthermore it has very good performance regarding the execution time [59].

3. Although PSA is an advisable tool for the stemming task, a disadvantage of this algorithm is that some important words are removed or over-simplified. The reason of this phenomenon is very obvious: although English is a natural language in many aspects and it is relatively straightforward to handle compared to other languages [60] [61], still plenty of knowledge is difficult to express at an algebraic level, and grammatical exceptions widely exist. As a consequence, after applying the algorithm some semantics is lost compared to the original documents. For instance, the meaning of "*letters*", standing for literature, is radically different from "*letter*".

As a result, a *preservation-base* has been constructed, which consists of the words which will be maintained rather than being stemmed.

4.3.2 Feature Creation

For the remaining words after the removal of the stop-words and normalization, certain features can be extracted in respect to the different dimensions of interest. Here the created features vary from case to case. The main functionality to be provided during this step is *aggregation*. For each document d, for example, $(\mathbf{Q}_d, \mathbf{C}_d)$ is created in form of a matrix recording the quotations and their codes in this document. They are used as the input of the word-topic model.

4.3.3 Topic Generation

Based on the previous discussion, our choice is oriented toward a variant of the *Latent Dirichlet allocation model*, which is a *Bayesian probabilistic model* introduced in [54]. The LDA model uses words in documents as elementary data, but the most noticeable advantage of this probabilistic generative model is its capacity to be extended, by including supplementary information. For this reason a variant of LDA is being considered - the *coding* based LDA model ($cLDA$) belonging to the same family which includes the *label-topic* model [62], the *author-topic* model [63] and the *tag-topic* model [64].

A very crucial property of LDA model is that this model is that there is no method to find out the parameters in a straightforward way, so estimation methods are necessary to estimate the parameters [54] [65] [66]. In order to estimate the parameters, methods such as *Expectation-maximization(EM) algorithm* [54] or *Gibbs sampling algorithm* [67] may be employed.

4.3.4 Topic Selection Algorithm

For the label-topic model approach in respect to the selection of the topics and their terms, there are two questions to answer: how are the significant topics opted and for each topic, which terms are closer to the themes compared to the others. The first question is crucial in reality since the number of topics produced by LDA is usually large, but only part of the topics are ultimately adopted by the users. They are usually interested by the relevant and informative topics. The second question aims to find out the most important terms in respect to the topics. In order to answer the first question, a *Topic selection algorithm* is designed as described in Algorithm 3 with the following steps:

1. *The result set of the LDA (G) is obtained* [1] *and an iterative process is started. For each loop, one topic is picked and initial weight is assigned to it.*

[1] The result set consists of a list of topic groups. In each group, there are several terms and the first term is the topic of this group.

Algorithm 3 Topic selection algorithm

Require: O, G, P as the input, *weight* as the output
Ensure:

```
1:  while G.hasNextChild do
2:     topicGroupI = G(i)            {get a group of terms from the result}
3:     topicI = topicGroupI(0)       {get the topic from the group}
4:     if topicGroupI.size < φ then
5:        weight(i) = weight(i) − α ;
6:     end if
7:     if ∃τ ∈ O.codes, topicI ∈ τ then
8:        weight(i) = weight(i) + ε1;
9:        if ∃τ ∈ O.codefamilies, topicI ∈ τ then
10:          weight(i) = weight(i) + ε2;
11:       end if
12:    end if
13:    if ∃τ ∈ O.quotations, topicI ∈ τ then
14:       weight(i) = weight(i) + γ1;
15:    end if
16:    while P.hasNextPage do
17:       if ∃τ ∈ P(pCount), topicI ∈ τ then
18:          count(i) = count(i) + P(pCount).count(topicI);   {count the frequency of
             topicI on the current page}
19:       end if
20:       pCount++;      {pCount represents the page number of the current document}
21:    end while
22:    if count(i) ∈ ω1 then
23:       weight(i) = weight(i) + η_ω1;
24:    else if count(i) ∈ ω2 then
25:       weight(i) = weight(i) + η_ω2;
26:    end if
27:    i++
28: end while
```

2. *The number of terms belonging to this topic is checked. If this number is too small, then the weight of this topic is reduced.*

3. *If the topic shows up for certain number of times in the code list given by the users' annotations (O), then its weight will be increased due to its importance for the primary document.*

4. *If the topic is part of the text selected as a quotation, then its weight will increase accordingly.*

5. *The weight of a topic which appears for a certain number of times in the primary*

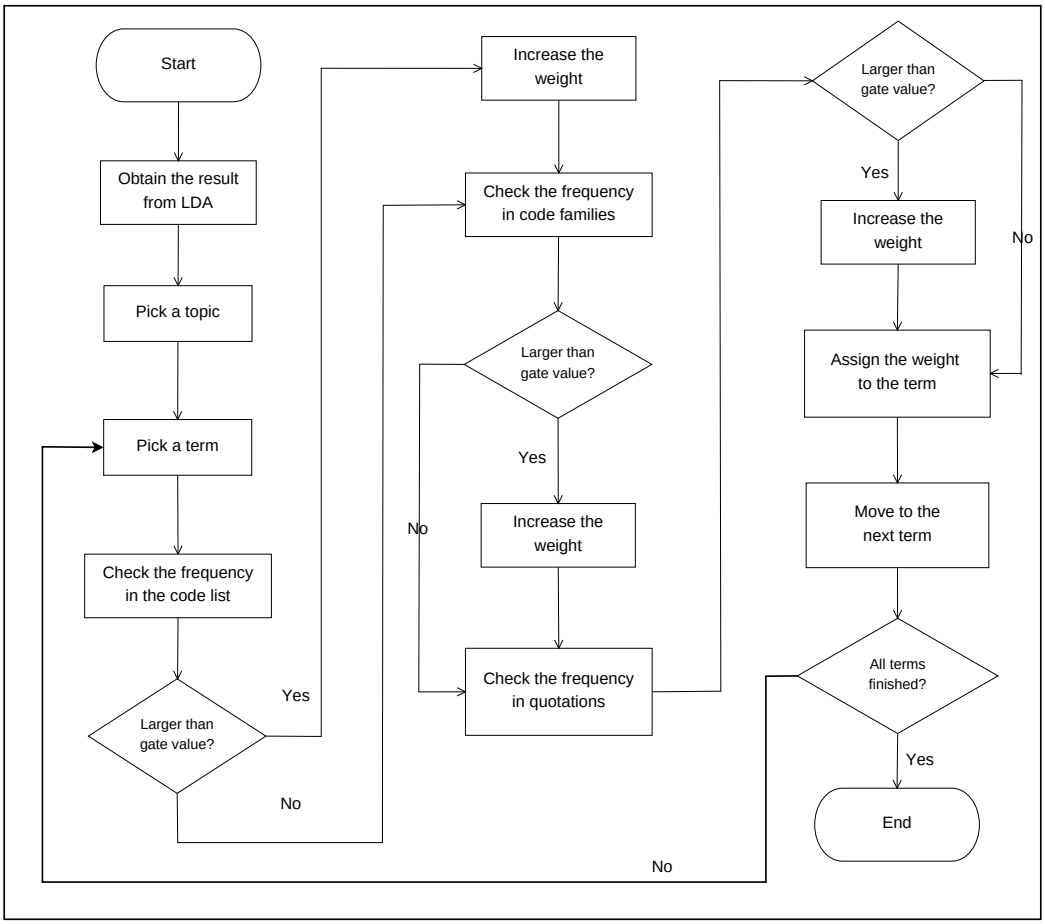

Figure 4.1: Term Ranking Algorithm

document (P) will be augmented.

4.3.5 Term Ranking Algorithm

The *Term ranking algorithm* targets at answering the second question stated above. This question is as important as the first one because once a topic has been fixed, it is desirable to refine its affiliated words extracted from the document. The principle idea of this algorithm follows a similar routine as the first one - i.e. to consider the annotation information. A sequential outline is summarized with in figure 4.1:

4.4 Conclusion

Founded on the basis of several fundamental probability concepts, the LDA model was discussed with Dirichlet distribution as the core element. Based on this model, a novel approach was established for topic discovery and refinery. This approach has several

advantages: it has a solid theoretical basis, it is straightforward to implement and it has a reasonable design considering the features of qualitative data.

A remaining issue concerns the question whether the acquired topics are somehow intertwined via implicit connections. In other words, not only the existence of these topics, but also how they are interconnected, matters to the comprehension of the data. In the next chapter a method will be introduced to show how these topics can be organized in a hierarchical structure.

5

Hiearchical Modelization
- Ontology Learning

In the previous chapters, ontologies as the knowledge representation tool have been set up for the purpose of formalization and modelling. The analytical annotations were designed to interact with domain experts and to be loaded into the data warehousing models via ETL methods. The topics were discovered and refined based on Latent Dirichlet Allocation as well as based on improved LDA algorithms. These approaches allowed to establish a mechanism for knowledge retrieval at the vocabulary level. In this chapter, on top of the work above, the next research question is addressed namely how to set up a more sophisticated framework to acquire the ontological hierarchies for the inferred vocabularies. Once this objective is achieved, it can be used for many applications such as ontology evaluation, recommendations, and inference.

5.1 Ontology Learning

Ontology learning is the focus of this chapter. The reason why this topic is stressed in this work is very simple. With the development of web technology, larger and larger vocabularies are discovered from text. However, without finding the semantic links among the acquired concepts, it is still not feasible to obtain the global view of these concepts and elaborate their relations. In the *Semantic Web*, hence, the challenge we are confronted with is exactly the issue of ontology population [68]. From a practical point of view, ontology learning is very promising and productive, because its output facilitates all the steps of production in industrial domains [69]. The learning process has an advantage in that it adopts multiple types of original data owning similar semantic framework and finally make them well integrated, which is highly beneficial for ontology population [70].

The research literature of ontology learning indicates that this topic is very active, depicted by the augmenting number of applications in this field. However enterprises are still eager to obtain more satisfactory results based on these techniques, leading to new challenges [71]. In [72], domain information, including some initial concepts, is utilized for building ontological hierarchies, implemented in a system. The proposed work is however limited to particular domains and to a large extent relies on the incorporation of domain knowledge. Another commonly-seen approach of ontology learning is *"hierarchical clustering"* [73]. This technique organizes the hierarchies upon groups of terms and then transforms them into prototypical ontologies with the assistance of *"distance measures"* [74]. In [70] is argued that the integration of the existing knowledge is crucial while some new ontologies are established, and the structure of the data is an important factor to consider. The nature of data we are considering here is intrinsically non-structured and consists of natural language text such as Word, PDF documents or Web pages. At present, even though some support is provided such as manual ontology development, the existing methods for learning from unstructured data have not shown satisfactory results for ontology acquisition in a purely automated manner [74].

Consequently, an innovative method for ontology learning is necessary to be established and elaborated in detail. As a benchmark for the proposed method, *Cobweb* [75], a popularly accepted algorithm for hierarchical clustering, will be used. The performance of the proposition will ultimately be compared with Cobweb to demonstrate its advantages regarding semantic aspects.

5.2 WordNet

5.2.1 Features and Advantages

In this section, the most important characteristics of WordNet will be detailed. WordNet is a lexical dictionary containing a vast vocabulary. The original objective of WordNet was to indicate the connection between psychology, social vocabularies, and research topics such as linguistics computation [76]. According to [77], WordNet provides the level-based contents of different types of vocabularies as well as their semantic relations. The internal organization of WordNet at a semantic level is basically a tree-liked one. At the bottom of this structure there are usually some words which are very specific - the words without any disambiguity, commonly those which clearly define concrete names of some objects or actions. Pervasively exploited as the most widespread lexical database, WordNet has a couple of plausible advantages:

- It contains a relatively complete vocabulary [78], delivering a solution for many challenging applications which demand for a large data set of vocabulary as training data or as experimental testbed.
- Its vocabulary is always growing to include new words. This dictionary is thus a dynamically evolving data set [79]. The up-to-date augmentation of WordNet enables its adoption for the fast development of Internet-based documentation.

- WordNet is not only a static database. Instead, approaches established based on this lexical dictionary can answer many unconfirmed questions [80]. To put it differently, WordNet supports semantic investigations based on its structural concepts.

5.2.2 Applications of WordNet

Many applications have been developed based on WordNet to be used in different domains. In [81], the authors designed an application to clarify the disambiguation of words. The data sets of this paper came from articles of the *Wall Street Journal* [82]. A case study was carried out to apply the approach in this paper in order to validate the theoretical propositions.

In [83] is proposed a very interesting idea to use *PageRanking* - the algorithmic model for search engines - to rank WordNet synsets. Graph theory was employed to express the semantics in the synsets. Several experiments were conducted to support the intuition. In another paper [84], the authors implemented an application for semantic zooming in WordNet, following a similar principle used in many web applications. Structurally dynamic visualization in WordNet is enabled by this application.

There are also some works conducted to bridge the gap between WordNet and ontologies [85]. For example, in [86], *YAGO* is a data set of ontologies representing the knowledge derived by Wikipedia and WordNet. It contains a large number of entities and their relations. This data set was designed following a logical model with a query interface. Likewise, the authors of [87] aimed at offering a solution for the cooperation between ontologies and linguistics. They thus utilized WordNet as well as EuroWordNet as the lexical dictionaries. Besides, fuzzy ontologies combining ontologies and fuzzy theory were employed to represent the hierarchical knowledge. According to several case studies, it was observed that this approach is particularly well suited for environmental applications. [88] introduced a method to incorporate new knowledge in form of ontologies into an existing vocabulary in WordNet. This method has been applied in *KYOTO* system, a project for information exchange on different domains, languages, and cultures.

5.3 Ontology Population

In this section, a new approach is going to be exhibited to set up and populate the ontologies in a hierarchical perspective. This approach is based on WordNet and its semantic interpretation. As depicted in the detailed descriptions, it provides many appealing advantages compared to prior methods.

5.3.1 Semantic Distance

As the first step of ontology population, a method for measuring the semantic distance of word pairs is introduced. Only after obtaining the values of this measurement, their relativeness can be evaluated and then used as the tool for knowledge construction. So far, some existing work has already attempted to establish methods to compute the distances of word pairs. One of the principal methods of measuring the semantic distances based on

WordNet is *Jiang-Conrath Measure* [89]. This method is a combinatorial one, founded on two classical tools:

Node-based approach [90] This approach leverages the *"information content"* of the nodes (concepts) - according to [91][92], the quantified information content is the negative *log* of a node c:

$$IC(c) = [-logp(c)] \tag{5.1}$$

The intuition behind the formula 5.1 is that as the level of a word in the lexical hierarchies rises, the word is getting more inconcrete and thus its information content decreases accordingly [92]. We then have the following equation [92]:

$$sim(c_1, c_2) = max_{c \in S(c_1, c_2)}[-logp(c)] \tag{5.2}$$

In the equation 5.2, $S(c_1, c_2)$ is the set of the nodes which contains both c_1 and c_2, so the *max* function is to find the lowest level node that contains c_1 and c_2 [92]. Speaking in the terms of WordNet, this equation aims to find the *direct hypernym* of c_1 and c_2.

Despite the fact that this approach has many promising advantages, e.g. it is intuitive and straightforward to implement, there is a significant shortcoming of the node-based approach [93]: This approach does not consider the useful information in other parts of WordNet. It only utilizes the relations between the words with similar meanings as well as their parents and children. From time to time, even though two words are not directly linked, their semantics may be related and the simpleness of this approach neglects this fact.

The approach stated above is also called *information content based approach*. It mainly relies on the semantic content of the nodes in WordNet. In order to solve the problems stated just before, another method was designed, named *edge based approach* or *distance based approach*.

Edge based Approach [90] As the name implies, this type of approach looks for the shortest path in the WordNet map between two words and then defines the length of this path as the distance of this word pair. Even though the original idea of this method is very intuitive and reasonable, it nevertheless has several disadvantages [94] [93] [95]: The main drawback being the heavy influence by the established layout of the ontologies as its basis, as if the ontologies are not well established, then the results will be inaccurate; WordNet, however, was originally not initialized for estimating the distances between word pairs, so its structure will not be suitable in some cases for distance calculation.

In order to improve the performance of edge based approach, new methods have been proposed founded on this original approach. Scriver proposed a new set of definitions and re-defined the calculation criterion in his thesis [96]. Two words may be very close to each other, such as *sports* and *player*, but not similar. Based on this concept, the author designed a couple of methods to compute the semantic distance from WordNet and

then evaluated the experimental results of different measures. The authors of [97] innovated another approach to overcome the problem of simple edge counting. They extended the edges with directions - upward, downward, and horizontal. With these directions, a more sophisticated mechanism was designed for the similarities between words and this approach was validated by real cases. In [98], the authors used results from experience to argue that the current definitions of depth and density had their limitations. These limitations lead to the failures of computing semantic similarities in many cases.

Jiang-Conrath Measure [89] Jiang-Conrath Measure takes both of these two approaches (node-based and edge-based measures) into account, since each of them has its own strength. Plenty of work has been carried out to show that Jiang-Conrath Measure performs better than a number of other methods when calculating the similarities between words in WordNet [99]. For example, an experiment was conducted to compare the performance of correlations among the three approaches stated above [89]. From the reported result, we can see that Jiang-Conrath Measure provides more reasonable result than the other two benchmark methods. This approach is thus up to now the most widely-used one when semantic distances of the WordNet vocabulary are required.

5.3.2 Hierarchy Construction Algorithm

The topics obtained from LDA [1] reveal the main themes of the primary documents. This generative model succeeds in acquiring the key concepts hidden in a text. Nevertheless, one issue of this process is that the results from LDA are flat, which means that all the discovered concepts (topics) are assigned at the same level. Clearly this is a limitation, because in most cases the semantic themes of a document exhibit a hierarchical structure. Only with topic structures we will be able to express more precisely and deeply the knowledge found in the input documents. A *hierarchy construction algorithm* is thus designed to establish these hierarchical structures as briefly illustrated in algorithm 4. This algorithm is essentially an *ontology learning* process. It takes the topics acquired from the LDA model as input, and sets up the hierarchical knowledge using the information from WordNet. The output of the algorithm permits plenty of usages, recommendations being one of the most important applications.

1. Initially, several *target words* (T) are selected. These target words reveal the main topics and ideas of a document. For example, an article which reports on the car industry may include the target words "*automobile*", "*factory*", and "*production*". There are several ways to find the target words, e.g. using the title of the document, the keywords, or the words having the highest frequency of occurrence, after the removal of irrelevant words. Our approach is not limiting the method used for finding the target words. However, a good choice of the target words is highly influential to the final effectiveness of the algorithm.

[1] For simplicity, in this chapter, "LDA" refers to the LDA-based model established in the previous chapter.

2. In the second step, the semantic distances between each word and the target words are measured. The objective of this step is to calculate the relevance or importance of all the words in a document archived in the ontologies (O) via their distances to the target words, which act as the representatives of the document. The closer a word is to the target words regarding the semantic distances, the more important it is to the document. The results are put in a hashmap called *disTarMap*.

3. In the next step, the distance between each pair of non-target words is calculated based on the semantics given by WordNet. Similar to the second step, the Jiang and Conrath measure is principally employed as the tool for the distance calculation.

4. The non-target word pairs are sorted according to their semantic distances from step 3 and stored in a hashmap named *sortedWordMap*. For each pair, the word having higher importance to the document (according to step 2) is assigned as *word_superior* and the other word as *word_subordinate*. By default, a pair of words is always formulated as {*word_superior, word_subordinate*}. This rule is respected for the sake of systematic investigation and construction.

5. A new hashmap named *oTreeMap* is set up to store the nodes of the ontological tree. Whenever a new node of the tree is created, as an information unit which has been decided as sufficiently interesting to be maintained, it will be stored in this hashmap. A *DOM* tree is used to represent the constructed hierarchy. Therefore each node is called an *element* of this tree during the implementation phase. The words are then loaded into their corresponding nodes to generate the ultimate hierarchy.

6. For *sortedWordMap*:

   (a) Take a pair of words from the head of *sortedWordMap*.

   (b) Get the two words from the key of the current instance, which is archived in the form of *word_superior*, *word_subordinate*, and denote these two words as w_{sp} and w_{sb} respectively.

   (c) Check if the w_{sp} has already been processed. If not, create a new element w_{sp} for the ontological tree; otherwise get the element corresponding to the w_{sp} from *oTreeMap*.

   (d) Check if the w_{sb} has already been processed. If not, create a new element w_{sb} for the ontological tree; otherwise get the element corresponding to the w_{sb} from *oTreeMap*.

   (e) Create a new element for the ontological tree. The children of this new element should be the two elements w_{sp} and w_{sb} obtained in the previous steps. The new element will correspond to w_{sp} in *oTreeMap*.

   (f) Remove from *sortedWordMap* all the word pairs which contain the word w_{sb} in the second position

   (g) If *sortedWordMap* is not empty, go back to step 6a.

7. Output the result of *oTreeMap* to an ontology or XML file for export. By default all the knowledge goes into ontologies. They serve as the basis for further analysis such as ontology inference and validation.

The proposed algorithm has several advantages:

- First of all, it is efficient for constructing the novel knowledge hierarchies. As the test cases in chapter 8 reveal, this algorithm succeeds in setting up the hierarchies using the acquired topics. With the attained hierarchical structures, it is very advantageous to discover in-depth facts through ontology inference techniques.

- Next, this algorithm integrates well with the LDA model and its derived algorithms. The hierarchy construction algorithm relies on the topics as its input. According to the walk-through in chapter 8, our proposed algorithm ingests the prior knowledge in a proper way.

- Moreover, it is straightforward to put into practice. This is important as we intent to implement it with a programming language and perform the case studies with domain experts using its implementation. This algorithm is an essential link between the software and the domain experts and therefore crucial for practical purposes.

- Finally, different from most prior works which results in trees of topics, our proposed methodology provides outputs on form of graphs, and a tree can be regarded as a special type of graph. The essential difference between the graph-based results and the existing ones is that nodes located at different branches may also be correlated. For example, *car* is a child of *auto-mobile* and *gas* is a child of *resource*. Meanwhile *car* and *gas* are associated by the action of *engine consumption*. A graph, which is semantically more expressive, can demonstrate this case whilst a tree cannot.

Algorithm 4 Hierarchy construction algorithm (abbreviated)

Require: O, T as the input, *weight* as the output
Ensure:
  $disTarMap \leftarrow \emptyset$
  $sortedWordMap \leftarrow \emptyset$
  while $O.hasNext$ **do**
    $temp \leftarrow O.children(i)$
    $dis \leftarrow distance(temp, T)$ {*calculate the semantic distance of the current concept with the targets*}
    $disTarMap.put(temp, dis)$
    $i++$
  end while
  while $O.hasNext$ **do**
    $tempA \leftarrow O.children(j)$
    while $O.hasNext$ **do**
      $tempB \leftarrow O.children(k)$
      $disPair \leftarrow distance(tempA, tempB)$ {*calculate the semantic distances of the word pairs*}
      $tempKey \leftarrow concate(tempA, tempB, disTarMap)$ {*set up the tupple of word pairs*}
      $wordMap.put(tempKey, disPair)$
      $k++$
    end while
    $j++$
  end while
  $sortedWordMap = wordMap.sort()$
  ...

5.4 Conclusion

By employing the hierarchical conceptualization of WordNet, a method is proposed in this chapter to calculate the semantic distances and then construct the hierarchies of the topics acquired from LDA. The produced structures allow many extended applications, for example reasoning over formulated rules. The ontologies, as a tool for knowledge representation, offer a convenient infrastructure for their inference, making the discoveries of new knowledge straightforward. The detailed processes of ontology inference will be presented in the next chapter.

6 Inferential Modelization
- Ontology Reasoning

In the previous chapter, an approach for building the hierarchical structures of the acquired topics was proposed by employing ontology learning techniques based on the WordNet vocabularies. An hierarchy construction algorithm was emphasized as the tool to achieve this objective. The established hierarchies serve for a wide variety of applications in which conceptual knowledge is required as well as its structural relations. One of the most practical ideas of how to use the attained hierarchies is to put them into a reasoning process to conduct *ontology inference* to produce new knowledge. Ontology inference mainly deals with loading the prior facts into the working systems and then generating a series of novel knowledge according to the domain set in advance. For example, the core element of an ontology inference is its rule sets while a rule engine is a system to action over the defined rules. In our case the inference on ontologies plays the role of the rule engines, while the ontological knowledge structures act as the rules. This chapter will start with the discussion of rule engines and then proceed to the proposed framework for ontology inference.

6.1 Rule Engine

6.1.1 Features

Rule engines are a critical element in the *Semantic Web*, which takes advantage of *"metadata"* as tags to express extra descriptions about web content at a semantic level [100]. These tags, with the assistance of ontologies, are expected to provide information usable in an inference process [100]. There is a potential advantage of the inference on metadata - when the number of web documents is increasing to a very large value, it is still effi-

cient to handle the metadata without considering its affiliated documents [101]. As a consequence, the reasoning processes based on the tags are developed.

Furthermore, rule engines are recently gaining interest in the context of *linked data*, a way to share correlated data via linked connections through Internet transportation protocols [102]. The role of rule engines assisting linked data is very promising to discover the latent patterns behind the visible links. Taking the example of the research on citations of academic articles, a typical goal is to capture the invisible associations between different authors and their research topics in order to reveal latest academic popularities. With rule engines, the implicit patterns will be discovered and new trends of the research will be indicated. This type of discovery will be very contributive to researchers for the potential development.

6.1.2 RuleML

An anticipated achievement of rule engines includes two aspects: rule expressions and engine implementation. In this section, the focus is given to the representation of the rules whilst engine implementation will be discussed in the next section.

Although a rule language can be considered as a specification of static documents, a reasonable choice is crucial and tricky, because at present there are many options popularly employed. Among all the candidates, *RuleML* [103] is favoured [104]. RuleML has many advantages. A decisive one is that it makes use of XML for marking up expressions and exhibiting object-oriented features. We can observe that this specification is organized in a highly structural way, with several levels of concepts and values. A reason is that RuleML adopts XML to completely express its content. Therefore the advantages of XML can accordingly be inherited. To put it from another angle, the documents represented in RuleML are easy to be integrated with other XML-based knowledge, such as ontologies, due to their shared semi-structured data models. Moreover, rule assertions exhibited in RuleML comprise also functional information linking their input variables. This pattern is consistent with standard programming languages such as Java. So a rule segment can be converted into a Java implementation smoothly via a *deserialization* process. The pattern matching task mentioned above is therefore facilitated.

Given these positive features, RuleML is employed as our principle format of rule representation. The programming tools for XML are greatly useful to parse and develop the rule documents. For instance, if later in the project rules in other formats (*e.g.* tables or plain text) are required, it is straightforward to use XSLT to convert the original RuleML segments to the target formats.

6.2 Ontology Inference

One of the most desirable functionalities of the proposed system is the inference over ontologies. Since in our system a great deal of knowledge has been acquired, new discoveries can be anticipated for a variety of practical purposes. For example, if two codes (denoted as *codeA* and *codeB*) are affiliated to the same quotations for dozens of times, it

could be concluded that these codes are heavily correlated. Accordingly, once *codeA* is found to have been assigned to a new quotation (namely Q_m), it is very likely that *codeB* will also be assigned to Q_m.

A typical domain in which ontology inference is commonly applied is the Semantic Web. This domain basically relies on two pillars - ontolgies as the structured knowledge and rules for reasoning. Ontological knowledge functions are the facts to be put into the inference process with different matching patterns. The ontology inference task thus has to take advantage of rule engines as its fundamental tool. To facilitate the reasoning process, an appropriate solution for the rule engine is vital. It has to be able to well integrate with the existing ontological knowledge in order to produce new discoveries. To achieve this objective, Jena [105], a rule engine particularly developed for ontological assignments, was chosen as the basis of our solution. As some rules have been created in advance to express domain knowledge, these files will be parsed before pattern matching is applied. For example, the following scripts demonstrate how the rules are formulated:

[r1: (?projA <hasdocument> ?docB),(?docB <hasquotation> ?quoC),(?quoC <haspage> ?pageD) -> (?projA project_page ?pageD)]

[r2: (?projA <hasdocument> ?docB),(?docB <hasquotation> ?quoC),(?quoC <hascode> ?codeD), strConcat(", ?codeD,?codeTemp), regex(?codeTemp,'#([A-Za-z0-9])_[A-Za-z0-9]*',?codeM) -> (?projA project_code ?codeM)]*

[r3: (?docB <hasquotation> ?quoC),(?quoC <hascode> ?codeD), (?codeD <cname> ?nameE)-> (?docB document_code ?nameE)]

6.3 Validation Framework

Apart from the inference assignments, another crucial task is to validate the ontologies. The term "*validation*" includes a wide range of functionalities and the criteria to validate an ontology vary depending on the concrete requirements. Here a framework is proposed considering the details of the entire methodology, shown in figure 6.1. Concretely speaking, there are three types of validations:

- *Concept validation.* The first requirement of an ontology file is that it has to be consistent with the OWL standard as defined in chapter 3. As the ontologies are founded on the XML specification, the requirements of XML documents have accordingly to be respected. This is to say that the syntax of each element, such as tags, IDs, attributes, and their values are supposed to follow the specification. Only the ontologies passing the syntax validation process can be considered as valid ones. Otherwise modifications have to be actioned to correct the errors. So far there are already a number of applications developed for ontology editing and most of them support the error notification mechanism. So it is relatively straightforward to avoid this kind of problems.

- *Property validation.* Concept validation is essentially an elementary process to verify the atom-level segments in an isolated way whilst property validation verifies the connection among concepts in ontology files. It analyses the links among

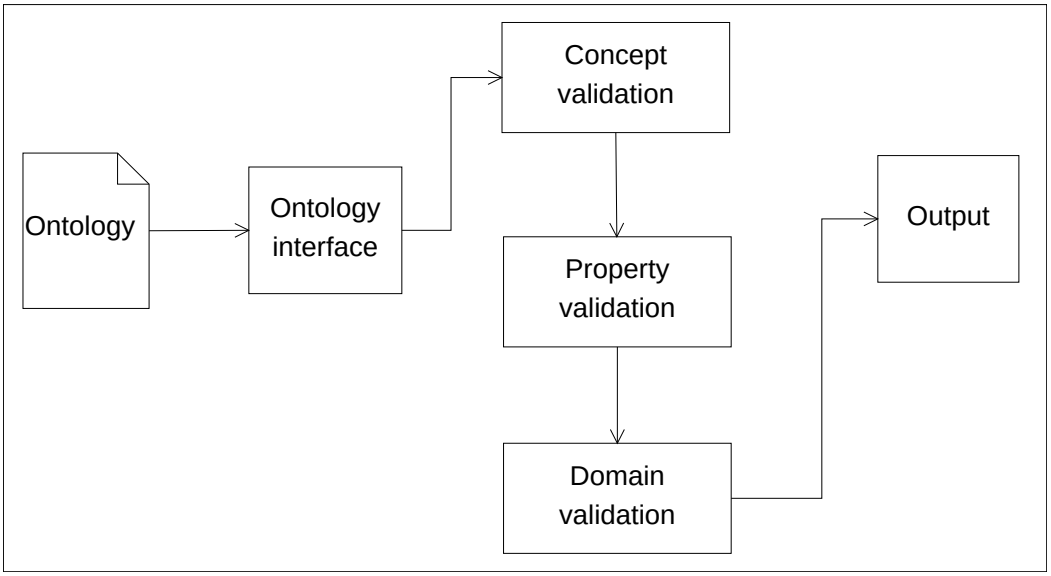

Figure 6.1: Validation Framework

ontological elements and is vital to guarantee the quality of knowledge in the system. For example, the *cardinality* constraints limit the number of instances that can be presented for a given property. This type of constraints is thus verified in our system.

- *Domain validation.* After the first two types of validations are performed, a higher-level validation is considered of particular importance if the project gets larger. In other words, even though after the first two validation steps the ontologies "seem" to be correct, they may still need some modifications. This mainly refers to the constraints imposed by the domain knowledge. For example, in some cases all the individuals in an ontology have to belong to the *gas* family because the objective of the project is to handle gas problems. Some external descriptions are thus desired to be incorporated into the validation process. Considering the concrete needs for this factor in our project, a *rule set* and a *knowledge set* were developed. The latter one expresses the necessary domain knowledge also in ontological forms and the former one specifies the requirements/constraints upon this knowledge set. The validation module navigates the rule set and verifies the consistency of the original ontologies.

6.4 Conclusion

This chapter mainly introduces the mechanism of inferencing over ontologies. The objective of this process is to extract the valuable information encapsulated in the acquired hierarchical concept structures. Ontology inference relies heavily on tools such as rule

engines, a powerful technique to discover new patterns of knowledge. A framework for ontology inference and validation was proposed.

7

Multilinguistic Modelization[0]
- Logographic Language Processing

7.1 Globalization and the Chinese Language

A set of methods have been established in the previous chapters exploiting the functionalities of ontologies to acquire and organize the thematic concepts from complex-structured data. Among all the influences on the relevant subject, *globalization* is considered as one of the most crucial factor in that the global economy growth has brought about substantial sustainability effects [107]. Furthermore, environmental issues have increasingly been of concern in an international context caused by the interdependence with global marketing activities [108].

The globalization is heavily affected by the popularity of the Internet and the English language [109]. Consequently, the Internet has to be updated constantly to function regarding the linguistic vector in order to promote globalization via cross-country activities [110] [109]. In this setting, for both economists and linguistics, it is vital to realize that the linguistic characteristics can be regarded as an issue of great importance to globalization [111]. For example, the Chinese language, as a language with a very large number of native speakers in the world [112], is drawing more focus as the economy of China is gradually getting more prosper [113]. Chinese has a history initializing from three thousand years ago, inherited by its content oriented teaching approach [114]. One of the most significant elements of this language is it is closely based on characters instead of alphabets inspired by the innate characteristics of its utilization in the daily activities [115]. We are thus motivated to discover some of them as the key elements in order to develop further investigation on documentation in Chinese.

[0]The work of this chapter has partially been presented in [106] and a series of more advanced amendment is added in this chapter.

A key difference between Chinese and English at the term level is that the former one connects a couple of symbols to shape new meanings whilst the latter one adds some additional letters to produce new forms of the same words [116]. Plenty of research indicates that Chinese can be well exploited in the comprehension cycles and it has recognition advantage over English [117] [118] [119]. Essentially, Chinese is a semantic-based language rather than a letter-based and phonetic one [120]. In other words, it is a language with evident qualitative features. Consequently, from the view of computer science, traditionally designed statistical and quantitative approaches are not sufficient to analyse documentations in the Chinese language. An innovative solution is therefore desirable.

7.2 Methodology

7.2.1 Paradigm Modelling

In order to study the unique features of the Chinese language, a data paradigm is established. This paradigm seeks to represent typical Chinese sentences with a list of variables. It is the necessary basis of our proposition in this chapter as all methods are built on this paradigm. Concretely speaking, the following six elements are included in the equation 7.1 [106]:

$$S = \{\phi, \eta, \gamma, \nu, \theta, \rho\} \tag{7.1}$$

In this paradigm, the subject, verb, and object of a sentence are denoted as η, ν, and θ respectively. These elements are the basic ingredients of a typical Chinese sentence. Besides, an adverb γ is placed before ν to add descriptive information about the verb. Two special elements - ϕ and ρ are added at the beginning and end of the sentence as the grammatical particles. They are adopted by our paradigm in that many sentences in Chinese have some particular words [1] at these two places to indicate the tense, negation, etc. as seen before. With this paradigm, machine based systems are able to handle the paradigm with techniques such as data mining to seek for in-depth knowledge encapsulated in the Chinese language. Besides, it works as the transitional bridge between Chinese and other alphabetical languages since most of these languages can be as well represented with this paradigm.

7.2.2 Ontology Establishment

On top of the defined paradigm, the task is to add some analytical approaches. In order to realize this task, three issues are involved - knowledge representation, feature extraction, and some advanced processing mechanisms. These issues will be discussed respectively in the following sections. At first, it is critical to set up a series of ontologies serving as the bridge between the proposed paradigm and the system implementation. Three types of ontologies are therefore established [106]:

[1] In this chapter, when we say a "word" in Chinese, it generally refers to a character, a short expression, or a short phrase of which the information amount is comparable to a word in English

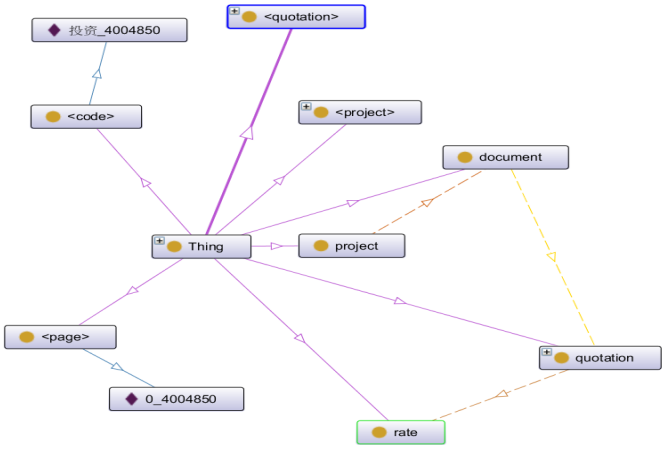

Figure 7.1: Content Ontologies

(Source: [106] (adapted))

1. *Content ontologies*. As shown in figure 7.1, this type of ontologies inherits the coding ontologies mentioned in chapter 3, which means that the annotation information from domain experts or other end users are embodied. All the APIs developed for English-based solutions can therefore be exploited for these ontologies as well. The implemented system can thus load these ontologies for semantic studies and visualization.

2. *Class ontologies*. Class ontologies, as the name implies, principally deal with the *speech of words* showing up in the content ontologies. For example, the word "快速地" (*rapidly*) is supposed to be labeled with a tag "*adverb*". These ontologies are leveraged not for the purpose of translating a sentence from Chinese to English. Instead, they seek to segregate the words in a sentence to establish maps according to the paradigm. Then the structure of this sentence will become clear for the systems. Figure 7.2 provides as example a part of a class ontology.

3. *Contributory ontologies*. The contributory ontologies are implemented to characterize the sentences with keywords at the syntactic level. For example, a character [2] "吗" from these ontologies indicate that a sentence is an interrogative one. These ontologies play an important role in distinguishing the tenses, tones, and structures of different sentences, not for the concrete content, but for their grammatical features. A segment of an contributory ontology is depicted in figure 7.3.

[2]In this chapter, "*symbol*" and "*character*" of the Chinese language can be used interchangeably, referring to the same meaning. As a general principle, when sentence structures are discussed, "*character*" is preferred to be used, whilst "*symbol*" is mainly mentioned to emphasize the graphical representation of Chinese comparing with the alphabetical languages.

7.2.3 Feature Extraction Algorithm

An algorithm (algorithm 5) is further designed to extract and identify the features from the sentences in Chinese [106]. The technical basis of this algorithm is to capture the keywords in the contributory ontologies with the assistance of the class ontologies. Content ontologies(O_{ct}), class ontologies(O_{cl}) and contributory ontologies(O_{cr}) are the input of this algorithm and the output will be the features m (tense, tone, etc) of the sentence q.

7.3 Amendment based on the Chinese Language

The methodology stated in the previous chapters provides a language independent capability of semantically studying the documented knowledge. However, adaptations have to be made in order to take into account some of the characteristics of Chinese in order to improve the effectiveness of our method:

1. Different from English, Chinese is based on symbols without spaces between each other. *Word segmentation* [121] has thus to be conducted considering this special feature. Once this step is completed, symbols will be separated into different symbol-compositions based on their semantic meanings.

2. Ontologies designed for the Chinese language are imported into our system following *facade patterns* [36] to maintain the extensibility. In this case, if later newly-designed ontologies are incorporated, the framework will remain stable.

3. For the analysis at the level of the sentences, we take advantage of the novel approach proposed in this chapter to retrieve their grammatical features. Topics produced by the LDA-based algorithms stated in chapter 4, plus the knowledge obtained from the feature extraction algorithm elaborated in section 7.2.3, will of-

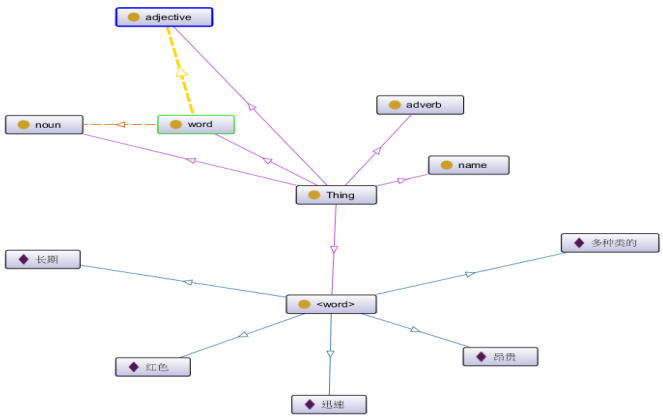

Figure 7.2: Class Ontologies

(Source: [106] (adapted))

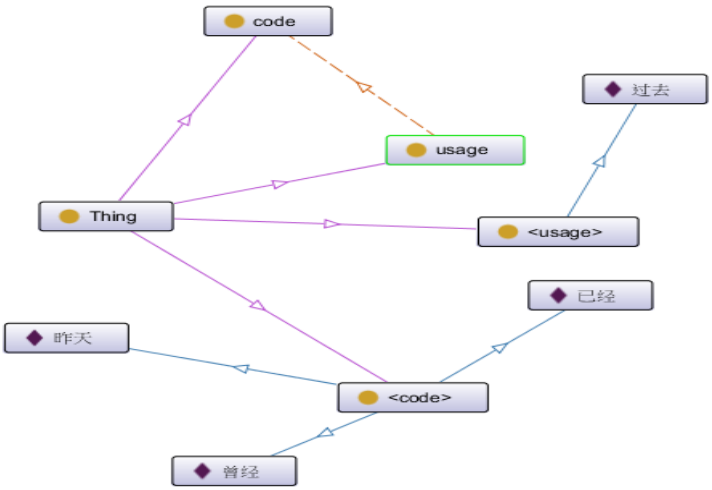

Figure 7.3: Contributory Ontologies

(Source: [106] (adapted))

fer a more specific explanation of the sentences at the grammatical level. When the Chinese processing module is invoked, the system starts to load the *contributory ontologies* to execute the extraction algorithm. Once some results have been produced, it will keep them in the working memory and push them to the output module for visualization and persistence.

4. Idioms assist the comprehension of topics. The composition of topics and idioms unveils more information at the level of concepts and statements. "乐不思蜀"(to be happy without missing home, from a historical story two thousand years ago), for instance, is frequently used for hospitality or tourism in modern business. A tuple *<tourism, fairly, positive>* is accordingly formulated and then used for inference later on. With this step, many latent topics can further be explored and developed.

5. With the objective of business demands, information of localization is occasionally needed. For example, from a certain expression in a business letter we can infer the region where a customer is from and as a consequence more local information can be integrated into the interaction.

7.4 Evaluation

7.4.1 Data Selection

In order to assess the methodology designed above, a critical assignment is to find some appropriate data for the semantic studies. The objective for this data is to act as the training set for the proposed algorithm to capture some key features in the Chinese language.

Algorithm 5 Feature extraction algorithm (Source: [106])

Require: O_{ct}, O_{cl} and O_{cr} as the input, OT as the output
Ensure:

```
O_temp ← O_ct
while O_temp hasNextChild do
    q ←O_temp.next
    while q hasNextChild do
        w ← q.next
        if ∃τ, τ ∈ O_cl ⋀ τ ∈ w then
            T_s = τ.getParent(O_cl)
        end if
        if ∃ϵ, ϵ ∈ O_cr ⋀ ϵ ∈ T_s then
            m = ϵ.getParent(O_cr).getKey
            t = ϵ.getParent(O_cr).getValue
            if OT_{q,m} is null then
                OT_{q,m} = t
            end if
        end if
    end while
end while
```

The selection of the data set is therefore highly influential for our final results. As a consequence, several criteria have to be fulfilled while the data is being selected [106]:

The main concern of this chapter is to discover the features of Chinese regarding its structures, and not a sentence-wise translation. Therefore the vocabularies adopted are supposed to be plain, without too many uncommon words. The abandoned words usually come from *combined ideograms*. For example, "高屋建瓴" (to pour water from the high roof) means that the speech or opinions from someone is very convincing. As "瓴" is a character irregularly emerging, we avoid these expressions in our dataset. Besides, One of the obvious difficulties in processing the Chinese language is that this language owns a lot of indirect expressions. An indirect expression means that the meaning of this expression is not intuitive even though the characters composing this expression are meaningful as such. Most of these expressions come from historical stories as well as local cultures. For example, "阳春白雪" (sunny spring and white snow) means decent music and art, which is far away from its literal meaning. This kind of expression will cause a series of problems. Therefore only the expressions straightforward to understand will be focused on in our data set. Another challenge of handling the Chinese language is that many of its expressions do not have a clear structure. Important elements of these expressions are omitted, allowing the readers fair amount of space to analyse the semantics by themselves. This approach makes sense for human readers, but is not practical for machine based systems. In order to simplify the established model, the focus of this thesis is attributed to those sentences with a clear structure and high comparability to English.

Considering these factors listed above, a dataset of Chinese language has been established. It contains many typical sentences for different levels of Chinese language courses. A number of typical sentences will be extracted from the text and then some new sentences will be derived following the same structure. They will be used as the input data of our model for systematic investigation.

7.4.2 Feature Filtering

With this selected dataset and its derived sentences, the next target is to extract/filter some key features from these sentences. The process of extraction will benefit from the paradigm established above and also take advantage of data mining techniques. Concretely, a decision tree approach will be employed to discover the hidden principles in the Chinese sentences. Once this step is accomplished, the attained characteristics will be interpreted using the tuples included in the paradigm. In detail, the following steps are conducted [106]:

1. A *feature map*, at the beginning empty but for the purpose of containing a list of keys and values, is brought in. The words enclosed in the feature map are a kind of "*catalyst*" in the process. They are used to inspire the emergence of the latent patterns in the Chinese language.

2. Fill in the feature map with a *bag of keywords* containing the elements which have a strong impact on the syntax of the language. For example, present = "正 在", past = "昨 天", tomorrow = "将 来", etc. For simplicity, its size is limited.

3. Navigate each sentence word by word. Whenever a word initially defined as the value of an element in the feature map is discovered, it is replaced by its key in the feature map.

4. Then all the remaining words are replaced by random values. The content of the table is shown in table 7.1

After the previous step, we decide to apply a method of classification - decision tree, to extract the features from the data set. In order to achieve this purpose, Weka [122] was used. Weka is an open source software project which is widely used both for industrial and academic domains. With the assistance of Weka, we were able to produce the decision trees as figure 7.4 and figure 7.5. These decision trees reveal the features from the original data set. Taking some examples [106]:

$$(\phi = "past") \Rightarrow (s \in past)$$

$$(\phi = "future") \Rightarrow (s \in future)$$

$$(\rho = "ma") \Rightarrow (s \in Question)$$

Table 7.1: Feature Table

ϕ	η	γ	ν	θ	ρ	S_t	S_q
fut	0.16	0.97	0.8	0.11	ma	fut	que
fut	0.16	0.97	0.8	0.11	0.17	fut	sta
past	0.19	0.03	0.11	0.72	le	past	sta
past	0.66	0.71	0.72	0.83	ma	past	que
fut	0.34	0.51	0.74	0.29	ma	fut	que
0.06	0.86	just	0.77	0.89	ma	prs	que
past	0.35	0.81	0.55	0.59	ma	past	que
past	0.55	0.34	0.39	0.45	le	past	sta
0.08	0.66	just	0.58	0.81	0.25	prs	sta
past	0.66	0.5	0.88	0.17	le	past	sta
fut	0.82	0.03	0.14	0.65	ma	fut	que
fut	0.82	0.03	0.14	0.65	0.64	fut	sta
0.76	0.28	just	0.25	0.33	0.62	prs	sta
0.37	0.65	just	0.5	0.66	ma	prs	que
fut	0.34	0.51	0.74	0.29	0.5	fut	sta
0.82	0.27	just	0.68	0.87	0.7	prs	sta
past	0.32	0.29	0.48	0.85	ma	past	que

(Source: [106])

7.4.3 Grammatical questions

For a language processing mechanism, many potential topics can be involved. It is therefore impractical to set the evaluation in a range too wide to focus on. The major concentration of this chapter is to capture some principles and common features of Chinese which very often show up in the regular documentation. These features are different from the English language at the syntax level, leading to the difficulties for using existing methods. It is therefore desirable to find some solutions for handling these features by adapting methods stated in the previous chapters. We will focus on the following issues [106]:

1. *Tense*. No matter which language is studied, tense is a critical issue, because it influences the meaning of verbs in almost all the sentences, and verbs are considered as the pivot of a sentence. The mechanism of expressing tenses in Chinese is very different from English. The latter one uses suffixes at the end of a verb, commonly "*-ed*", for the past tense. Chinese, however, keeps infinitives in all the cases. It adopts some temporal adverbs and phrases as supplementary information for the verbs. For instance, "*The company* ***launched*** *their training last week*" in English is translated as "*The company launch their training* ***last week***" in Chinese [3] because

[3] In this chapter, we use the direct translation from Chinese to non-grammatical English to represent the

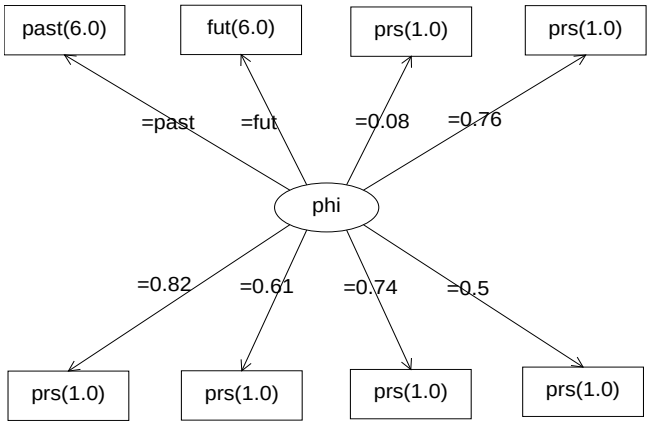

Figure 7.4: Decision Tree 1

(Source: [106] (adapted))

"*last week*" implies the past tense of the action "*launch*". This phenomenon can be formalized in the following way:

$$[EN]v \sim ed = [CN]v + time \tag{7.2}$$

The direct reason for this difference is that Chinese has many *pictographs*.

2. *Interrogation*. Another issue to consider is the interrogation. The ways of constructing interrogative sentences greatly differ in English from Chinese. In English, it is necessary to invert the order of a sentence as "***Do*** *you like this book*?". In Chinese, instead, there are several ways to pose a question:

original Chinese sentences to facilitate the understanding.

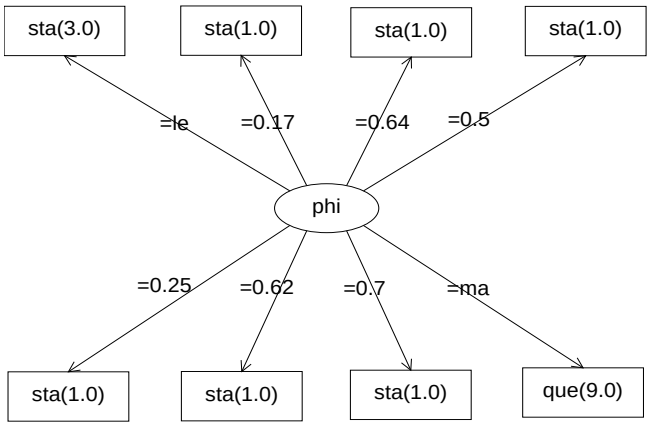

Figure 7.5: Decision Tree 2

(Source: [106] (adapted))

- *An interrogative particle*. For *General questions*, Chinese uses an interrogative particle "**吗**" to query the answers. "**吗**" is always placed at the end of a sentence, meaning "*You like the book,* ***right****?*".
- *An interrogative pronoun*. For *Special questions*, in Chinese the sentence is kept in a declarative order with the interrogative pronouns included - "*You like* ***which*** *book?*"
- *To interrogate the verb*. When a confirmation is needed between two options, the interrogation of a verb is presented right after the verb - "*You like* ***or not*** *the book?*"

7.4.4 Experiment and Discussion

With the purpose of validating the presented methodology, an experiment was carried out with ten sentences [106]. These sentences follow the principles stated in section 7.4.1. The target is to test whether our method will be able to discover the tense and tone of these sentences.

n	Sentences
1	昨天，我去商场了。(Yesterday, he go shops already)
2	他去上课了。(He go to the course already)
3	他去比赛了吗？(You go to the match already right)
4	这本书好不好看？(This book is very interesting or not)
5	我正在讨论问题。(We just discuss questions)
6	每天他都跑步。(Every day, he go running)
7	他跑步去了？(He go running already)
8	你要去哪里？(He want go where)
9	我准备看这部电影 (I plan watch this movie)
10	上次比赛以后，我很累 (After last match, I be tired)

Figure 7.6: Input Sentences

(Source: [106])

The results are presented in table 7.2 [106]: Values of tense and tone underlined are wrongly identified by the system, while the others are the correct answers. From the experimental results, the conclusion can be drawn that we are able to correctly detect the tense and tone for most of the sentences. In detail, for past tense, the results are more precise because the past tense in Chinese uses contributory adverbs as a strong indicator. This is more straightforward for systems to comprehend. For future tense, some results are imprecise not because the algorithm does not fit the question sets. Instead, the reason is that in Chinese sometimes a variety of regular verbs are used to express the intention in the future. For example, "*I would like read the book*" uses "*would like*" in place of "*will*". A reasonable solution is to design some more advanced algorithm to accumulate more verbs from the training data to the contributory ontologies. Then the results will be more accurate.

Table 7.2: Comparison Table

n	tense_1	tone_1	tense_2	tone_2
1	past	sta	past	sta
2	past	sta	past	sta
3	past	que	past	que
4	prs	que	prs	sta
5	prs	sta	prs	sta
6	prs	sta	0	sta
7	past	que	past	sta
8	fut	que	0	que
9	fut	sta	0	sta
10	past	sta	0	sta

(Source: [106])

7.5 Conclusion

As for the question of documentation processing, the influence of globalization, especially the Chinese language was discussed. Then a set of ontological norms as well as methods have been set up. These elements were designed based on a representative paradigm particularly suited for Chinese sentences. An algorithm was then proposed founded on this paradigm aiming to capture the typical features and to carry out topic discovery on the documentation with systematic approaches. The obtained results show that this approach owns reasonable capabilities in handling practical sentences in the Chinese language.

8 Implementation and Evaluation[0]

In the previous chapters, an approach has been proposed based on LDA and ontology learning to acquire topics from qualitative data and to build up novel knowledge. This approach tries to satisfy a need particularly formulated in the context of business data to find strategies in order to investigate the qualitative information embodied in business documentation. Functioning as a basic tool, ontologies play a vital role in the overall steps of this approach for knowledge formalization, annotation, transformation, and learning. According to the discussion in the previous chapters, we are convinced that the approach offers a promising solution in many associated domains. However the proposed approach has to be validated with the cooperation of domain experts since its ultimate goal is to offer analytical capabilities in the experts' regular working processes. In order to turn this idea into practice, a prototypical application has been implemented on which a number of tests were carried out. These tests are designed to verify the integration of ontologies set up manually and systematically. Furthermore, the discovered topics and derived knowledge will be evaluated for their effectiveness. Also, it is our intention to offer satisfactory usability to the end users, especially regarding the graphical interface, in order to improve the usability of the prototype.

8.1 General Framework

The developed system, named *Qualogier*, is established as a prototypical platform to fulfil the requirements stated above. Figure 8.1 and 8.2 present respectively the architecture and a screen-shot of the system.

[0]Part of the work has been presented in [123] and [32]. In this chapter there are more experiments, enhancement and details.

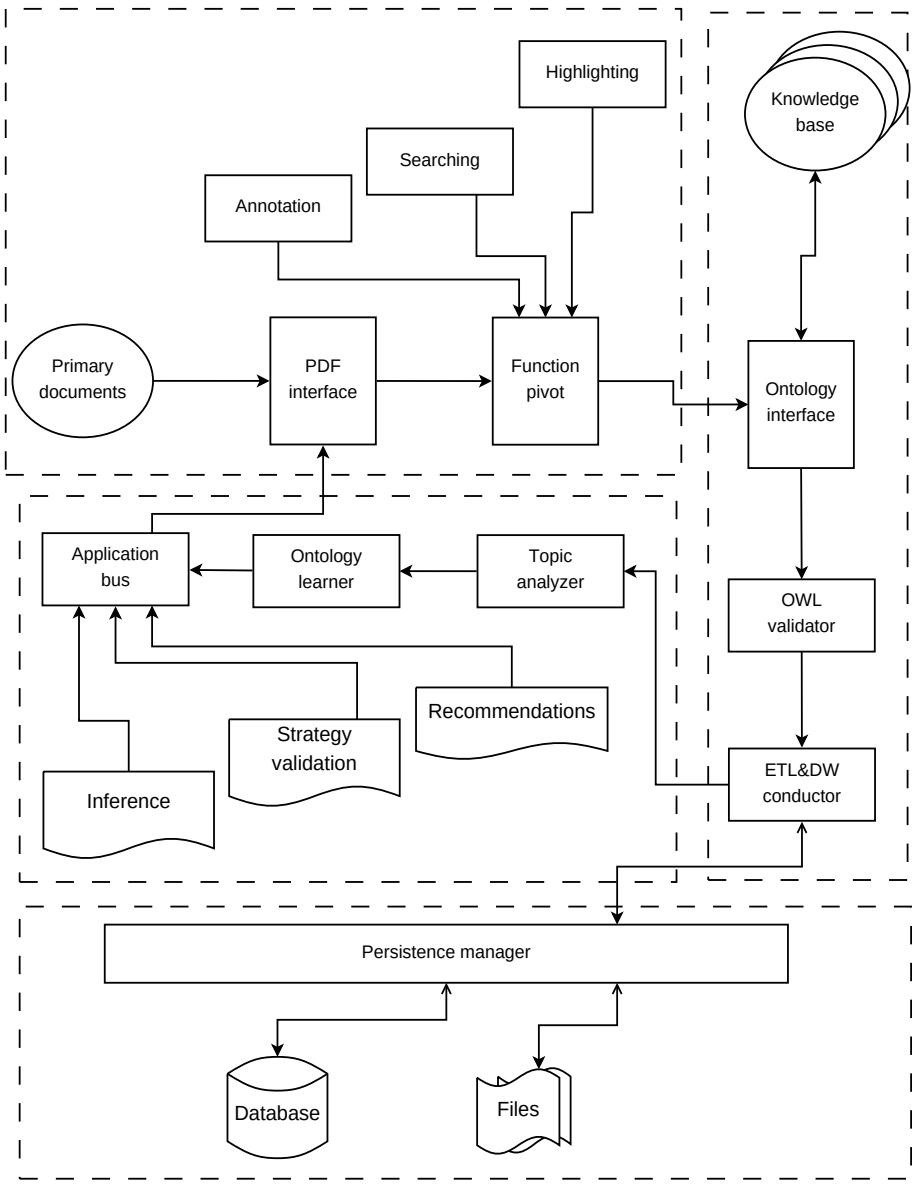

Figure 8.1: System Architecture

(Source: [123] [106] (improved))

8.1.1 Annotation Interface

The direct interface between the original data and the implemented system is called *Annotation interface*. It gives the users the functionality to operate on the primary documents. Without loss of generality, *PDF* files are specified as the format of the input. The reason is that PDF is a standard widely used in the domains of finance, enterprise man-

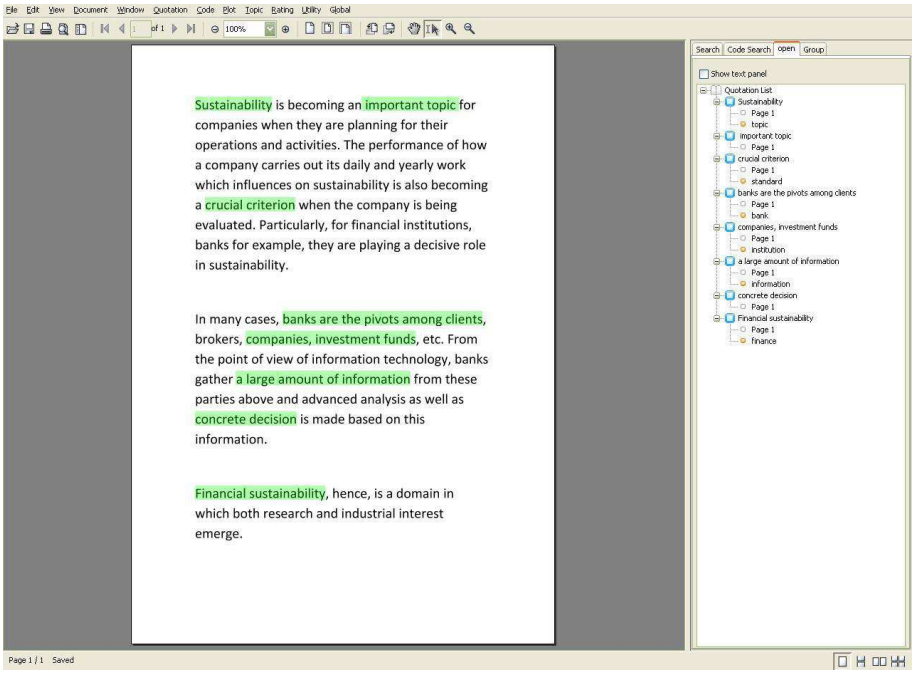

Figure 8.2: Screenshot

(Source: [31] [32] [106] [123] (adapted))

agement, etc. Meanwhile, many other input formats such as TXT, MS Word, HTML, and XML can automatically be converted to PDF. Another great advantage of the PDF standard is that all the phrases, words, and even characters can be located with two-dimensional coordinates. This opens the possibility to maintain the location - (X, Y) - of a specific part of the text of interest to the users. The PDF interface is established based on *ICEpdf* [124], a Java-based project for PDF navigation. The implementation is conducted based on formalization in appendix B. In detail, it offers the following functionalities:

- *Quotation selection and Quotation Tree* [32]. An essential functionality of *Qualogier* is to select useful words, phrases, and sentences as quotations [125][26]. The system provides a mechanism as follows: with the mouse's drag-and-drop functionality, a user highlights the text in a PDF file displayed by the graphical interface. Then he/she could right click the mouse to select this text as a quotation. Another frame on the right hand of the interface presents a *quotation tree* where all the quotations are visualized as outline. Whenever a new quotation is selected, the tree will automatically add a new node at its first level. Then all the quotations affiliated to the host document are displayed.

- *Coding* [32]. On top of the quotations, the users conduct coding actions. The codes they add are the keywords and labels to mark the quotations [46][26]. Here two types of code sources are provided: an external tree is provided by the domain

knowledge out of which the users can select codes to add. Furthermore the users can add any words according to their own opinions. Once a new code is added to a quotation, a corresponding node will be created in the quotation tree at the second level as child of the quotation. This is always the case since a code is always affiliated with a quotation.

- *Output* [123]. In *Qualogier*, data output in different formats is offered, e.g. the most frequent words in the navigated documents, the most frequent words showing up in the quotations, the frequencies of all the codes, and the numbers of codes of each quotation. All these types of output are stored by default in spreadsheets and can be converted to XML and HTML files.

- *Visualization* [32]. In order to give the users a direct impression of the resulting information, visualization functions are implemented. For example, a graph is used to visualize the distribution of quotations over the pages of a document. The X coordinate represents each page from the beginning to the end of the document, and the Y coordinate displays the text amount of quotations on each page. From this curve, the importance of each page can be inferred, measured by the text quantity of their quotations. Similar functionalities include the numbers of codes and the average values of ratings in each page.

8.1.2 Ontology Interface

While the users are operating on the documents for navigation and annotation through the PDF interface, another functional component is the ontology interface. This interface is significant because the proposed methodology relies substantially on ontologies. It is therefore desirable to have reasonable capabilities to handle the ontological data. In detail, the following functionalities are included in this component:

- *Project ontology loading*. At the beginning of the cycle, when *Qualogier* is initializing, a template defined as the project ontology is loaded into a specific directory of the system. During the entire process, the users' behaviour as well as the analytical output will follow this project ontology as the guideline of the overall system.

- *Reference ontology incorporation*. Besides the project ontology, another critical element are the reference ontologies that come from domain knowledge. A *schema* has been formulated to regulate the format of reference ontologies. All the collaborating experts will follow this schema in order to ensure the compatibility of the system. For ontologies or knowledge in other formats which already exist, transformation tools are proposed, such as *Powershell* and *XSLT* scripts, to convert this data into the format in accordance with the defined schema.

- *Ontology library*. In order to facilitate the basic operations on ontology files, a library was implemented to manage all the necessary functionalities. Technically, *Dom4J* and *JDom* are selected as the fundamental tools to implement this library.

The developed functionalities include loading an ontology file, creating a new tree from scratch, inserting a node into the existing tree, adding attributes and values to a node, deleting a node from a tree, updating the attributes, etc. This library is capable of dealing with the challenges when a large amount of data is being processed (see the experiments presented in section 8.3).

- *Ontology validation*. The ontologies incorporated into *Qualogier* are supposed to be validated in order to ensure that their syntax follows correctly the project specifications. Consequently, a component is added into the system to validate the input ontologies. Then the ontologies are loaded into the component to validate the consistency of their content with OWL and with the specific requirements of our project. The feedback is eventually given to the users for necessary modifications.

8.1.3 Recommendation

Besides, *Qualogier* offers the capability to propose recommendations to the users based on their previous behaviour and the characteristics of the knowledge [32]. The functionality of recommendations is of great use since the coding process requires the support of the system to find the appropriate codes for annotating a given document. In detail, a *collaborative filtering* algorithm [126] is employed to produce the recommendations. This algorithm assumes that for two people (denoted as A and B) who like a group of similar items, if later A likes a new item, then it is very likely that B will like this item as well. Plenty of work has been conducted on this algorithm to prove its validity and extend it into many new applications. With the assistance of the reasoning engine, this algorithm is implemented based on the ontology inference.

8.2 Experiment Objective

With *Qualogier*, a series of experiments have been conducted to validate the proposed methodology [1]. In the experiments more than thirty PhD and master students majoring in *International Business Development* and related majors have participated. They were divided into seven groups and each group worked on a number of reports from different banks in different years. The participants first designed an analytical framework based on which they conducted the task of report coding. Also they carried out the evaluation steps for each company according to their performance of sustainability indicated in the reports. All the data was maintained in ontologies and could be exported in form of CSV spreadsheets and XML files. The main objectives of the experiments are characterized as follows:

1. We intend to involve domain practitioners into the process of report examination based on our implemented prototype. The involved tasks must be also conducted on other systems such as ATLAS.ti and Nivo in order to have a comparison with state-of-the-art software.

[1]This work is supported by Swiss National Science Foundation (SNSF) project *"Formal modeling of Qualitative Case Studies: An Application in Environmental Management"* (No. CR21I2_132089/1).

2. The system will be entirely based on ontologies, as argued in chapter 3. Ontologies recording user behaviour are validated for their conformity in respect to the project ontologies and to the integration of domain knowledge. These ontologies serve for data persistence, knowledge incorporation, domain integration, and analytical studies.
3. Real cases are used to verify the established models and to asses the performance of the ETL processes in the case of massive data produced by multiple participants. One of the aims is to test whether the demanded data can successfully be retrieved from the data models. In addition, these processes should be operated in a highly automatic way.
4. The capabilities of outputting different data formats will be analyzed by the users for its re-usability in other external tools. The goal is to keep this approach as generic as possible.
5. Learning algorithms will be used to produce derived information targeting domain experts. These algorithms mainly focus on the latent topics for their innate links and semantic structures. The learning techniques in this thesis come from generative theories plus the accumulative knowledge from the users' action on the original data. These techniques are expected to discover new knowledge, not too obvious for the domain experts but still comprehensible. The ultimate goal is to improve their analytical work on qualitative data for case studies and theory conclusions.

8.3 Scalability

When the implemented system was being evaluated, the first issue that came to mind was *scalability*. When data is getting larger, computational complexity is increasing accordingly. A couple of experiments were designed and implemented to validate the efficiency of the system when large data sets and complex operations were involved. The experimental environment was set up on a PC with a 3.00GHz CPU and 4GB physical memory running on Windows 7.

We have first conducted an experiment concerning the loading of PDF documents into our system. As the textual documents used in enterprises becomes longer and longer, a problem might arrive in an implemented system if loading of raw data is slow. It is thus necessary to test this issue in order to ensure that the loading time is fast enough for the users. Ten documents were generated containing 100 pages, 200 pages up to 1,000 pages and then loaded into *Qualogier*. As depicted in figure 8.3, the X-axis represents the number of pages in each file and the Y-axis indicates the processing time to load these documents in millisecond. It can be seen from that the response time of loading 1,000 pages is smaller than *120 ms*. According to the opinions of domain experts, 1,000 pages is already a large number and should cover almost all practical cases. So the conclusion can be drawn that *Qualogier* is fast enough to initially load documents with large number of pages.

A second experiment concerns the users' interaction with the proposed system. In *Qualogier*, a typical operation is to select some interesting text using the mouse. For

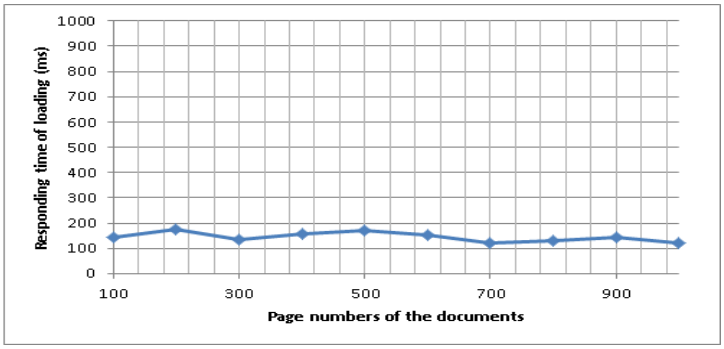

Figure 8.3: Loading PDF Files of Multiple Pages

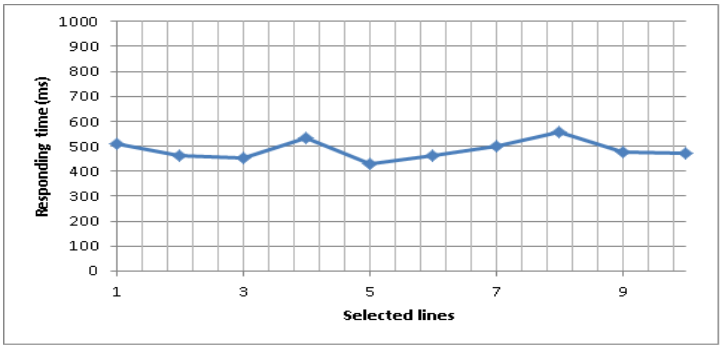

Figure 8.4: Multiple Selected Lines to be Painted

PDF files it is always a challenge to precisely extract and handle this text at the level of words and characters. When the users select text, they expect rapid response from the graphical interface of the system As a result, this experiment was designed to test the response time for selecting characters. The results are depicted in figure 8.4. The X-axis represents the number of lines selected and the Y-axis indicates the processing time of the graphical interface in millisecond. It can be observed that even if multiple lines are selected, requiring more memory, the response time is fast enough with the maximum lower than *600 ms*. An acceptable time for an interactive user.

Another experiment, which follows a similar idea as the previous one, focuses on the persistence efficiency, or the speed of ontology storage. After some text is selected and highlighted by a user, he/she is able to save this text as a quotation. This information is stored in an ontology file. As one can imagine, when the text amount of this quotation is large, the processing time may be slow for the I/O operation. In figure 8.5, the X-axis represents the number of lines selected as a quotation and the Y-axis indicates the processing time to archive this quotation in ontologies in millisecond. From this plot, it can be observed that when quotations are getting larger up to 10 lines, the corresponding time always stays on a small level, lower than *70 ms*.

We have in addition carried out an experiment for the case were multiple quotations are selected continuously. The reason why this experiment was designed is the following:

after a quotation is saved, the system takes certain time to process the user's action. At the same time, however, if the same user selects another quotation while the previous request is still being processed, then this may cause a problem of synchronization. In order to provide user friendliness and satisfaction and also to avoid system blockage, the following test was conducted. The number of continuous quotations are increases from 1 to 50, as 50 is already a number large enough for the concurrency test. In figure 8.6, the X-axis represents the number of continuously added quotations and the Y-axis indicates the processing time to store these quotations into ontologies in millisecond. From the result shown in this figure, it can be seen that even with 50 quotations added non-stop, the system still responses efficiently.

Based on the previous experiment, another test was implemented concerning the situation in which a large number of quotations exist [2]. When a user has already added many quotations stored in the ontologies, the system has heavier burden to process new requests compared with the work on an empty system. It is therefore of interest to verify whether the response time is still fast enough in this case. As a general rule, when n quotations have been added and archived in ontologies, the $(n+1)th$ quotation is added and the processing time is tested. The variable n starts from 0 to 999, which is a number large enough for our projects. From the result shown in figure 8.7, in which the X-axis stands for the number of existing quotations and the Y-axis indicates the processing time in millisecond, the conclusion can be drawn that, even if the number of existing quotations is large, the system still manages to handle the newly-coming requests and the response time is very satisfactory from the point of view of an end user.

The following experiment is testing the highlighting function. After a quotation is selected in a PDF file and stored in the ontologies, the system provides a utility that allows users to highlight these quotations in the graphical interface. At the technical level, the system loads the ontological files and then parses all the information, for the quotation's content, update time, and coordinates. The challenge here comes from the coordination because the amount of data is large, possibly leading to slow response. In

[2]The previous test mainly dealed with the question whether the system can manage a number of requests at the same time whilst this test concerns more adding new data with plenty of existing data.

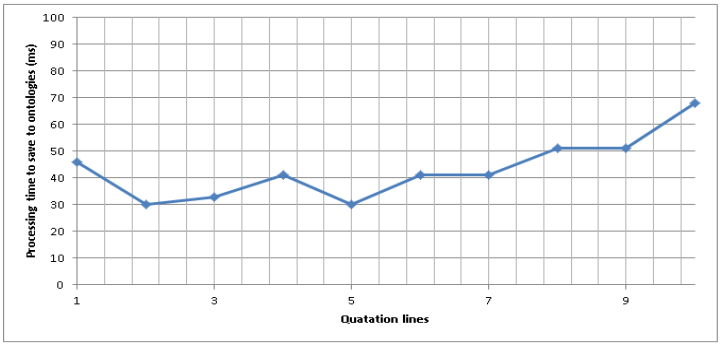

Figure 8.5: Quotations of Multiple Lines Saved in Ontologies

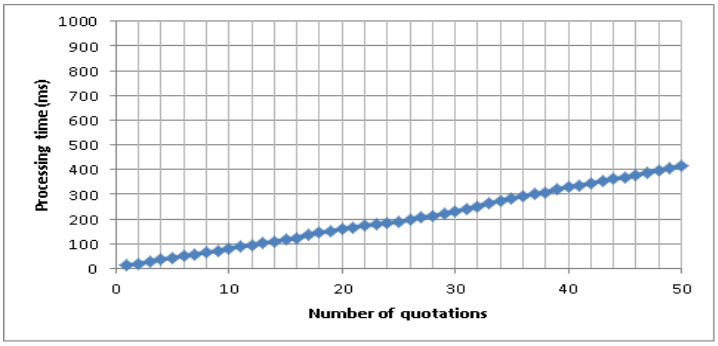

Figure 8.6: Continuously Added Quotations

figure 8.8, the X-axis indicates the lines of the quotations whilst the Y-axis represents the response time for highlighting the text of these quotations in millisecond. From the test results shown in this figure, it is apparent that even though the responding time increases as the lines of quotations increase, it always remains in a very satisfactory range, lower than 0.06 second. This result can be explained by the system architecture and the OWL APIs which provide reasonable efficiency.

Later on, another experiment was carried out to measure the efficiency for adding new codes [32]. This experiment aims to find out the processing time for inserting a new code when a large number of quotations already exists. To describe this process, the system needs first to figure out which quotation this new code belongs to by searching all the quotations and then add a new node to this quotation segment. For this purpose, 100 codes were generated automatically to conduct the above-stated process and then the average value was calculated. In figure 8.9, the X-axis standards for the number of quotations which already exist in the ontologies when a new code arrives, and the Y-axis represents the processing time in millisecond for inserting the new code. It can be seen that even in the case of a large number of existing quotations archived, inserting a new code can still be processed in an efficient manner.

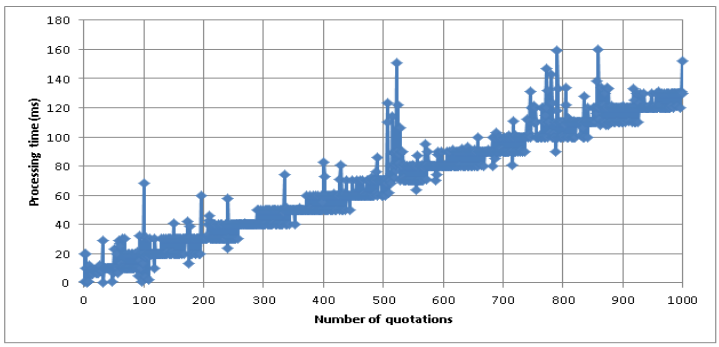

Figure 8.7: Adding a New Quotation upon Existing Quotations

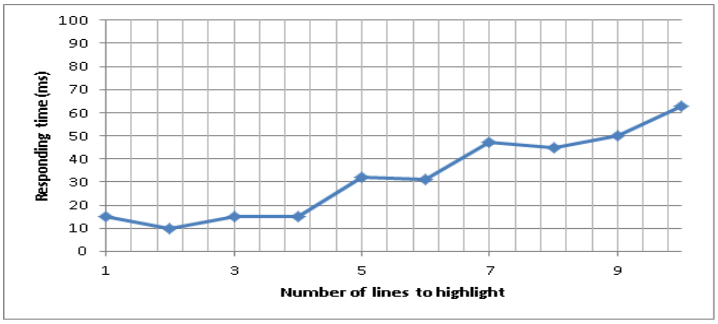

Figure 8.8: Multiple Lines Highlighted

8.4 Topic Discovery

After the efficiency of the proposed system was tested, the evaluation task then turns its focus on topic discovery from the qualitative data. As stated in the previous chapters, the LDA model has been selected to acquire the latent topics and enhanced algorithms were designed to improve the results from LDA. It thus is interesting to know the effectiveness of these algorithms. A couple of experiments have been designed for this purpose. The first thing of interest is to compare the original LDA and the improved algorithms. The following steps have been developed to conduct the evaluation process:

1. ***Validation data selection***. At the first step, some validation data is to be selected. The purpose of employing this validation data is to verify the effectiveness of the acquired topics. It is not an easy task to choose the validation data because there exist many criteria that influence this decision. We argue that domain opinions are the most important factor to consider, in the sense that the acquired topics will eventually serve for domain participation and development. Thus a method firmly based on domain analysis was developed to select the validation data. One way to achieve this is to investigate the coding information produced by the users. Table 8.1 shows

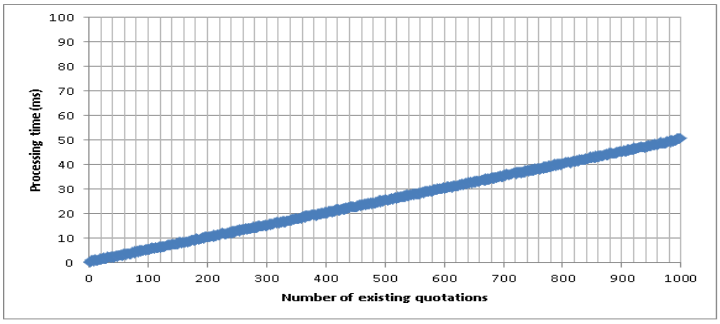

Figure 8.9: Inserting New Codes to Existing Ontologies

(Source: [32])

the total numbers of codes assigned to each document. Concretely speaking, the following method is proposed to compare the different sets of validation data:

(a) If a document contains much more thematic codes than another, then it is selected as validation data. Here the meaning of "*thematic codes*" depends on the concrete demand in each project. In our case, for example, codes such as "*sustainability*", "*environment*", and "*energy*" are considered.

(b) If a document contains substantially more codes than others, then it is selected as validation data.

(c) If a document was coded by much more domain experts than others, then it is selected as validation data.

(d) Otherwise select the document with the largest number of codes.

The number of documents to select as validation data can be set in advance. If multiple documents are engaged, then they will accordingly be merged as a combinatorial validation data set. It is worthwhile to notice that this method is highly extensible and adaptable. That is to say that the criteria based on code numbers can be replaced by any other standards if necessary, for example, the length of quotations. Some documents recommended by domain experts will be involved in the validation data set even though they were not selected based on the previously stated criteria.

With the selected data set for validation, we will be able to check whether the acquired knowledge is reasonable. The main idea is to input the topics of this validation data and then apply certain evaluation steps to obtain an output that compares the existing methods with the one we proposed.

2. ***Counting frequencies***.

Table 8.1: Relational Table

Code	Company	Year	Number
Applying standards	BKA	2006	14
Communication	BKA	2006	47
Community programs	BKA	2006	6
Continuous improvement	BKA	2006	16
Customer relationships	BKA	2006	13
Customers policies	BKA	2006	1
Developed Service	BKA	2006	23

Definition $count(w, d)$ is a function to count the number of occurrences of the word w in a document d. Then the frequency of this word w is:

$$F_d^w = count(w, d) \tag{8.1}$$

Definition $T_j(D_i)$ is the set of terms discovered by the method T_j over document D_i.

Assume that N is the number of terms in each topic group. For all $i_z \in T_z(D_i)$, "*evaluation value*" δ_z is defined to evaluate the performance of the terms produced by method z:

$$\delta_z = \Sigma_{k=1}^{N}(F_{D_i}^{i_{zk}} * (N - k)) \tag{8.2}$$

The intuition behind this method is straightforward - if a term shows up in a document for a large number of times, then this term is an important one. Furthermore, if a word ranks before other words in a topic group, it will be given a higher weight, because it has a position showing more importance. Then the overall importance δ is calculated by summing all the words. This strategy was executed on *Qualogier* to check the computational results. More detailed experimental results will be reported in the next paragraphs.

3. ***Comparison and interpretation***.
   The next task, on top of the previous preparation steps, is to compare the proposed LDA-derived method and the classical LDA method. As the new method we proposed offers more flexibility to adjust the ranks of the topics and their affiliated terms, from a theoretical perspective we could estimate that this method should produce better solutions than the original one.

   The idea behind conducting the comparison is to apply the algorithms to multiple reports annotated by the experts and to produce a series of terms, based on the frequency counting approach using the selected validation data. The results are shown in figure 8.10 and 8.11 as some examples. The blue dotted curves represent the performance of the proposed method whilst the red solid ones are for the classical LDA. The X-axis represents different terms produced by the algorithms and the Y-axis stands for their evaluation values (δ). From the included graphs, it can be observed that the proposed method has better performance than the classical one, which is consistent with our theoretical analysis.

8.5 Ontology Learning

After the LDA-based method was evaluated in section 8.4, in this section, the ontology learning (hierarchy construction) approach is going to be assessed for its performance of setting up the hierarchies upon the acquired topics. The main concern during this assessment is to verify whether the proposed method manages to construct the hierarchical

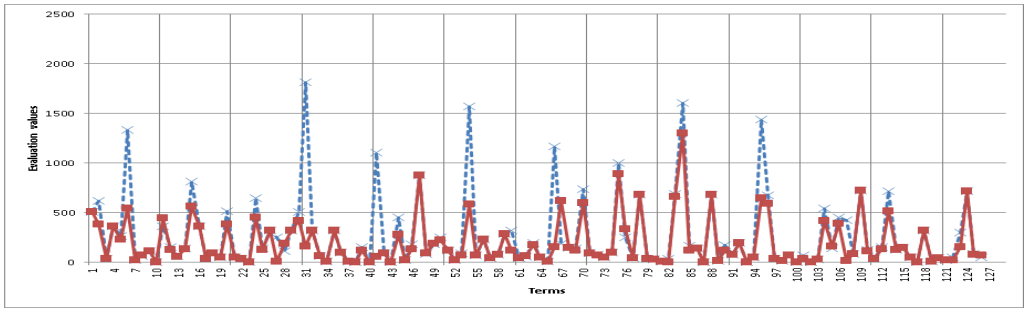

Figure 8.10: Robank 2010

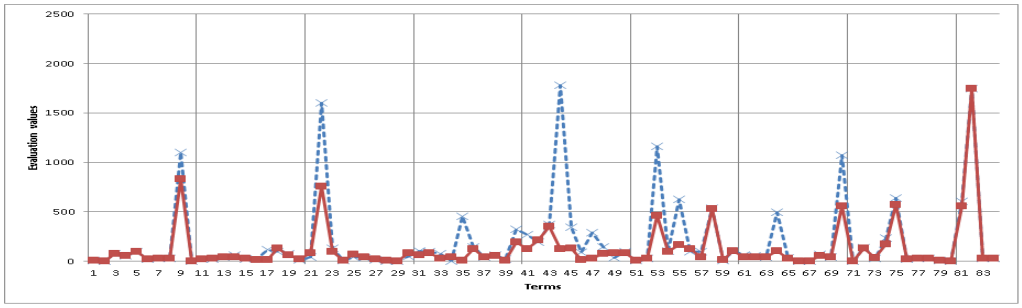

Figure 8.11: UBS 2010

representation to form a new ontology. In order the achieve this objective, the following evaluation process is proposed.

8.5.1 Walk-through

To begin with, a series of terms are supposed to be retrieved from the primary documents. These terms come from the topic analysis described chapter 4. The acquisition process has already been evaluated in section 8.4. Therefore it can be taken for granted that these terms are properly selected. They are from several groups of topics, reflecting some correlated ideas, but not synonyms. For example, assume that in a walk-through the following words have been acquired:

sustainability, energy, automobile, pollution, enterprise

The next critical issue is to set the *target words*. As stated in chapter 5, our proposed method adopts a set of target words as the "*coordinate*" to measure the semantic distances. Here a simple but reasonable way is to count the most frequent words (not including the words in the stop-list) as the target words. In the tested documents, the three words are:

environment, customer, investment

Next, the semantic distances among the selected terms and the target words are calculated as shown in table 8.2. Then the ontology learning algorithm as stated in chapter 5 is im-

plemented in Java. After the entire algorithm is executed, the following result is obtained (see figure 8.12).

The *Cobweb* algorithm [127] was used as benchmark for our proposed method. The same data set is then be used with Cobweb to generate a cluster-based hierarchy as shown in figure 8.13. Then a comparison of the learned trees is going to be carried out. At this phase, two types of evaluation are involved - observation phase and performance evaluation phase. The former one includes the analytical comments from the point of view of human experts whilst the latter one uses systematic ways to evaluate these results.

While conducting the observation phase, several aspects have to be taken into account. First, the structures of the results is analysed. As the principal objective of ontology learning is to set up the hierarchical structures, the parameter *structure* is the most decisive factor. When comparing the results of our proposed hierarchical construction algorithm (HCA) and Cobweb, it is not difficult to observe that Cobweb has a much flatter structure. Most of the nodes connect directly to the root node. Hence these nodes were grouped as a single cluster without substantial correlations with other nodes except for the root. This means that less information is provided by the learned results compared to HCA. In HCA, on the contrary, in-depth hierarchies are inferred to reveal the internal connections between the different topics. Compared with Cobweb, HCA has a first advantage that the hierarchies it sets up provide more illustrative information regarding the structural levels of the topics.

The second advantage of HCA over Cobweb also is related to the established structures. If HCA and Cobweb are reviewed in more detail, it can be seen that HCA is able to set up the structures in form of a *graph* whilst Cobweb can only builds a *tree* structure. This means that HCA offers more generic capabilities for hierarchy construction than Cobweb. In many cases, trees are not sufficient to express the semantic links between

Table 8.2: Semantic Distance Table

sustainability	environment	0.06803406031903386
sustainability	customer	0.05034590489827559
energy	environment	0.09803573471125152
energy	customer	0.07789730162809413
automobile	environment	0.07645240384707779
automobile	customer	0.08053489580367604
...	...	...
government	sustainability	0.05434083824775188
government	energy	0.14587787356743462
government	pollution	0.07493264953709526
motorcar	sustainability	0.04924856115437957
motorcar	energy	0.07530127156148918
motorcar	automobile	4.8317970238510275E9

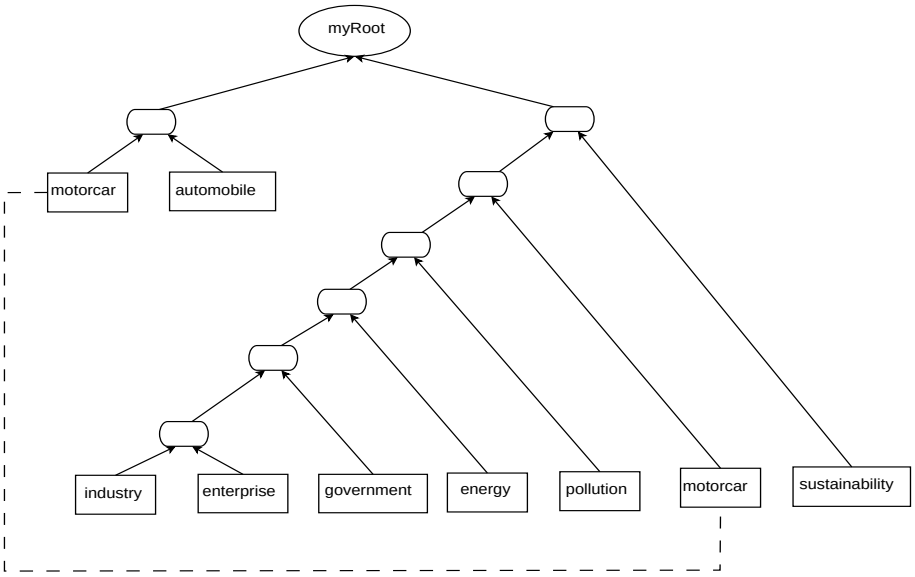

Figure 8.12: The Tree Produced by HCA

topics, because very often it is insufficient for a node to be simply linked to its parental node. It might well be connected to nodes from another branch of the tree. A graph provides a flexible structure to reflect the semantic relations of a set of topics. HCA thus is better suited for structure establishment.

Another strength of HCA is the in-depth relations among topics it can infer. This point is less evident than the previous two features, but worthwhile to be discussed. In Cobweb, when two topics (denoted as A and B) are assigned to a cluster (C_a), another topic (D) is needed as the parent of A and B, and also as the semantic representation of C_a. D has to be one of the candidate topics. HCA, instead, has two potential way to carry out this step:

- A word is taken as the parent of A and B. This word, as the representative of cluster C_a, can be either a new word or one of words A and B.

- A new cluster (C_b) is generated containing C_a and D. Similar to the previous step, a new word or one of the words A, B or D can be referred to as the representative topic of C_b.

The mechanism applied by HCA to set up the hierarchies offers more flexibility - by incorporating new words or reorganizing the candidate topics in the established hierarchies - than the one applied by Cobweb, which only allows topics as the nodes. In this sense, HCA outperforms Cobweb.

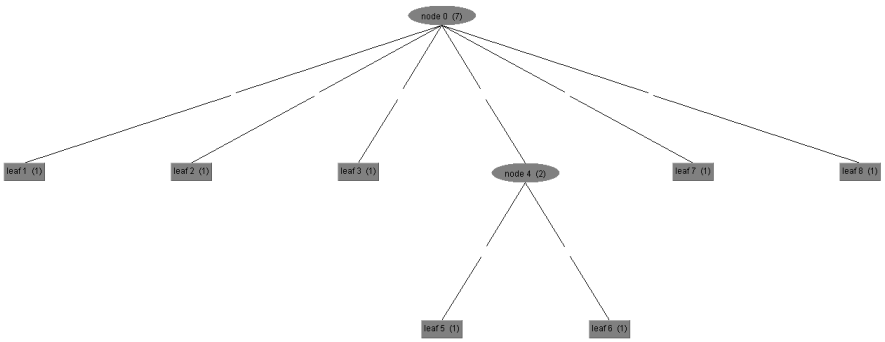

Figure 8.13: Hierarchy Produced by Cobweb

8.5.2 Performance Evaluation

The next test to conduct is the performance evaluation phase, which takes some quantitative measures of performance to compare the two algorithms. In order to illustrate more clearly the comparison between these two approaches, the following steps are carried out:

1. *Candidate topic selection*. In this step, a list of topics is picked out as the topics. The selection of these topics is critical because they influence the final performance as well as the returned values of the experiment. The topics should cover the main themes of the documents incorporated in the project. This is important as the acquired ontological knowledge has to be useful for the analytical work of domain experts. To achieve these goals, a LDA-based topic discovery process is first conducted, then the obtained topics with the highest frequency are used for the hierarchy construction. These topics can be considered as multi-dimensional descriptors of the documents.

   As the result of this step, a list of words is produced and stored in a textual file for each document. The programs serving for experimental purposes will then read these textual files to set up the semantic hierarchies for evaluation. In order to achieve the comparison, Cobweb and HCA will be applied to each word list. Then the acquired results of these two methods will be used as input for the evaluation framework to assess their performance.

2. *Path distance definition.* Furthermore, another important issue is to define a measure for assessing the performance of the two approaches. As a fundamental tool, *path distance* will be used to define the closeness of two given topics in the established hierarchies. The basic principle of the path distance is to capture the routines of two words in the hierarchical structures after the semantic construction. It is easy to understand that the closer two words are, the more semantically relevant they are. Correspondingly, the following definitions are proposed:

   Definition (*One-distance pair*) For two adjacent nodes (N_a and N_b) linked with

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
1	government	12	12	13	22	24	24	29	31	33	34	38	42	43	43
2	organization	12	13	14	25	28	28	40	42	42	49	51	64	93	101
3	history	9	12	13	21	23	27	34	39	41	42	49	58	64	73
4	business	7	11	17	27	31	32	36	42	43	47	51	63	73	75
5	training	7	13	13	16	20	20	22	23	23	25	28	35	40	41
6	material	7	12	12	14	15	15	16	16	16	16	16	21	22	23
7	monitor	7	8	8	11	11	12	13	14	14	14	15	30	37	42
8	plan	7	10	10	12	16	17	21	25	25	25	27	31	31	34
9	customer	6	11	11	16	19	22	26	28	28	53	64	79	86	113
10	employee	6	8	9	15	15	16	27	29	29	45	46	62	62	80
11	community	5	8	8	10	13	14	16	16	17	17	17	17	17	17
12	management	5	6	6	9	10	13	13	14	15	21	22	28	28	34
13	industry	4	4	4	6	6	6	6	6	6	6	6	6	13	13
14	energy	4	6	7	7	7	8	9	9	10	10	13	27	27	28
15	infrastructure	4	8	9	9	12	14	15	17	18	21	21	30	39	39
16	integration	4	9	9	15	20	21	24	29	31	58	69	81	94	113
17	Leadership	4	4	4	6	6	6	7	7	7	22	26	35	41	54
18	Coordination	2	2	2	2	2	2	2	2	2	2	2	7	7	7
19	relationship	2	6	9	14	18	20	25	29	32	36	38	41	42	48
20	issue	2	2	2	3	4	4	5	5	5	5	5	13	13	13
21	involvement	2	2	2	3	3	5	11	11	11	11	14	18	18	20
22	Planification	2	2	2	11	13	17	21	25	26	27	28	38	38	45
23	product	2	2	2	2	2	2	2	2	2	4	7	8	10	12

Figure 8.14: Code Frequencies

a direct edge, their path distance is defined as 1, denoted as $L_{ab} = 1$. These two nodes are called a *one-distance pair*.

Definition (*Multiple-distance pair*) For two linked nodes (N_c and N_d), their path distance is equal to the length (i.e. the number of edges) of the shortest path between them, denoted as $L_{cd} = |N_c, N_d|$. If $L_{cd} > 1$, then these two nodes are called a multiple-distance pair.

Definition (*Merge point*) A node (N_e) which is the ancestor of two other nodes (N_f and N_g) is called their *merge point*. This node N_e is significant for the calculus of the path distance between N_f and N_g.

Definition (*Real merged nodes*) If the merge point of two nodes is not the root of a hierarchy, then these two nodes are called *real merged nodes*.

Definition (*Rooted merged nodes*) If the merge point of two nodes is the root of a hierarchy, then these two nodes are called *rooted merged nodes*.

Based on these definitions, a *path iteration approach* is proposed to calculate the path between two topics. The simplified idea of this approach is to find the merged node of two words and then count the distances between these two words and the merged node.

3. *Information benefit function*. Taking the advantage of the path distance and its affiliated definitions outlined above, the next step is to formulate a standard function as the criterion to evaluate whether the performance of the proposed method is better than the existing classical method. The principle idea is to compare:

   - for a pair of words m and n, the difference between the path distance in the established hierarchies (L) and the semantic distance from WordNet (S).

- compare the lengths produced by the two algorithms (CobWeb and HCA)

The exact implementation is the following:

$$P_z(m,n) = \frac{1}{|\alpha^{-1}L_{mn} - \beta^{-1}S_{mn}|} \tag{8.3}$$

$$F_{xy}(\varrho) = count_{\varrho}(P_x(m,n) > P_y(m,n)) \tag{8.4}$$

In formula 8.3, the performance of the two nodes (m and n) for method z is evaluated in respect whether they are assigned to a reasonable positions in the established hierarchy. The idea of this formula is to compare the path distance in the hierarchy with the semantic distance based on WordNet. In this formula, when two nodes are semantically related and their path distance is consistently short, P_{mn} will return a large value, and vice versa. So the larger a P_{mn} value is, the better the performance.

A critical issue in the formula 8.3 are the two parameters involved, α and β, which are two adaptive coefficients. As the scales of of the two types of distances (L and S) are different, normalization is necessary, and the parameters have to be adapted accordingly. After comparing several possibilities, the following solution is retained:

(a) calculate all the path distances in the hierarchy

(b) sort the values from step 3a (the previous step)

(c) take the largest value as α

(d) do the same for β

In formula 8.4, the complete list of word pairs is denoted as ϱ. The number of all word pairs (m, n) from ϱ for which $P_x(m,n) > P_y(m,n)$ (i.e. the performance of the method x is greater than the performance of the method y) is established, in order to assess, for a given document, whether the hierarchy produced by x is more reasonably than the one produced by y. We call this measure *information benefit function*. For two methods, the one with the larger F value can be considered as the better solution. Furthermore, we have the following property

$$F_{xy}(\varrho) + F_{yx}(\varrho) = count(\varrho) \tag{8.5}$$

i.e., the sum of the information benefit values of two methods is equal to the number of word pairs in ϱ.

4. *Performance evaluation.* Once the evaluation method is constructed, a final task has to be achieved. In order to do this, the following steps have to be carried out:

(a) Get the annotations of all the documents from all the participants involved in the project. In our test project, there are seven groups in total, each group with three or four participants annotating three documents. The overall number finally reaches 72 *document instances* [3].

(b) For each document instance:

   i. List the top 10 topics which show up most frequently.

   ii. Input these topics into the Hierarchy Construction Algorithm (HCA) to produce a hierarchy H_a.

   iii. Input these topics into Cobweb to produce a hierarchy named H_b.

   iv. Input H_a and H_b into the information benefit function and compare the results.

(c) Compare all the documents involved in the project to assess the performance of these methods.

The results of the complete process can be found in figure 8.15. The blue dotted curve stands for the results of HCA whilst the red solid one represents the Cobweb results. In most cases depicted by the curves, the information benefit values from HCA are greater than those from Cobweb. With this data, it can be concluded that HCA performs better compared with Cobweb based on our evaluation approach. Furthermore, the relation 8.5 is satisfied by the values of these two curves on each document instance.

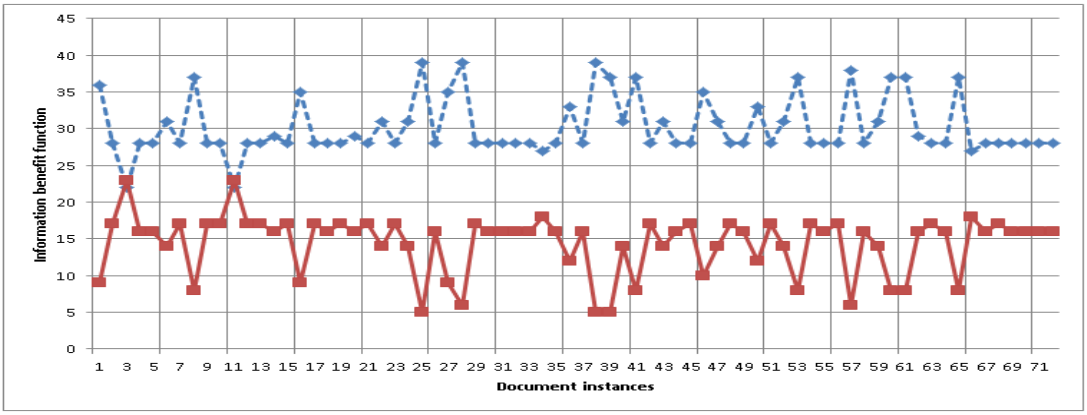

Figure 8.15: HCA and Cobweb Comparison

8.6 An Exemplary Demonstration

At the end of this chapter, also the end of this thesis, we would like to demonstrate a simple but typical example showing how ontology inference can help to produce useful

[3] A "document instance" means a participant coding a document. So one document coded by two participants will be considered as two different document instances.

Company	Year
CREDIT SUISSE	2009
CREDIT SUISSE	2010
CREDIT SUISSE	2011
USB AG	2009
USB AG	2010
USB AG	2011
ZURCHER KANTONALBANK	2009
ZURCHER KANTONALBANK	2010
ZURCHER KANTONALBANK	2011
RABOBANK GROUP	2009
RABOBANK GROUP	2010
RABOBANK GROUP	2011
BANCO SANTANDER SA	2006
BANCO SANTANDER SA	2007
BANCO SANTANDER SA	2008
DEUTSCHE BANK	2009
DEUTSCHE BANK	2010
DEUTSCHE BANK	2011
EIB	2009
EIB	2010
EIB	2011

Figure 8.16: Companies and Years

results, i.e. how analytical conclusions should be obtained from the inferred knowledge structures.

The original idea of this demonstration comes from real cases namely green services. For the companies analysed in our project, a series of indicators are used to assess their performance. For example, the indicator "*paper consumed*" is one of the most important factor for sustainability. A number of actions or operations the companies apply are influencing the paper consumption. Intuitively we should be able to capture the description of some of them, for example "*questionnaires*", "*printing*", etc. These words are evidently linked with paper consumption issues. This is clear even without conducting systematic investigation. The experts involved in the project have practical expectations to find some implicit links between paper consumption and other indicators which might be less evident. For instance, hypothetically, if an enterprise sets up a new data center to digitize all/most of the data, then its paper consumption will possibly decrease to a certain degree. On the other hand, if this company organizes several employee training, workshops or other similar activities, the amount of paper consumption might be increasing, or, if these activities are conducted in a sustainable way, they might stay stable. Therefore it is necessary to study the correlations between the various activities of the companies and their performance in respect to sustainability issues.

To demonstrate the operations in our system the following steps are executed. First, a table is created to store the inventory of all primary documents available. Each row contains the company which produced the report as well as the year of the report (see figure 8.16). This table can be extend as new documents are integrated.

During the next step, a table of codes' frequency is established (see figure 8.17). This table inherits the design of "*relational tables*" as described in chapter 3. It mainly aims to aggregate the coding information of all involved documents. From the frequencies of codes, information about the companies' activities over multiple dimensions are inferred.

Likewise, another table, which records the data for the performance indicators, was

set up (see figure 8.18). These indicators are the standards ones, published by Global Reporting Initiative, for describing different strategies of companies to evaluate their activities regarding sustainability. Following these indicators, the experts summarized the performance with quantified values.

Subsequently, another table for connecting codes and performance indicators are set up based on the opinions of the domain experts. This table mainly shows for each indicator which codes are related to it. This allows interpreting performance indicators as categories abstracted from the existing codes. The content of the table is illustrated in figure 8.19.

Then a piece of SQL scripts was developed to produce an intermediary table. This illustrates the possibility for further treatment using external tools based on the obtained results. This table, shown in figure 8.20, aims to join all the tables mentioned above for the purpose of report produced in the final step.

Finally a report is going to be generated. The report targets at revealing some implicit factors influential to the sustainability performance of a company. As an illustration, the following steps can be conducted for this objective:

1. Set up two ontology files. The first one comes from the intermediary table (figure 8.20) and the second one is formed based on the code frequency table (figure 8.17).

2. Apply a collaborative filtering algorithm with a set of rules upon these ontologies. In our example the collaborative filtering algorithm is principally conducted based on the code frequency, which means that if two codes have similar frequencies, then they are considered as correlated. Ontology inference is used in this step to provide the connections among different codes and performance indicators.

3. For each performance indicator of each company, find the implicitly relevant codes using the results of step 2 and output a textual document as the report, depicted in figure 8.21.

Code	Company	Year	Frequency
Energy	RABOBANK GROUP	2009	93
Targets or objectives	DEUTSCHE BANK	2009	82
Energy	CREDIT SUISSE	2010	82
Involvement	CREDIT SUISSE	2010	80
Energy	CREDIT SUISSE	2011	79
Targets or objectives	DEUTSCHE BANK	2008	72
Energy	RABOBANK GROUP	2011	66
Emissions	CREDIT SUISSE	2009	63
Energy	DEUTSCHE BANK	2008	63
Involvement	CREDIT SUISSE	2011	62
Energy	RABOBANK GROUP	2010	60
Targets or objectives	BANCO SANTANDER SA	2007	59
Targets or objectives	BANCO SANTANDER SA	2008	59
Emissions	RABOBANK GROUP	2009	59
Energy	CREDIT SUISSE	2009	58
Global	CREDIT SUISSE	2010	57
Emissions	DEUTSCHE BANK	2010	57
Emissions	CREDIT SUISSE	2011	56
Managerial vision	CREDIT SUISSE	2011	54
Historical orientation	DEUTSCHE BANK	2010	54

Figure 8.17: Code Frequency

Company	Year	Performance Indicator	Description	Amount	Unit
CREDIT SUISSE	2010	EN5	Energy saved	718737000	kg/employee
CREDIT SUISSE	2008	EN5	Energy saved	682992000	MWh
CREDIT SUISSE	2009	EN5	Energy saved	682465046	MWh
ZURCHER KANTONALBANK	2010	EN29	members of the workforce	4427085	MWh
ZURCHER KANTONALBANK	2009	EN29	members of the workforce	4284628	MWh
ZURCHER KANTONALBANK	2007	EN29	members of the workforce	4258825	MWh
ZURCHER KANTONALBANK	2008	EN29	members of the workforce	4239323	MWh
CREDIT SUISSE	2010	EN8	Water consumed	1900000	MWh
CREDIT SUISSE	2009	EN8	Water consumed	1567600	MWh
ZURCHER KANTONALBANK	2008	EN1	Paper consumption	1342524	MWh
ZURCHER KANTONALBANK	2010	EN22	Quantity of waste	1130274	MWh
CREDIT SUISSE	2010	EN3	emises energy consumption	718700	kg
CREDIT SUISSE	2008	EN3	emises energy consumption	683000	kg
CREDIT SUISSE	2009	EN3	emises energy consumption	682500	CHF million
CREDIT SUISSE	2007	EN3	emises energy consumption	664500	MWh
ZURCHER KANTONALBANK	2010	EN4	direct energy consumption	650143	MWh
CREDIT SUISSE	2010	EN3	Electricity consumed	605800	kg
CREDIT SUISSE	2008	EN3	Electricity consumed	578210	kg
CREDIT SUISSE	2009	EN3	Electricity consumed	571700	kg
CREDIT SUISSE	2007	EN3	Electricity consumed	559100	kg

Figure 8.18: Performance List

4. Distribute this report to domain experts as the suggestive material for their subsequent analysis. For example, regarding the paper consumption of Credit Suisse in 2009, the code *"volunteer program"* was highlighted. It indicates that this issue may influence the paper consumption for this company, presumably this program consumed plenty of paper [4]. The experts could then go on to study this topic based on the machine-produced opinions.

[4] As mentioned above, the meaningfulness of the demonstration in this section is not to find some conclusions as domain discoveries, but to reveal how the proposed system can be leveraged.

Performance indicator	Code
EN1	Materials consumption
EN1	Internal environmental care
EN1	Processes
EN1	Supplier policies
EN1	Industry symbiosis
EN1	Employee empowerment
EN1	Product Service Stewarship
EN1	Continuous improvement
EN2	Materials consumption
EN2	Internal environmental care
EN2	Processes
EN2	Supplier policies
EN2	Industry symbiosis
EN2	Employee empowerment
EN2	Product Service Stewarship
EN2	Continuous improvement
EN3	Energy
EN3	Internal environmental care
EN3	Processes
EN3	Supplier policies

Figure 8.19: Performance Indicators and Codes

Company	Year	Indicator	Description	Amount	Unit	Code	Frequency
CREDIT SUISSE	2009	EN1	paper consumed	5900	tons	Processes	2
CREDIT SUISSE	2009	EN1	paper consumed	5900	tons	Employee empowermen	15
CREDIT SUISSE	2009	EN1	paper consumed	5900	tons	Internal environmen	12
CREDIT SUISSE	2009	EN16	Green Gas Emissions(Scope 1-3	272680	tons	Emissions	63
CREDIT SUISSE	2009	EN16	Scope1: directly through burn	18060	tons	Emissions	63
CREDIT SUISSE	2009	EN16	Scope2: indirectly from energ	183950	tons	Emissions	63
CREDIT SUISSE	2009	EN16	Green Gas Emissions(Scope 1-3	272680	tons	Targets or objectiv	30
CREDIT SUISSE	2009	EN16	Scope1: directly through burn	18060	tons	Targets or objectiv	30
CREDIT SUISSE	2009	EN16	Scope2: indirectly from energ	183950	tons	Targets or objectiv	30
CREDIT SUISSE	2009	EN16	Green Gas Emissions(Scope 1-3	272680	tons	Processes	2
CREDIT SUISSE	2009	EN29	Business travel	492	Moi km	Waste	1
CREDIT SUISSE	2009	EN29	Business travel	492	Moi km	Materials Consumpti	13
CREDIT SUISSE	2009	EN29	Business travel	492	Moi km	Water	8
CREDIT SUISSE	2009	EN29	Business travel	492	Moi km	Pollution	1
CREDIT SUISSE	2009	EN29	Business travel	492	Moi km	Biodiversity	6
CREDIT SUISSE	2009	EN29	Business travel	492	Moi km	Climate change	34
CREDIT SUISSE	2009	EN29	Business travel	492	Moi km	Emissions	63

Figure 8.20: Intermediary Table

8.7 Conclusion

In order to test the proposed methodology and provide a prototypical platform, a software program has been implemented in Java. It mainly emphasizes the interaction with end users and the semantic processing of ontologies. A number of experiments were performed to validate the efficiency and effectiveness of the implementation. The resulting data reveals that the reported methodology produces satisfactory feedbacks in many aspects.

Company: "CREDIT SUISSE", 2009

(1) EN1

* indicator descriptions:

- name: "paper consumed", amount: 5900, unit: tons

* codes:

- "Employee empowerment", frequency: 15
  - # Socially responsible shareholders frequency: 14
  - # Product service development frequency: 17
  - # Volunteer programs frequency: 14
  - # Long term frequency: 17
  - # Materials Consumption frequency: 13
  - # Not global frequency: 19
- "Internal environmental care", frequency: 12
  - # Environmental reasonabilities of Top management frequency: 10
  - # Materials Consumption frequency: 13
  - # Environmental management systems frequency: 10
  - # Volunteer programs frequency: 14
  - # Environmental management system frequency: 10
  - # Socially responsible shareholders frequency: 14
- "Processes", frequency: 2
  - # Lending frequency: 2
  - # Medium term frequency: 2
  - # Short term frequency: 2
  - # Operations organization frequency: 3
  - # Service or Product Design Interface frequency: 2

Figure 8.21: Report

9 Future Work and Conclusion

9.1 Summary and Contribution

A suite of methodologies has been in detail depicted in the previous chapters. It illustrates an entire approach with the purpose to document the investigation concerning qualitative and complex-structured data. Here we will review and summarize the main contribution of this thesis.

The starting point of this work was a collaborative effort of researchers in the domain of logistics and computer science in order to overcome some of the current limitations of the methodologies and the tools used in qualitative case studies. Qualitative data was the cornerstone for this work. To better understand this type of data, grounded theory - one of the tradition ways to deal with this kind of data - was selected as the guideline for the methodology to be developed. Even though it has been a while that this theory was developed, its applicability on qualitative data from an IT point of view is still underutilized and has allowed to jump-start some of the methodologies proposed in this thesis.

In order to establish a baseline for the proposed methodology, several globally accepted software products were studied for their applicability to qualitative data. The main contributions of the approach presented here, compared to the state of the art tools, are firstly the structural representation of the overall process as well as the specific knowledge with formal semantics and learning capabilities, and secondly an open architecture allowing for inter-operability. The proposed formalization extends grounded theory into a system-implementable way and is considered as the elementary meta-data structure for the entire methodology. Upon these elements, an overall workflow was proposed to allow a smooth integration of the proposed methodology in a general research context but also to guarantee a replicability of the produced results.

From the beginning it was a declared goal that the defined approach had to be imple-

mented in form of an operational system. This constraint added some extra complexity to the achieved work. One of the main challenges consisted in finding a formalization allowing, on one hand, the necessary theoretical work but being, on the other hand, implementable. A thorough analysis of the knowledge typically used in case studies and of the type of formalism needed to express the concepts of grounded theory has led us to select ontologies as the fundamental tool. The ontological elements, in compliance with the OWL specification, were used on one hand to formalize the system and on the other hand as meta-data structure to store all the data, information and knowledge available in the system. Ontologies were used to specify the approach given by grounded theory. At the same time, the output of this approach will again be ontologies, allowing for a closed loop. While the grounded theory mainly describes several steps, we designed ontologies converting these steps into a series of semi-data structures serving as the raster for qualitative data investigation. They can be regarded as the advisable guideline for many similar projects.

The purpose of using these ontologies is not only to maintain knowledge, but also to collect potential new information and data. Therefore we have to extract the necessary information from the ontologies, transform it into appropriate formats, and then load it into different data warehousing models - which defines an ETL process. Instead of using the traditional ETL methods, a novel approach is proposed, based on the scenarios of case studies. This approach fully supports the qualitative characteristics of the input data and attempts to automatize the process of transformation. On top, a series of data warehousing models were established to refine the relational links in the data. These models provide the possibilities to record all involved data even with frequent updates. Even though the techniques of data warehousing are based on existing theories, the solutions - for the data models and transformation methods - encapsulated in our methodology are more generic in many respects.

The data archived in the data warehousing models serves as basis for the extraction of more abstract information. One promising idea is to seek for latent topics reflecting the thematic concepts of the raw data. These topics reveal the main subjects as well as some details of the annotated documentation. Inspired by the LDA approach, a novel model was built up integrating the coding mechanism (based on grounded theory) as a key factor of the generative reasoning process. A couple of improved algorithms were presented to alleviated the practical deficiencies of the traditional model. A number of experiments have shown that the new model outperforms the classical one in many regards. Afterwards, based on the extracted topics, an ontology learning process was introduced. This process takes advantage of WordNet-based semantics to measure the diversities among the vocabularies. Then an innovative algorithm was proposed to organize the extracted topics in form of hierarchies based on their semantic distances. Different from more traditional approaches based on statistical methods, the WordNet approaches allows to take into account the more in-depth correlations among words. An ontology inference system was later designed based on a rule engine. This system adopts the hierarchical knowledge produced in the previous steps and then matches patterns formulated as rules with the ontological facts. The newly-discovered information can then be exported to a variety of

applications in order to assist them in their specific tasks. The overall method depicted above provide the "*learning*" capabilities, defined as an expectation in the first chapter. It is a novel method for ontology processing and construction, with domain participation and analytical assignments.

The proposed methodology relies to a large extend on semantic comprehension of text, so the syntactical representation of the documentation is fundamental. For this reason we decided to consider different natural languages. This consideration was also motivated by the application domain of the project, particularly in the context of globalization. As the basic approach was based on documents in English, it is relatively easy to understand that this approach works for most alphabetical languages. To generalize our methodology we decided to approach the Chinese language. To achieve the extension of our techniques to the Chinese language, a prototypical paradigm was proposed formalizing typical Chinese sentences. Then a set of ontologies were established inheriting the ideas of this paradigm. An algorithm was designed to extract features from the training data. A series of experiments have shown that the proposed method has promising capabilities to handle the sentences in Chinese language and to integrate smoothly with the LDA based topic discovery described earlier. Although the proposed approach is not able to cover all the cases with complex context, its basic principles offer an innovative and extensible perspective for machine processing on this non-alphabetical language.

A prototypical application, named *Qualogier*, was developed as a working platform for field tests. This system implements the functionalities described above, thus providing more powerful tools compared to the state-of-the-art software. A series of criteria were defined to assess the efficiency and effectiveness of the methodology encapsulated in the system. The usability of this prototype was verified by an intensive collaboration during the case study process between the management and the IT side. It was straightforward to show that the system scales very well along all dimensions (size of data, number of users, complexity of inference, etc). In addition, the proposed methodology was assessed regarding its learning capabilities, with satisfactory results. Some insightful discoveries were meanwhile made during these case studies.

9.2 Limitations and Future Work

Although encouraging progress was made as reported in this thesis, several issues still remain to be solved. In the first place, ontology establishment and incorporation is concerned. In this thesis ontologies are set up via manual implementation and system acquisitions. The former approach is mainly associated with domain references and data annotations whilst the latter one refers to new discoveries without *a priori* knowledge. The established ontologies demonstrated their effectiveness in the case studies. Yet, there is still a difficulty in making these two approaches to collaborate. The criteria for which method to adopt in which moment are not clarified yet. Confusion and even contradiction may occur when both types of ontologies are engaged concurrently. An enhanced approach is necessary to coordinate and validate these ontologies. A promising yet simple idea is to extend the upper ontology with a more specific functional component to inter-

connect these ontologies under concrete conditions. This component would follow the *bridge pattern*, a paradigm well known in design patterns. Then the project ontologies will be supplemented with the embedded conditions. In this case, the overall project would be clearly guided when it needs to decide which method to initiate and employ, because the conditions are already formulated at a higher level.

The next limitation concerns the data warehouse models related on the ETL process. Once the model is established, it is expressed in a relational representation, regulated by database theories. However, even if the model complies perfectly with the normalization and cardinality requirements, there is no guaranty that this model is correct in all respects. In our context it is crucial to be able to validate the consistency between interconnected entities at the semantic level. This demand is challenging as there are so far no known theories to support this idea. The validation task is supposed to be designed firmly based on the characteristics of the concrete cases, but on the other hand the specific constraints have to be abstracted into a generic approach. Otherwise the validation is almost impossible to be carried out. For the purpose of solving this problem, a possible solution is to define a rule set on top of the data warehousing models. The advantage of this solution would be that no matter which kinds of entities are involved and how complicated their links are, it is always feasible to operate the validation in a standardized way as long as the rules can be verified.

In addition, limitations exist in the core section - the ontology learning. When the hierarchies of the topics are constructed, two concepts are merged into a single node whose children are these two concepts. There are a number of ways to identify the parental concept. The method used here takes advantages of the semantic distances between these two concepts in respect to a set of target words. This is an acceptable solution but not an optimal one, because sometimes it is not sufficient to express the semantics only with the provided list of words. It is thus of interest to find another word, beyond the scope of the original ones, to identify the newly acquired node. For example, CO_2 and NO_2 lead to the higher level concept *gas*. If this goal can be achieved, it will provide richer information as derived concepts are easier to be associated with the existing knowledge. A possible way to satisfy this need is to develop an algorithm based on the network of WordNet to walk through the different paths among the concepts, analysing the words along these paths to find the most promising generalization. As WordNet integrates a lot of conciliative aspects it might be that concepts found this way are better connected to the pre-existing knowledge.

9.3 Conclusion

Even though a number of methods have been developed for data studies, analyzing complex-structured data is still challenging. With the assistance of ontologies, many of the tasks become easier, by taking advantage of structural knowledge representation and discovery. However, the process of appropriately establishing the necessary ontologies and the reasoning over them is still challenging. Furthermore, the learning capability of machines on ontological knowledge is in high demand from different sides, but has not reached a

satisfactory level.

The goal of this thesis can be summarized as follows: we would like to contribute with ontology-based approaches to improve the methodologies of data modelization and analysis. Considering this objective, a formal model as well as its processing methods have been established based on grounded theory. This model is presented in form of ontologies respecting domain references and analytical annotations. On this basis, ETL processes are conducted to retrieve the ontological knowledge in a standardized manner, for the establishment of a series of data warehousing models as regularized data sources, serving for cross-references queries. Using this input data, a proposition for an improved LDA model was made. This model manages to capture a series of implicit topics in the original text. These topics are then organized in hierarchies to represent the more profound semantic relations among these topics. Moreover, inferences are carried out to produce novel facts based on the constructed hierarchies. The model is further extended to non-alphabetical languages, such as Chinese, to broaden the application scope of the approach to globalization scenarios. In compliance with the theoretical design, a prototype implementing the ontology layers and supplementary functionalities have been developed and used as a working tool for a series of experiments.

To summarize, we have been applying an qualitative methodology originated from information science to enable and facilitate the modelization and analytical process oriented on data investigation.

Appendices

A
Big Data

Research indicates that modern business, instead of merely providing materialized products, depends to a large extend on *"service"* as a non-material type of products [128]. The daily operations of enterprises, no matter simple or intricate, relies on the assistance of software since software itself is already becoming a special type of service [129]. As a large number of marketing wars occur in the competition of information processing, the utilization of software promotes the efficiency of a company as well as its competence. In the domain of computer science, especially software engineering and business intelligence, a suit of enterprise frameworks have been proposed in the past decades for the purpose of facilitating enterprise activities. These framework have already achieved satisfactory progress and have widely been applied in industry. *J2EE*, *.Net*, *grid computing*, and *cloud computing* are some of the typical ones [130][131][132][133]. The last one, developed based on classical mechanisms such as distributed systems as well as network theories and innovative ones - including *"SOA"* and *"virtuliazation"* [133] - has become advisable in numerous domains. For cloud computing, a pivotal issue is how to handle data of large amount and high complexity, namely *"big data"* [134]. Big data is recently becoming a prominent branch of research in computer science.

The concrete features of big data are summarized as "*4-Vs*" [135]. The primary feature is its *volume*. The dramatic growth of information has been a phenomenon for years [136]. For instance, the giant of information technology, Cisco, estimates that the monthly mobile traffic will arrive at a very high level by 2017 [137]. The amount of data is so massive that not only it is difficult for human beings to navigate and substantially utilize, but with the fastest computing capacity at present it is not an easy task. Besides, it is not the ultimate objective only to maintain and go through this information. Instead, the useful elements are supposed to be captured and exploited, making this question more challenging. Due to the volume of the data, additionally, it is tricky to keep it without

consuming too many resources such as facilities, preservation, energy, etc [138]. Optimization approaches will be contributive to the theoretical research and the benefits of enterprises for practical purposes.

The next factor of big data is its *variety*. Traditionally, numeric data depicting financial flows and performance was highly valued [139]. Business analysts were expected to focus on these numbers to investigate the background information. It was later realized that plenty of data without clear structure or with complex structures also contains rich information [140]. As revealed in [141], enterprises produce a large number of daily documents, business transaction records, and e-mails. The contained qualitative factors are considered as the necessary basis of the numeric data to describe procedures [142]. However, there is so far no enough support to investigate this type of data. As the fast development of Internet-based file storage and delivery, the usage of textual data is increasing dramatically, leading to the obvious demands of an innovative approach to handle this matter. Our objective is thereby to study the qualitative data to discover its invisible rules, and to provide applicable knowledge as the feedback to domain experts.

Another concern of big data takes its *velocity* into account, which considers the changes and evolution of the data. As the data sources are extended to different equipment with active interoperability, such as stream, the speed of data input and output is growing promptly [143]. Archetypal approaches are not sufficient any more to solve the emerging problems [144]. For example, commonly there are systems and facilities deployed in supermarkets to record the information pieces of customers, products, and transactions. However, nowadays on-line retail firms receive a large number of orders on a daily base. These on-line firms have been developing a suit of methods based on information techniques to investigate the massive data, which is so far not common in physical stores [145]. This data allows producing conclusions for decision making so as to promote the business in the near future [146]. Another case is *Twitter*, a popular social networking service, which produces an unusually large data volume per day from millions of tweets [147]. These examples imply the significance of modeling the data into an adequate representation in advance [148]. Only if the data model is effectively established beforehand, the real-time demands from the practical scenarios can be fulfilled and the problems of velocity are thus feasible to be handled.

Meanwhile, the *value* of big data is necessary to be considered. For On-Line Analytical Processing (OLAP) applications, although the amount of data is large, only part of the data brings benefits to the ultimate usage - in other words, the value of this data is very sparse [149]. Hence, conventional data mining approaches are not satisfactory to apply without pre-processing procedures [150]. In addition, for big data, some new values are advisable to be created, for example, recommendations to the users [151]. A list of well-known websites such as *Youtube* and *Amazon* offer recommendations of their published items to the users based on the great amount of browsing history [152]. Likewise, social network applications including *Facebook* and *Twitter* take advantage of this functionality and have gained extraordinary success [153]. Consequently, a new methodology is set up in this thesis to investigate the data in order to discover its valuable ingredients and convert them into meaningful knowledge. With this knowledge, many applications required

by end users become easy to design and implement.

B

Formalization of Analytical Data Annotation

A desirable step is to formalize the research objects - the original data and the annotation action - as indispensable elements of our presented work. This phase is crucial in that subsequent evolvement will confront great challenges if these elements are not clarified. Problems such as term confusion and disambiguity will arise. Furthermore, the formalization can limit the research scope in a precise way. Figure B.1 gives a direct idea in the form of bar layers for the concepts to be formalized in this section.

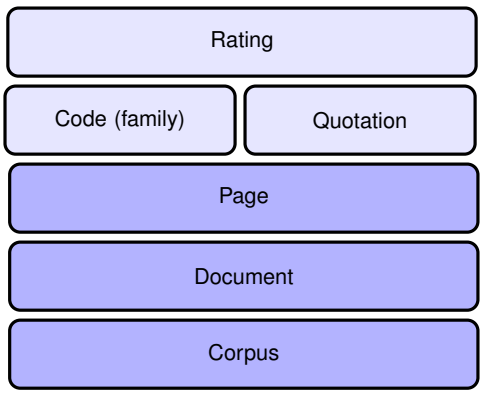

Figure B.1: Formalized Concepts

Corpus A corpus, for instance $Corpus_{\alpha}$, is a collection of archived records [45]. Usually these records concern similar themes while their content can be greatly different. Technically speaking, a corpus consists of a couple of documents (D). Each document is

a textual file and the system will load the document into its working memory for navigation. When new documents emerge, a corpus should be open to integrate these documents into its scope incrementally. The overall system takes a corpus as its initial input and deliver it to the users for the analytical processes.

$$Corpus_{\alpha} = \{D_1, D_2, ..., D_i, .., D_n\} \tag{B.1}$$

Document A document is an element of a corpus and a collection of pages (P). The number of pages varies from very small such as one or two to very large as several hundred. A document main contain different forms of data, including pure text, images, tables, and figures. By default, a document belongs to only one corpus. To continue with the previous example of $Corpus_{\alpha}$:

$$D_i \in Corpus_{\alpha} \tag{B.2}$$

$$D_i = \{P_1, P_2, ..., P_j, ..., P_n\} \tag{B.3}$$

Page A page is an element of a document and a matrix of words (w) with their *coordinates* (cor) on the page. Here two dimensions - X and Y - are used to represent the location of each character. Page as well as its affiliated coordinates are the properties to identify the geographical distributions of the words and sentences.

$$P \longleftarrow \{w, cor\} \tag{B.4}$$

Taking the example of P_j:

$$P_j = \begin{bmatrix} w_1 & cor_{w_1} \\ w_2 & cor_{w_2} \\ \dots & \\ w_n & cor_{w_n} \end{bmatrix} \tag{B.5}$$

and equation B.5 is derived as:

$$P_j = \begin{bmatrix} w_1 & (X_{11}, Y_{11}), ..., (X_{1t_1}, Y_{1t_1}) \\ w_2 & (X_{21}, Y_{21}), ..., (X_{2t_2}, Y_{2t_2}) \\ \dots & \\ w_n & (X_{n1}, Y_{n1}), ..., (X_{nt_n}, Y_{nt_n}) \end{bmatrix} \tag{B.6}$$

The formalization manifested above facilitates the process of ontology learning comprising two aspects: *data annotation* and *systematic analysis*. Data annotation is at the beginning carried out by the users as interaction with the primary documents. They operate on the PDF files with actions such as drag-and-drop, highlighting, coding, and rating. This will provide refined data on top of the original text. The data generated from their actions is cached into ontologies. Furthermore a couple of algorithms are applied for knowledge learning and acquisition.

Quotation A quotation is a collection of selected words [125][26]. These words are not necessarily continuous. They indicate some useful text sections for the users from the primary documents and are highlighted compared with other text. Quotation is the link between the original data and the users' insights. Q_β is an example as a quotation:

$$Q_\beta = \{w_1, w_2, ..., w_n\} \tag{B.7}$$

The description of this quotation can be derived based on formula (B.4) to the following matrix. Here t represents some additional information such as users, update time, etc.

$$Q_\beta = \begin{bmatrix} w_1 & P_{w_1} & cor_{w_1} & t_{w_1} \\ w_2 & P_{w_2} & cor_{w_2} & t_{w_2} \\ \dots & & & \\ w_n & P_{w_n} & cor_{w_n} & t_{w_n} \end{bmatrix} \tag{B.8}$$

Code A code is a key word assigned to a quotation to label its properties in different dimensions [46][26]. The usage of code is consistent with the main idea of grounded theory in which codes function as a decisive role. As there are several modes of coding defined in grounded theory, the principles of this theory are inherited by supporting all of these modes in our proposed methodology. In formula B.9, $\overline{C}$ is the set which contains all the codes attributed to quotation Q. The coding process, denoted as $\hat{C}$, is essentially a reflection combined by Q and $\overline{C}$

$$\hat{C} \longleftarrow \{Q, \overline{C}\} \tag{B.9}$$

and can further be illustrated with a concrete example:

$$\hat{C_\eta} = \begin{bmatrix} Q_1 & c_{11} & c_{12} & \dots & c_{1m_1} \\ Q_2 & c_{21} & c_{22} & \dots & c_{2m_2} \\ \dots & & & & \\ Q_n & c_{n1} & c_{n2} & \dots & c_{nm_n} \end{bmatrix} \tag{B.10}$$

Code family When some single codes refer to the same subjects, they are designated as a code family due to their semantic interconnections [26]. Code families are leveraged primarily for ontology inference with some rules set in advance. A general rule for a code family will be applied to all the codes in this family. Code families can be designed partially beforehand and gathered dynamically with clustering and ontology learning techniques while the annotation is being carried out.

$$CF_\delta = \{c_1, ...c_m\} \tag{B.11}$$

Rating Ratings are scores assigned to the elements in order to evaluate their importance and relevance to the main themes. Ratings are applied in addition to evaluate the contribution of the users by giving scores to their quotations and codes. The usage of ratings is very flexible regarding their maximum and minimum values. Users should be

able, with the support of the system, to personalize these values according to the concrete needs. Some *aggregation* functions, for example *average*, will be provided for the sake of output and visualization.

$$\hat{R} \longleftarrow \{Q, \overline{R}\} \tag{B.12}$$

or

$$\hat{R} \longleftarrow \{C, \overline{R}\} \tag{B.13}$$

$\overline{R}$ above standards for the set containing all the rating values over given quotations/codes, and $\{Q, \overline{R}\}$ can be characterized in more detail as:

$$\hat{R_\rho} = \begin{bmatrix} Q_1 & r_{11} & r_{12} & r_{1m_1} \\ Q_2 & r_{21} & r_{22} & r_{1m_2} \\ \dots & & & \\ Q_n & r_{n1} & r_{n2} & r_{nm_n} \end{bmatrix} \tag{B.14}$$

Bibliography

[1] N. Wirth, *Algorithms + Data Structures = Programs*. NJ: Prentice-Hall, 1976.

[2] "Global reporting initiative." Online, retrieved September 2013. Available: `https://www.globalreporting.org/Pages/default.aspx`.

[3] L. I. Millett and D. L. Estrin, *Computing Research for Sustainability*. D.C.: The National Academies Press, 2012.

[4] V. L. Lemieux, "Envisioning a sustainable future for archives: A role for visual analytics," in *the International Council on Archives Congress*, pp. 1–10, 2012.

[5] "Enterprise-level indicators for resource productivity and pollution intensity," tech. rep., United Nations Industrial Development Organization, 2010.

[6] I. S. Dhillon and D. S. Modha, "Concept decompositions for large sparse text data using clustering," *Journal Machine Learning*, vol. 42, pp. 143–175, 2001.

[7] "Selecting the right project management methodology," tech. rep., KLR Consulting Team, 2006.

[8] M. Basson, "Evolution of business analysis in standard bank PBB," in *Keynotes of Business Analysis Summit Africa*, 2012.

[9] "The Sigma guidelines - tookit - Sigma guide to sustainability issues." Online, retrieved September 2013. Available: `http://www.projectsigma.co.uk/Toolkit/SustainabilityIssuesGuide.pdf`.

[10] K. Lyons, "2012 EL program: Sustainable manufacturing." Online, 2011 (retrieved September 2013). Available: `http://www.nist.gov/el/msid/lifecycle/upload/SMprogram2012.pdf`.

[11] Wikipedia, "Quality function deployment — Wikipedia, the free encyclopedia." Online, retrieved September 2013. Available: `http://en.wikipedia.org/wiki/Quality_function_deployment`.

[12] T. Kivinen, "Applying QFD to improve the requirements and project management in small-scale project," Master's thesis, University of Tampere, 2008.

[13] E. J. Marchiori, A. Serrano, A. del Blanco, I. Martinez-Ortiz, and B. Fernandez-Manjon, "Integrating domain experts in educational game authoring," in *Proceedings of the Fourth IEEE International Conference on Digital Game and Intelligent Toy Enhanced Learning*, pp. 72–76, 2012.

[14] "Transforming BI: Subway IPC turns to balanced insight consensus." Online, retrieved September 2013. Available: `http://www.balancedinsight.com/wp-content/uploads/2011/06/Balanced-Insight-Inc-IPC-Subway-Case-Study.pdf`.

[15] B. Jalender, A. Govardhan, and P. Permchand, "A pragmatic approach to software reuse," *Journal of Theoretical and Applied Information Technology*, vol. 3, pp. 87–96, 2010.

[16] B. Glaser and A. Strauss, *The Discovery of Grounded Theory: Strategies for Qualitative Research*. NJ: Aldine Transaction, 1967.

[17] N. Mavetera and J. H. Kroeze, "Practical considerations in grounded theory research," *Sprouts: Working Papers on Information Systems*, vol. 9, pp. 1–23, 2009.

[18] J. Saldana, *The Coding Manual for Qualitative Researchers*. CA: Sage Publications Ltd, 2009.

[19] A. L. Strauss, *Qualitative Analysis for Social Scientists*. UK: Cambridge University Press, 1987.

[20] S. H. Khandkar, "Open coding," tech. rep., University of Calgary.

[21] N. K. Rohatinsky, *Committing to Mentorship: Nurse Managers Perceptions of their Roles in Creating Mentoring Cultures*. PhD thesis, University of Saskatchewan, 2012 (retrieved September 2013).

[22] M. T. T. Thai, L. C. Chong, and N. M. Agrawal, "Straussian grounded-theory method: An illustration," *The Qualitative Report*, vol. 17, pp. 1–55, 2012.

[23] B. G. Glaser, "Conceptualization: On theory and theorizing using grounded theory," *International Journal of Qualitative Methods*, vol. 1, pp. 1–31, 2002.

[24] S. Seidel and J. Recker, "Using grounded theory for studying business process management phenomena," in *Proceedings of the 17th European Conference on Information Systems*, pp. 490–501, 2009.

[25] A. Boyd, J. Pucciarelli, and M. Webster, "Organizational organizational blind spot: Blind spot: The role of document-driven business processes in driving top-line growth," tech. rep., IDC, 2012.

[26] "ATLAS.ti: The qualitative data analysis & research." Online, retrieved September 2013. Available: `http://www.atlasti.com/index.html`.

[27] "QSR international." Online, retrieved September 2013. Available: http://www.qsrinternational.com/.

[28] Microsoft, "Email and calendar software Microsoft Outlook." Online, retrieved September 2013. Available: http://office.microsoft.com/en-us/outlook/.

[29] A. S. Incorporated, "Adobe reader." Online, retrieved September 2013. Available: http://get.adobe.com/reader/.

[30] D. Romano, "Data mining leading edge: Insurance & banking," in *Proceedings of Knowledge Discovery and Data Mining*, 1997.

[31] D. Han and K. Stoffel, "Ontology based model and procedure creation for topic analysis in Chinese language," in *Proceeding of Joint Conference of the Sixth Chinese Semantic Web Symposium (CSWS) and the First Chinese Web Science Conference (CWSC)*, pp. 67–73, Springer Publishing Company, 2012.

[32] D. Han and K. Stoffel, "Ontology based qualitative case studies for sustainability research," in *Proceedings of Workshop "A.I. for Intelligent Planet", IJCAI 2011*, ACM, DOI: 10.1145/2018316.2018322, 2011.

[33] XBRL, "An international standard language for communicating business data simply and quickly." Online, retrieved September 2013. Available: http://www.xbrl.org/.

[34] S. M. Nunez, J. Emilio, L. Gayo, J. D. A. Suarez, and P. O. D. Pablos, "Analysis of XBRL documents representing financial statements using semantic web technologies." Online, retrieved September 2013. Available: http://di002.edv.uniovi.es/~labra/FTP/Papers/mtsr07LabraXBRL.pdf.

[35] UNEP, "UNEP environmental data explorer - the environmental database." Online, retrieved September 2013. Available: http://geodata.grid.unep.ch/.

[36] Wikipedia, "Facade pattern — Wikipedia, the free encyclopedia." Online, retrieved September 2013. Available: http://en.wikipedia.org/wiki/Facade_pattern.

[37] P. Vassiliadi and A. Simitsis, "Extraction, Transformation, and Loading," *Encyclopedia of Database Systems 2009*, pp. 1095–1101, 2009.

[38] P. Vassiliadis, "A Survey of ExtractTransformLoad Technology," *International Journal of Data Warehousing and Mining*, vol. 5, pp. 1–27, 2009.

[39] M. R. D. Giusti, N. F. Oviedo, and A. J. Lira, "Extract, transform and load architecture for metadata collection." Online, retrieved September 2013. Available: http://www.istec.org/wp-content/uploads/2011/10/ISTEC-GA-XVIII-Extract-Transform-and-Load-Architecture-for-Metadata-Collection.pdf.

[40] Wikipedia, "Table (database) — Wikipedia, the free encyclopedia." Online, retrieved September 2013. Available: http://en.wikipedia.org/wiki/Table_(database).

[41] Wikipedia, "Pivot table — Wikipedia, the free encyclopedia." Online, retrieved September 2013. Available: http://en.wikipedia.org/wiki/Pivot_table.

[42] D. Calvanese, G. De Giacomo, M. Lenzerini, D. Nardi, and R. Rosati, "Data integration in data warehousing," *International Journal of Cooperative Information Systems*, vol. 10, pp. 237–271, 2001.

[43] B. Husemann, J. Lechtenborger, and G. Vossen, "Conceptual data warehouse design," in *Proceedings of the International Workshop on Design and Management of Data Warehouses*, pp. 1–11, 2000.

[44] C. Imhoff, N. Galemmo, and J. G. Geiger, *Mastering Data Warehouse Design: Relational and Dimensional Techniques*. NJ: Wiley, 2003.

[45] Wikipedia, "Text corpus — Wikipedia, the free encyclopedia." Online, retrieved September 2013. Available: http://en.wikipedia.org/wiki/Text_corpus.

[46] Wikipedia, "Code — Wikipedia, the free encyclopedia." Online, retrieved September 2013. Available: http://en.wikipedia.org/wiki/Code.

[47] Wikipedia, "Bernoulli trial — Wikipedia, the free encyclopedia." Online, retrieved September 2013. Available: http://en.wikipedia.org/wiki/Bernoulli_trial.

[48] Wikipedia, "Bernoulli distribution — Wikipedia, the free encyclopedia." Online, retrieved September 2013. Available: http://en.wikipedia.org/wiki/Bernoulli_distribution.

[49] Wikipedia, "Binomial distribution — Wikipedia, the free encyclopedia." Online, retrieved October 2013. Available: http://en.wikipedia.org/wiki/Binomial_distribution.

[50] Wikipedia, "Multinomial distribution — Wikipedia, the free encyclopedia." Online, retrieved October 2013. Available: http://en.wikipedia.org/wiki/Multinomial_distribution.

[51] Wikipedia, "Beta distribution — Wikipedia, the free encyclopedia." Online, retrieved October 2013. Available: http://en.wikipedia.org/wiki/Beta_distribution.

[52] Wikipedia, "Gamma function — Wikipedia, the free encyclopedia." Online, retrieved October 2013. Available: http://en.wikipedia.org/wiki/Gamma_function.

[53] Wikipedia, "Dirichlet distribution — Wikipedia, the free encyclopedia." Online, retrieved October 2013. Available: http://en.wikipedia.org/wiki/Dirichlet_distribution.

[54] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *Journal of Machine Learning Research*, vol. 3, pp. 993 – 1022, 2003.

[55] M. Burton, "The joy of topic modeling." Online, retrieved October 2013. Available: http://mcburton.net/blog/joy-of-tm/.

[56] "Biostatistics for the clinician," tech. rep., University of Texas-Houston Health Science Center, 1997.

[57] D. M. Blei, "Introduction to probabilistic topic models," *Communications of the ACM*, vol. 55, pp. 77–84, 2011.

[58] M. Porter, "The porter stemming algorithm." Online, 2006 (retrieved October 2013). Available: http://tartarus.org/martin/PorterStemmer/.

[59] I. Smirnov, "Overview of stemming algorithms," tech. rep., DePaul University, 2008.

[60] S. Houthuys, "The differences between flemish pupils acquiring English and walloon pupils acquiring English before the input of formal instruction," Master's thesis, University of Ghent, 2011.

[61] "Schultimes czyli denglish w zstu," tech. rep., ZSTU, 2012.

[62] D. Ramage, D. Hall, R. Nallapati, and C. D. Manning, "Labeled LDA: A supervised topic model for credit attribution in multi-labeled corpora," in *Proceedings of Conference on Empirical Methods in Natural Language Processing*, pp. 248–256, 2009.

[63] M. Rosen-Zvi, T. Griffiths, M. Steyvers, and P. Smyth, "The author-topic model for authors and documents," in *Proceedings of the 20th conference on Uncertainty in artificial intelligence*, pp. 487 – 494, 2004.

[64] J. Tang, H. Leung, Q. Luo, D. Chen, and J. Gong, "Towards ontology learning from folksonomies," in *Proceedings of the 21st international jont conference on Artifical intelligence*, pp. 2089 – 2094, Morgan Kaufmann Publishers Inc., 2009.

[65] X. Wei and B. Croft, "LDA-based document models for ad-hoc retrieval," in *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*, pp. 178 – 185, 2006.

[66] S. K. Lukins, N. A. Kraft, and L. H. Etzkorn, "Source code retrieval for bug localization using latent dirichlet allocation," in *Proceeding of the 15th Working Conference on Reverse Engineering*, pp. 155 – 164, 2008.

[67] T. L. Griffiths and M. Steyvers, "Finding scientific topics," in *Proceedings of the National Academy of Science*, pp. 5228 – 5235, 2004.

[68] B. Omelayenko, "Learning of ontologies for the web: the analysis of existent approaches," in *Proceedings of the International Workshop on Web Dynamics*, 2001.

[69] Z. Li, M. C. Yang, and K. Ramani, "A methodology for engineering ontology acquisition and validation," *Artificial Intelligence for Engineering Design, Analysis and Manufacturing*, vol. 23, pp. 37–51, 2009.

[70] A. Maedche and S. Staab, "Ontology learning for the semantic web," *Journal of IEEE Intelligent Systems*, vol. 16, pp. 72–79, 2001.

[71] A. Lenci, S. Montemagni, V. Pirrelli, and G. Venturi, "NLP-based ontology learning from legal texts. a case study," in *CEUR Workshop Proceedings*, pp. 113–129, 2007.

[72] M. Hazman, S. R. El-Beltagy, and A. Rafea, "Ontology learning from domain specific web documents," *International Journal of Metadata, Semantics and Ontologies*, vol. 4, pp. 24–33, 2008.

[73] L. Karoui, M.-A. Aufaure, and N. Bennacer, "Context-based hierarchical clustering for the ontology learning," in *Proceedings of the 2006 IEEE/WIC/ACM International Conference on Web Intelligence*, pp. 420–427, 2006.

[74] C. Biemann, "Ontology learning from text: A survey of methods," *LDV Forum*, vol. 20, pp. 75–93, 2005.

[75] M. Li, G. Holmes, and B. Pfahringer, "Clustering large datasets using Cobweb and K-Means in Tandem," *Advances in Artificial Intelligence*, vol. 3339, pp. 368–379, 2005.

[76] R. Mandala, T. Takenobu, and T. Hozumi, "The use of WordNet in information retrieval," in *Proceedings of Usage of WordNet in Natural Language Processing Systems*, pp. 31–37, 1998.

[77] Wikipedia, "WordNet — Wikipedia, the free encyclopedia." Online, retrieved October 2013. Available: `http://en.wikipedia.org/wiki/WordNet`.

[78] Princeton University, "WordNet 3.0 database statistics." Online, retrieved October 2013. Available: `http://wordnet.princeton.edu/wordnet/man/wnstats.7WN.html`.

[79] G. A. Miller, "Nouns in WordNet: A lexical inheritance system," *International Journal of Lexicography*, vol. 3, pp. 245–264, 1990.

[80] M. Pasca and S. Harabagiu, "The informative role of WordNet in open-domain question answering," in *2nd Meeting of the North American Chapter of the Association for Computational Linguistics*, pp. 138–143, 2001.

[81] S. Harabagiu, "An application of WordNet to prepositional attachment," in *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pp. 361–362, 1996.

[82] WSJ, "The Wall Street Journal." Online, retrieved October 2013. Available: `http://www.wsj.com`.

[83] A. Esuli and F. Sebastiani, "Pageranking WordNet synsets: An application to opinion mining," in *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pp. 424–431, 2007.

[84] M. Abrate, C. Bacciu, A. Marchetti, and M. Tesconi, "WordNet atlas: a web application for visualizing WordNet as a zoomable map," in *6th International Global Wordnet Conference*, pp. 23–29, 2012.

[85] I. Niles and A. Pease, "Linking lixicons and ontologies: Mapping WordNet to the suggested upper merged ontology," in *Proceedings of the International Conference on Information and Knowledge Engineering*, pp. 412–416, 2003.

[86] F. M. Suchanek, G. Kasneci, and G. Weikum, "Yago: A large ontology from wikipedia and WordNet," *Web Semantics: Science, Services and Agents on the World Wide Web*, vol. 6, pp. 203–217, 2008.

[87] P. L. Arauz, J. Gomez-Romero, and F. Bobillo, "A fuzzy ontology extension of wordnet and eurowordnet for specialized knowledge," in *Proceedings of the 10th Terminology and Knowledge Engineering Conference*, pp. 139–154, 2012.

[88] M. Cuadros, E. Laparra, G. Rigau, P. Vossen, and W. Bosma, "Integrating a large domain ontology of species into WordNet," in *Proceedings of the International Conference on Language Resources and Evaluation*, pp. 2310–2317, 2010.

[89] J. J. Jiang and D. W. Conrath, "Semantic similarity based on corpus statistics and lexical taxonomy," in *Proceedings of 10th International Conference on Research In Computational Linguistics*, pp. 1–15, 1997.

[90] G. Liu, R. Wang, J. Buckley, and H. M. Zhou, "A Wordnet-based semantic similarity measure enhanced by internet-based knowledge.," in *SEKE*, pp. 175–178, Knowledge Systems Institute Graduate School, 2011.

[91] S. Ross, *A First Course in Probability*. UK: Pearson, 1976.

[92] P. Resnik, "Using information content to evaluate semantic similarity in a taxonomy," in *In Proceedings of the 14th International Joint Conference on Artificial Intelligence*, pp. 448–453, 1995.

[93] R. Richardson and A. F. Smeaton, "Using WordNet in a knowledge-based approach to information retrieval," tech. rep., Doublin City University, 1995.

[94] T. Slimani, B. B. Yaghlane, and K. Mellouli, "A new similarity measure based on edge counting," in *Proceedings of world academy of science, engineering and technology*, pp. 773–777, 2006.

[95] H. Yang and J. Callan, "Learning the distance metric in a personal ontology," in *Proceedings of the 2nd international workshop on Ontologies and information systems for the semantic web*, pp. 17–24, 2008.

[96] A. D. Scriver, "Semantic distance in WordNet: A simplified and improved measure of semantic relatedness," Master's thesis, the University of Waterloo, 2006.

[97] G. Hirst and D. St-Onge, "Lexical chains as representation of context for the detection and correction malapropisms," *WordNet: An Electronic Lexical Database (Language, Speech, and Communication)*, 1998.

[98] T. Wang and G. Hirst, "Refining the notions of depth and density in WordNet-based semantic similarity measures," in *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pp. 1003–1011, 2011.

[99] S. Mohammad and G. Hirst, "Distributional measures of concept-distance: A task-oriented evaluation," in *Proceedings of the conference on Empirical Methods in Natural Language Processing*, pp. 35–43, 2006.

[100] H. Hassanzadeh and M. Keyvanpour, "A machine learning based analytical framework for semantic annotation requirements," *International Journal of Web & Semantic Technology*, vol. 2, pp. 27–38, 2011.

[101] T. Finin, Y. Peng, R. Scott, C. Joel, S. A. Joshi, P. Reddivari, R. Pan, V. Doshi, and L. Ding, "Swoogle: A search and metadata engine for the semantic web," in *Proceedings of the Thirteenth ACM Conference on Information and Knowledge Management*, pp. 652–659, ACM Press, 2004.

[102] Wikipedia, "Linked data — Wikipedia, the free encyclopedia." Online, retrieved October 2013. Available: `http://en.wikipedia.org/wiki/Linked_data`.

[103] RuleML Inc, "RuleML home." Online, retrieved October 2013. Available: `http://wiki.ruleml.org/index.php/RuleML_Home`.

[104] H. Boley, "The RuleML family of web rule languages," in *PPSWR'06 Proceedings of the 4th international conference on Principles and Practice of Semantic Web Reasoning*, pp. 1–17, 2006.

[105] Apache, "Apache Jena -Home." Online, retrieved October 2013. Available: `http://jena.apache.org/`.

[106] D. Han and K. Stoffel, "Ontology based qualitative methodology for Chinese language analysis," in *Proceeding of the 26th International Conference on Advanced Information Networking and Applications Workshops (WAINA)*, pp. 839 – 844, ©2012 IEEE., DOI: 10.1109/WAINA.2012.141, 2012.

[107] U. Zollinger and K. Zollinger, "The effects of globalization on sustainable development and the challenges to global governance," tech. rep., Swiss Agency for Development and Cooperation, 2007.

[108] "Globalization and development," tech. rep., United Nations, 2002.

[109] W. Li and B.-C. Dan, "The impact of globalization and the internet on English language teaching and learning." Online, 2006 (retrieved October 2013). Available: `http://www.academia.edu/188911/The_Impact_of_Globalization_and_the_Internet_on_English_Language_Teaching_and_Learning`.

[110] D. Crystal, *Language and the Internet*. UK: Cambridge University Press, 2001.

[111] A. Schwegler, "Language and globalization." Online, 2006 (retrieved October 2013). Available: `http://www.globalization101.org/uploads/File/Syllabus-Lang-Globalization.pdf`.

[112] Wikipedia, "Chinese language — Wikipedia, the free encyclopedia." Online, retrieved October 2013. Available: `http://en.wikipedia.org/wiki/Chinese\language`.

[113] Y. Chen, "Research of the spread of Chinese as a foreign language," *Asian Social Science*, vol. 4, pp. 118–122, 2008.

[114] H. C. Lam, "A critical analysis of the various ways of teaching Chinese characters," *Electronic Journal of Foreign Language Teaching*, vol. 8, pp. 57–70, 2011.

[115] The British Museum, "Chinese symbols." Online, retrieved October 2013. Available: `http://www.britishmuseum.org/pdf/Chinese_symbols_1109.pdf`.

[116] A. Burk, C. Coleman, C. Wimberly, and J. Zapata, "The Chinese Language Manual." Online, 2008 (retrieved October 2013). Available: `http://languagemanuals.weebly.com/uploads/4/8/5/3/4853169/chinesemanual.pdf`.

[117] G. Bell, ed., *Asian perspectives on mathematics education*, ch. Setting the theme: Researching Asian mathematics education, pp. 1–16. The Northern Rivers Mathematical Association, 1993.

[118] L. Galligan, "Possible effects of English-Chinese language differences on the processing of mathematical text: A review," *Mathematics Education Research Journal*, vol. 13, pp. 112–132, 2001.

[119] R. Hoosain, *Psycholinguistic implications for linguistic relativity: a case study of Chinese*. China: Foreign Language Study, 1991.

[120] J. Yin and D. Sun, "Chinese house," tech. rep., the University of Vermont.

[121] Wikipedia, "Text segmentation — Wikipedia, the free encyclopedia." Online, retrieved November 2013. Available: `http://en.wikipedia.org/wiki/Text_segmentation`.

[122] M. L. G. at the University of Waikato, "Weka 3: Data mining software in java." Online, retrieved November 2013. Available: `http://www.cs.waikato.ac.nz/ml/weka/`.

[123] D. Han and K. Stoffel, "An interactive working tool for qualitative text analysis," in *Proceeding of the 12th Francophone International Conference on Knowledge Discovery and Management*, pp. 599–602, 2012.

[124] ICESOFT, "ICEpdf overview." Online, retrieved November 2013. Available: `http://www.icesoft.org/java/projects/ICEpdf/overview.jsf`.

[125] Wikipedia, "Quotation — Wikipedia, the free encyclopedia." Online, retrieved September 2013. Available: `http://en.wikipedia.org/wiki/Quotation`.

[126] P. Resnick, N. Iacovou, M. Suchak, P. Bergstrom, and J. Riedl, "GroupLens: an open architecture for collaborative filtering of netnews," in *Proceedings of the 1994 ACM conference on Computer supported cooperative work*, pp. 175–186, 1994.

[127] Wikipedia, "Cobweb_(clustering) — Wikipedia, the free encyclopedia." Online, retrieved November 2013. Available: `http://en.wikipedia.org/wiki/Cobweb_clustering`.

[128] P. Ciancarini, "Architectures for software services and cloud computing." Online, retrieved September 2013. Available: `http://www.slideshare.net/kronat/9-architetture-software-soa-cloud`.

[129] M. Turner, D. Budgen, and P. Brereton, "Turning software into a service," *Computer*, vol. 36, pp. 38–44, 2003.

[130] S. Viriri, "Dynamic caching design proto-pattern for J2EE web component development," *Journal of Object Technology*, vol. 2, pp. 113–117, 2003.

[131] "The .NET framework for Java developers," tech. rep., Microsoft Corporation, 2011.

[132] A. M. Riad, A. E. Hassan, and Q. F. Hassan, "Design of SOA-based grid computing with enterprise service bus," *International Journal on Advances in Information Sciences and Service Sciences*, vol. 2, pp. 71–82, 2010.

[133] R. Sharma and M. Sood, "Modeling cloud software-as-a-service: a perspective," *International Journal of Information and Electronics Engineering*, vol. 2, pp. 238–242, 2002.

[134] J. Bughin, M. Chui, and J. Manyika, "Clouds, big data, and smart assets: ten tech-enabled business trends to watch," tech. rep., McKinsey Global Institute, 2010.

[135] J. Gantz and D. Reinsel, "Extract ing value from chaos," tech. rep., IDC, 2011.

[136] B. Jungwirth, "Information overload: threat or opportunity," *Journal of Adolescent & Adult Literacy*, vol. 45, pp. 90–91, 2002.

[137] "Cisco visual networking index: global mobile data traffic forecast update, 2012-2017," tech. rep., Cisco Systems, Inc, 2012.

[138] R. P. Mohamad, D. Kolovos, and R. Paige, "Modeling workloads, slas and their violations in cloud computing," in *Proceeding of Fourth York Doctoral Symposium on Computer Science*, pp. 71–76, 2011.

[139] M. O. Adams, *Online Numeric Data-Base Systems: A Resource for The Traditional Library*. IL: Graduate School of Library and Information Science. University of Illinois at Urbana-Champaign, 1982.

[140] B. F. Cooper, N. Sample, M. J. Franklin, G. R. Hjaltason, and M. Shadmon, "A fast index for semistructured data," in *Proceedings of the 27th VLDB Conference*, pp. 341–350, 2001.

[141] "Peer research big data analytics," tech. rep., Intel IT Center, 2012.

[142] W. Trochim and J. P. Donnelly, *The Research Methods Knowledge Base*. OH: Atomic Dog, 2006.

[143] P. Domingos and G. Hulten, "Mining high-speed data streams," in *Proceedings of the sixth ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 71–80, 2000.

[144] P. A. Goloboff, "Analyzing large data sets in reasonable times: Solutions for composite optima," *Cladistics*, vol. 15, pp. 415–428, 1999.

[145] C. Matthews, "Future of retail: how companies can employ big data to create a better shopping experience." Online, 2012 (retrieved September 2013). Available: `http://business.time.com/2012/08/31/future-of-retail-how-companies-can-employ-big-data-to-create-a-better-shopping-experience/`.

[146] D. Selinger, "Why big data means big business for online retailers." Online, 2012 (retrieved September 2013). Available: `http://www.guardian.co.uk/media-network/media-network-blog/2012/dec/19/big-data-new-big-business`.

[147] "Oracle: big data for the enterprise," tech. rep., Oracle, 2012.

[148] "HPCC systems: models for big data," tech. rep., LexisNexis, 2011.

[149] J. S. Vitter and M. Wang, "Approximate computation of multidimensional aggregates of sparse data using wavelets," *ACM SIGMOD Record*, vol. 28, pp. 193–204, 1999.

[150] U. Fayyad and R. Uthurusamy, "Evolving data into mining solutions for insights," *Communications of the ACM*, vol. 45, pp. 28 – 31, 2002.

[151] R. R. Sinha and K. Swearingen, "Comparing recommendations made by online systems and friends," in *DELOS Workshop: Personalisation and Recommender Systems in Digital Libraries*, 2001.

[152] G. Linden, B. Smith, and J. York, "Amazon.com recommendations: item-to-item collaborative filtering," *Internet Computing, IEEE*, vol. 7, pp. 76 – 80, 2003.

[153] "Facebook ads getting started guide," tech. rep., Facebook, 2011.

Publications of the Author

[1] Kilian Stoffel, Paul Cotofrei and Dong Han. "Fuzzy Methods for Forensic Data Analysis. Integration Techniques in Computational Forensics". *Proceedings of 2010 International Conference of Soft Computing and Pattern Recognition (SoCPaR)*, pp. 23-28. 2010

[2] Gil Gomes Dos Santos, Dong Han, Gerald Reiner and Kilian Stoffel. "Method for Screening Sustainability Reports in the Financial Services Industry". *14th Toulon - Verona Conference*. 2011

[3] Dong Han and Kilian Stoffel. "Ontology based Qualitative Case Studies for Sustainability Research". *Proceedings of Workshop A.I. for Intelligent Planet, IJCAI 2011*. 2011

[4] Dong Han and Kilian Stoffel. "An Interactive Working Tool for Qualitative Text Analysis". *Proceeding of the 12th Francophone International Conference on Knowledge Discovery and Management*, pp. 599-602. 2012

[5] Dong Han and Kilian Stoffel. "Ontology based Qualitative Methodology for Chinese Language Analysis". *Proceeding of the 26th International Conference on Advanced Information Networking and Applications Workshops (WAINA)*, pp. 839-844. 2012

[6] Kilian Stoffel, Paul Cotofrei and Dong Han. "Fuzzy Clustering based Methodology for Multidimensional Data Analysis in Computational Forensic Domain". *International Journal of Computer Information Systems and Industrial Management Applications (IJCISIM)*. Volume 4, pp. 400-410. 2012

[7] Dong Han and Kilian Stoffel. "Ontology based Model and Procedure Creation for Topic Analysis in Chinese Language". *Proceeding of Joint Conference of the Sixth Chinese Semantic Web Symposium (CSWS) and the First Chinese Web Science Conference (CWSC)*, pp. 67-73. 2013