ISBN 978-82-326-3188-9 (printed ver.)
ISBN 978-82-326-3189-6 (electronic ver.)
ISSN 1503-8181

NTNU
Norwegian University of
Science and Technology

Mahsa Mehrpoor

Doctoral theses at NTNU, 2018:199

NTNU

NTNU
Norwegian University of Science and Technology
Thesis for the Degree of
Philosophiae Doctor
Faculty of Engineering
Department of Mechanical and Industrial
Engineering

Doctoral thesis

Doctoral theses at NTNU, 2018:199

Mahsa Mehrpoor

An Ontology-Driven Recommender System for Engineering Projects

NTNU
Norwegian University of
Science and Technology

Mahsa Mehrpoor

An Ontology-Driven Recommender System for Engineering Projects

Thesis for the Degree of Philosophiae Doctor

Trondheim, June 2018

Norwegian University of Science and Technology
Faculty of Engineering
Department of Mechanical and Industrial Engineering

NTNU
Norwegian University of
Science and Technology

NTNU
Norwegian University of Science and Technology

Thesis for the Degree of Philosophiae Doctor

Faculty of Engineering
Department of Mechanical and Industrial Engineering

ISBN 978-82-326-3188-9 (printed ver.)
ISBN 978-82-326-3189-6 (electronic ver.)
ISSN 1503-8181

Doctoral theses at NTNU, 2018:199

Printed by NTNU Grafisk senter

To my family

Abstract

Knowledge and information resources in enterprises are rapidly growing. The International Data Corporation (IDC) forecasts that significant yearly growth of data will result that the so-called global datasphere will have grown to 163 zettabytes (ZB) by 2025, which is 10 times of the 16.1 ZB of data generated in 2016. This happens while IT staff to manage it will grow less than 1.5 times (Reinsel, Gantz, & Rydning, 2017). A substantial number of these resources are documents that are potentially valuable for intentional reuse. Knowledge workers and engineers in particular, require specific knowledge and information embedded in different types of knowledge objects stored in internal or external resources (Hertzum & Pejtersen, 2000). However, identifying relevant knowledge from a large number of unstructured enterprise resources is challenging for users. There is a strong need for an approach that identifies users' required information and automatically explores their preferred documents.

This PhD project focuses on improving knowledge access, sharing, and reuse challenges that people, engineers, are faced with in their daily (knowledge-based) work tasks. The proposed solution is a recommender system in professional settings to provide relevant documents for users in specific work contexts based on domain-specific ontologies. A prototype has been developed and validated on a multidisciplinary engineering use case and its performance has been evaluated. The results show that the developed system is a useful tool for improving information access in traditional engineering projects compared to the currently applied solutions. The main contributions of this thesis are:

C1: In-depth analysis of the context of users and the document corpus in an engineering setting by applying information retrieval tools and semantic annotation.

C2. Proposing a framework for a knowledge access system combining recommendation approaches, ontologies, and information retrieval and extraction tools.

C3. Construction of a contextual ontology as knowledge domain, derived from users' work contexts and evaluating its retrievability and coverage against existing documents as resources of knowledge and information.

C4. Validation of the concept of the recommender system for improving knowledge and information access in engineering context by developing a system that uses the proposed ontology-based profiling approach and evaluating the performance of the developed system on a case-study.

Preface

This PhD thesis is submitted to the Norwegian University of Science and Technology (NTNU) for partial fulfillment of the requirements for the degree of philosophiae doctor.

This doctoral work has been performed at the Department of Mechanical and Industrial Engineering, NTNU, Trondheim, Norway with Professor Ole Ivar Sivertsen as the main supervisor from Department of Mechanical and Industrial Engineering and with co-supervisors Professor Jon Atle Gulla from Department of Computer Science, and Adjunct Associate Professor Kjetil Kristensen from Department of Mechanical and Industrial Engineering.

Acknowledgements

I would like to thank my supervisor Professor Ole Ivar Sivertsen for giving me the opportunity to pursue a Ph.D., for being always kind and supportive so that I felt less to be far from the family.

I thank my co-advisor Professor Jon Atle Gulla for his valuable research discussions, being always patient to explain new concepts and ideas, and guiding me through my research. I would like to thank Adjunct Associate Professor Kjetil Kristensen for his valuable feedback and help in contributing to the research activities of the EU-project, LinkedDesign, which was a great opportunity to learn new things.

A special thanks to Dr. Dirk Ahlers, research scientist at Department of Computer Science, who taught me how to proceed with my research and always guided me with his invaluable feedback and constructive criticisms.

I thank the students of the DNV GL fuel fighter 2015 team for their collaboration and cooperation with the experiments and their feedback. I thank my friends Ivar Marthinusen and Christos Kalavrytinos with whom I shared an office for being positive, fun, and energetic.

My deep appreciation belongs to my parents, Mahroo and Javad, for being supportive in all stages of my life. These acknowledgments would not be complete without mentioning my husband, Soroush, who has been always supportive of the challenges and difficult moments that I had during my Ph.D.

Table of contents

List of figures

List of tables

List of abbreviations

AP: Average Precision

CBR: Content-Based Recommendation

CF: Collaborative Filtering

ES: ElasticSearch

IR: Information Retrieval

IS: Information Systems

IE: Information Extraction

KM: Knowledge Management

MAP: Mean Average Precision

MOE: Measures of Effectiveness

OWL: Web Ontology Language

RDF: Resource Description Framework

RS: Recommender System

SE: Search Engine

SUS: System Usability Scale

SW: Semantic Web

TF-IDF: Term Frequency–Inverse Document Frequency

TAM: Technology Acceptance Model

VSM: Vector Space Model

1 Introduction

Knowledge and information resources play a pivotal role in enterprises and they are rapidly growing. As shown in Figure 1.1, IDC forecasts that by 2025 the global data sphere will grow by 163 zettabytes (a trillion gigabytes). That's ten times the 16.1ZB of data generated in 2016 (Reinsel, Gantz, & Rydning, 2017). The world's information is significantly growing while IT staff to manage it will grow less than 1.5 times. These valuable resources have a reuse potential in similar works and projects to save significant time that is being spent on searching and recreating them. However, since the amount of archived knowledge and information is extensive and usually not very well annotated, it has become a challenge for people to find existing pieces of information that are reusable.

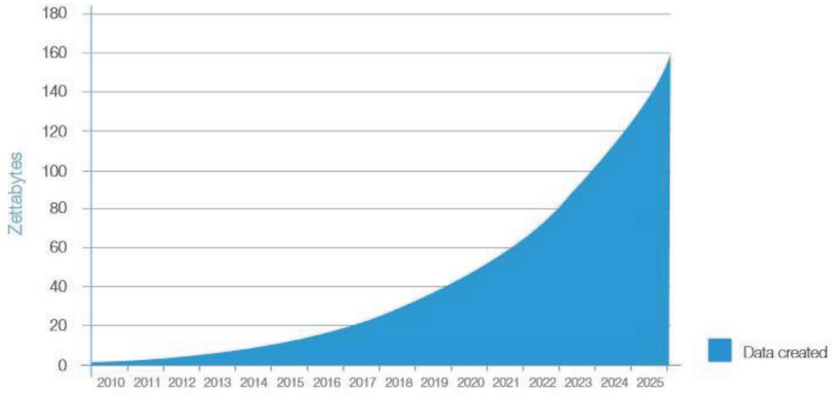


Figure 1.1. The trend of digital data growth. (Reinsel, Gantz, & Rydning, 2017)

As knowledge workers, people in an enterprise – engineers in particular – require specific knowledge and information embedded in different types of knowledge objects stored in internal or external resources. Specifically, in any discipline, engineers with different tasks and level of expertise need to find their required information (Hertzum & Pejtersen, 2000) – within and especially across different engineering disciplines in large, multidisciplinary design and engineering projects.

However, despite being a regular or frequently occurring activity of critical importance to most engineers, effective and efficient search is often not straightforward. Searching for information (defined as the search barrier; people are unable to easily find what they seek) is identified as one of four main barriers to collaboration in a study of 107 companies from various industries including manufacturing, financial services, high-tech, consumer goods/retail, healthcare, professional services and energy (M. Hansen, 2009; M. T. Hansen & Nohria, 2004).

The survey performed in (Williams, Figueiredo, & Trevelyan, 2013) discusses different types of interactions that engineers have for accessing knowledge and information in an organization. These interactions are classified into three groups of face to face, through documents, and interactions with abstract systems and data. Such interactions are identified as searching for information in a file system, databases, the Web etc. for design, modeling, simulation and creating software code etc. (Williams et al., 2013). The interactions with systems are the focus of this research.

This Ph.D. thesis addresses the importance of knowledge and information management in enterprises in terms of reusing the existing knowledge and information in similar projects and products. The methods of improving knowledge and information access and reducing the spent time for acquiring knowledge from experienced people are discussed. As stated by IDC, the number of unstructured data in enterprises are rapidly growing compared to structured data as shown in Figure 1.2.

There is a strong need for an approach that identifies users' required information and explores their preferred documents. The main focus of this research is to exploit relevant knowledge and information from large number of archived unstructured documents that meet users' information needs in varied work contexts. In order to achieve this, a recommender system has been developed which is named ProRecSys. The overall framework of the proposed system for improving knowledge and information access in the engineering settings is presented in Figure 1.3. This framework consists of the essential components that are identified to develop the target system. Detailed description is addressed in (Mehrpoor et al., 2015).

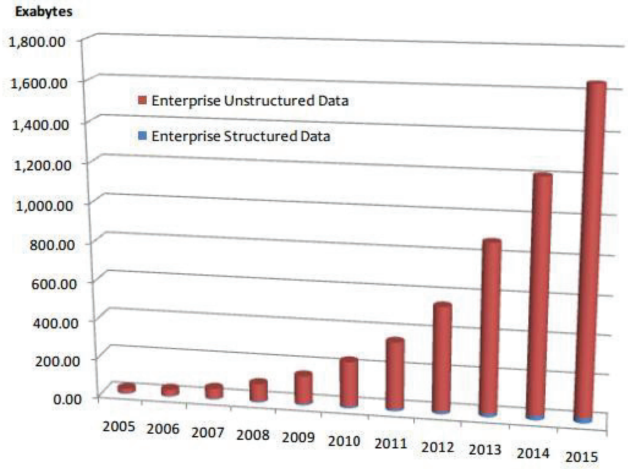


Figure 1.2. The trend of Structured and unstructured data growth in enterprises. IDC digital universe study, sponsored by EMC, 2011

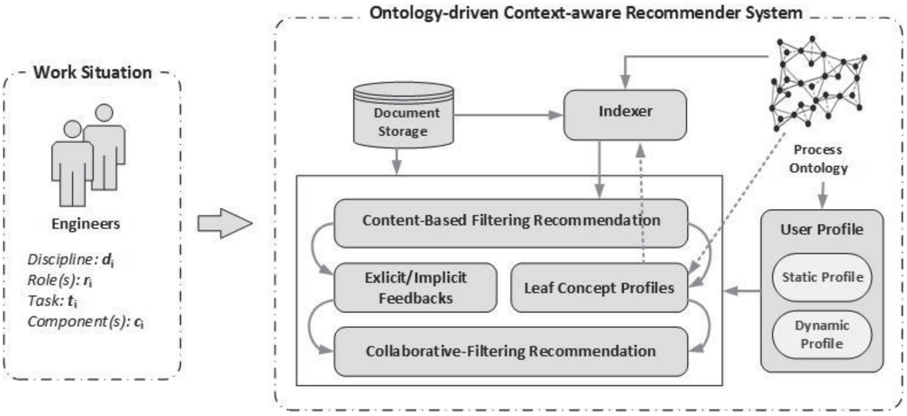


Figure 1.3. The framework of the ontology-driven context-aware recommender system (Mehrpoor et al., 2015)

This chapter starts with describing the motivation and research goals of this Ph.D. work. Afterwards, the research questions, the contributions, and published results to address the research questions have been described. At the end, the structure of the thesis has been presented.

1.1 Motivation

Data repositories in enterprises contain a vast number of documents in different types that may contain valuable knowledge and information to be reused in future works and projects which may have common or similar parts. As projects proceed, the number of generating documents are growing and being stored in common or individual repositories and they are not managed appropriately (Louis-Sidney, Cheutet, Lamouri, Puron, & Mezza, 2012) (Denkena, Shpitalni, Kowalski, Molcho, & Zipori, 2007). This large number of documents (that might be structured or unstructured) lead to the challenge of information overload in enterprises. Many of the enterprises admit that they are not aware of the actual extent of knowledge that exists in their organizations (Le Duigou, Bernard, Perry, & Delplace, 2012). It is evidenced that significant amount of time is dedicated to searching and acquiring the knowledge that actually already exist in the organization (Lowe, McMahon, & Culley, 2004). Depending on what method is used for storing the documents, finding and searching existing documents might be more challenging for accessing the potentially reusable resources.

One of the solutions that are used in enterprises to find the reusable knowledge and information is to rely on the experiences and memory of the personnel that have worked on a specific field for years (Bruno, 2015). However, this may lead to a time-consuming communication process and also these people may not remember all the similar past works or even the relevant people might have left the organization and their knowledge is taken away, thus not accessible. To enhance engineers' productivity, available data sources should be efficiently re-useable and refundable without expensive user annotations to avoid wasting time on searching knowledge that already exists within the organization, contributing to a lean enterprise (Kristensen, Krogstie, Ahlers, & Mehrpoor, 2016).

In each engineering discipline, people have different levels of expertise and deal with different tasks. To accomplish their tasks, they search through the relevant existing knowledge and information that meet their information needs to find out about facts, best practices, procedural information for doing a task, relevant tools to use, similar examples to reuse, and any other inputs that help them perform the assigned tasks. Many studies and research have been done on this issue to assist the people in enterprises in finding the related knowledge and information required to perform their tasks.

In general, a method for automating the identification of reusable and relevant resources could be beneficial to speed up design and manufacturing a product. In particular, individualizing these resources for each user can be helpful since each one has specific requirements depending on the specifications of the assigned tasks. Therefore, more detailed understanding of users' needs could lead to better identification of those pieces of knowledge and information that match better with their information needs.

In this thesis, major challenges that enterprises are faced with in accessing and identifying existing knowledge and information are addressed particularly in engineering projects to use them in their tasks. The methods and approaches that are applied to improve knowledge management will be discussed and the applied solution on the case-study (cf. Section 3.1) will be presented from early stages of system analysis to late stages of system development, experiments, and evaluation.

1.2 Research goals and research questions

In this research, the main objective is to explore different approaches that are applied in engineering settings for improving knowledge and information access during project development processes by identifying the relevant resources that meet users' information needs. Based on the motivation stated in the previous section, the research questions are addressed in this thesis and correspond to a number of large themes that have been discussed below.

Users' context and their challenges for knowledge access in engineering settings

The situation of the users in a professional setting has a set of characteristics or factors that define the specific context of a user. Inspecting what factors describe the context of a user and influence on his/her information needs, is one of the key points of this work. Thus, the first research questions are:

RQ1.What are the challenges that users deal with to identify their information needs in engineering settings?

RQ2. What factors characterize the context of a user in a professional setting and their effect on the type of information that he/she needs?

Knowledge and information context in engineering settings

Knowledge and information are documented in different methods in enterprises. Different types of repositories are used to store and manage these resources. Based on the case-study which is an example of a traditional engineering project, this work examines networked shared file system databases and how documents of different formats are managed and organized in these types of repositories. Hence, the research question on this topic is:

RQ3. What are the features and specifications of knowledge repositories and knowledge resources in a traditional engineering setting? And how they can be explored to meet users' information needs?

Knowledge and information management tools and technologies in engineering settings

Enterprises employ different methods and tools for managing and organizing the growing amount of knowledge and information. As mentioned, the aim is to enable systematic identification of relevant documents that match users' professional context. Therefore, recent methods of information retrieval and access tools and more particularly search engines, recommender systems, and ontologies have been studied and investigated how well these tools can be utilized in a traditional engineering context to improve knowledge and information access. In this context, the research questions are as follows;

RQ4. What is a tailored framework for the development of target recommender system according to the analyzed specifications of users and document corpus?

RQ5. How can enterprises utilize and combine the synergy of ontologies, information retrieval techniques, and recommendation approaches to improve knowledge management in professional settings?

RQ6. How can the developed system for knowledge access improvement in professional settings be validated and evaluated?

1.3 Contributions

The contribution of this PhD research is to combine existing solutions with the proposed approach which is applying ontology-driven concept profiles in the process of recommendation. These contributions answer the research questions as follows:

C1. The context of users and knowledge in a traditional engineering context has been investigated through a detailed analysis on these concepts using the engineering case-study (cf. Section 3.1).

C2. The framework of a knowledge access system has been proposed that consists of the synergy of recommendation approaches, ontology as knowledge domain, and information retrieval and extraction tools.

C3. The proposed contextual ontology derived from aspects of users' work context has been developed and validated through the proposed ontology-based content matching approach and the level of retrievability and coverage of the ontology has been evaluated against the document corpus.

C4. The concept of the recommender system for improving knowledge and information access in engineering context has been validated and the proposed system has been developed using the proposed ontology-driven concept profiles and evaluated in an engineering case-study.

The exploration of the research questions was published through several articles and the main contribution of each article is listed in the Table 1. The detailed steps will be discussed later in the Section 3.2, Figure 3.4.

Table 1. Research contributions

RQ	Contribution	Paper	Focus
R1, R2	C1	P1 (Mehrpoor, Gjarde, & Sivertsen, 2014)	Conceptual
R4	C1, C2	P2 (Mehrpoor, Gulla, Ahlers, Kristensen, Ghodrat, & Sivertsen, 2015)	Analysis, Design
R3	C1, C3	P3 (Mehrpoor, Ahlers, Gulla, Kristensen, & Sivertsen, 2017)	Analysis, Evaluation
R5, R6	C4	P4 (submitted to KAIS journal)	Development, Evaluation

To give an overview on the contributions of this Ph.D., the list of main publications are represented in Table 2 and each publication will be described in more detail in chapter 4. In addition, the results of further contributions are listed in Table 3 as supplementary publications to the Ph.D. work. Paper 5 (Ahlers, Mehrpoor, Kristensen, & Krogstie, 2015) discusses the challenges of data management in larger engineering scales, engineers' information needs, and search tasks. It discusses the use of information retrieval, recommender systems, and knowledge management methods and tools to improve the daily information seeking workflow in knowledge-intense disciplines. Paper 6 (Ahlers & Mehrpoor, 2015) discusses methods of managing and sharing documents in professional settings and the challenges that engineers have to find their information needs using ordinary search tools. An approach is proposed to improve professional search by joining content and metadata analysis, link derivation, grouping, and other measures to arrive at high-level features suitable for semantic similarity and retrieval to improve information access. The related contribution in the book chapter (Kristensen, Krogstie, Ahlers, & Mehrpoor, 2016) refers to proposing context-aware recommender systems for knowledge access improvement in multi-disciplinary engineering projects. The idea of investigating the context of the engineers in workplaces is proposed to provide better solutions for semantically identifying pieces of knowledge and information the meet their information needs by applying recommender systems and ontologies. Such systems improve the shortcomings of ordinary search engines and help engineers access to their information needs.

List of papers in the main body (Appendix A)

Table 2. List of publications in the main body

Paper I	**Intelligent Services: A Semantic Recommender System for Knowledge Representation in Industry**
	Authors: Mahsa Mehrpoor, Andreas Gjærde, Ole Ivar Sivertsen
	ICE Conference, 2014 (Peer-reviewed)
Paper II	**Using Process Ontologies to Contextualize Recommender Systems in Engineering Projects for Knowledge Access Improvement**
	Authors: Mahsa Mehrpoor, Jon Atle Gulla, Dirk Ahlers, Kjetil Kristensen, Soroush Ghodrat, Ole Ivar Sivertsen
	ECKM Conference, 2015 (Peer-reviewed)
Paper III	**Investigating contextual ontologies and document corpus characteristics for information access in engineering settings**
	Authors: Mahsa Mehrpoor, Jon Atle Gulla, Dirk Ahlers, Kjetil Kristensen, Soroush Ghodrat, Ole Ivar Sivertsen
	Journal of Information Technology Case and Application Research, 2017
Paper IV	**Development and Evaluation of a Knowledge Access System for Engineering Workspaces Based on Recommendation and Filtering**
	Authors: Mahsa Mehrpoor, Dirk Ahlers, Jon Atle Gulla, Ole Ivar Sivertsen
	Submitted to Journal of Knowledge and Information Systems, Feb. 2018

List of supplementary papers (Appendix B)

Table 3. List of supplementary papers

Paper V	**Challenges for Information Access in Multi-Disciplinary Product Design and Engineering Settings**
	Authors: Dirk Ahlers, Mahsa Mehrpoor, Kjetil Kristensen, John Krogstie
	ICDIM Conference, 2015
Paper VI	**Everything is Filed under File: Conceptual Challenges in Applying Semantic Search to Network Shares for Collaborative Work**
	Authors: Dirk Ahlers, Mahsa Mehrpoor, Kjetil Kristensen, John Krogstie
	ACM Conference on Hypertext and Social Media, 2015
Book chapter	**LEAP Collaboration System**
	Authors: Kjetil Kristensen, John Krogstie, Dirk Ahlers, Mahsa Mehrpoor
	Taking the LEAP book: The Methods and Tools of the Linked Engineering and Manufacturing Platform (LEAP), 2016

1.4 Thesis organization

This thesis is divided into 5 chapters. The first chapter presents the motivation of the research and states the problem, the research goals and research questions are represented and afterwards, the contributions to achieve the goals and address the research questions have been described. The second chapter is about the research areas that have been studied in this Ph.D. work. The third chapter briefly introduces the case study in the engineering context and continues with describing the applied research method. The fourth chapter represents an overview of the results of the research published in conferences and journal papers and the last chapter draws the conclusions and discusses future work.

2 Background and state-of-the-art

This chapter discusses related work from relevant research areas that are a basis to understand the contributions in the following chapters. First, the concept of knowledge and information in enterprises is described along with the applied methods of knowledge and information management. Then, recent technologies that have been utilized to improve knowledge access, sharing, and reuse in the professional settings are elaborated.

2.1 Knowledge and information management in organizations

Knowledge and information are valuable resources in enterprises and have a significant role in their success (Davenport & Prusak, 1998; Wellman, 2009). Therefore, managing and organizing them is highly important in an enterprise. In this section, the concept of knowledge and information is defined along with the type of knowledge that is the emphasis of this research. Furthermore, the challenges that stakeholders deal with for managing, sharing, and accessing the required knowledge will be discussed.

2.1.1 Knowledge and information definition

For describing information and knowledge, we first need to define what data is, since their definitions are closely related. According to (Thierauf, 1999), data means "unstructured facts and figures that have the least impact on the typical manager." To exemplify in an engineering context, 19mm is known as data that represents the measurement of an object or artifact. The definition of information is "For data to become information, it must be contextualized, categorized, calculated and condensed" as stated in (Davenport & Prusak, 1998). In our example, we know what data means: The master cylinder is 19mm. Here, we have more input about the object that is 19mm. And regarding knowledge, "Knowledge is closely linked to doing and implies know-how and understanding." as stated in (Davenport & Prusak, 1998). To follow the example, a master cylinder with 19mm is the right size to be combined with the caliper to make a brake pedal. Therefore, knowledge is information about information which guides people on how to accomplish a task.

2.1.2 Various types of knowledge

Knowledge can be categorized into two groups of explicit knowledge and tacit knowledge (Nonaka & Takeuchi, 1995). Explicit knowledge is a type of knowledge that is formalized and codified and it is usually referred to as know-what (Brown & Duguid, 1998). Knowledge management tools are usually effective in storing and retrieving them. However, it has been also a challenge to ensure that explicit knowledge is accessible to the people who need it, it is appropriately stored and can be identified and retrieved easily. These type of knowledge is found in databases, memories, documents, notes, and so on.

Tacit knowledge is a type of knowledge that is usually referred to as know-how. It is intuitive, hard to define and mainly based on experience (Brown & Duguid, 1998). It is known as the most valuable source of knowledge particularly in organizations (Wellman, 2009). This type of knowledge is challenging to convey from experienced people to beginners in a field. Tacit knowledge is found in the mind of people which include expertise, skills and capabilities, attitudes, values, mental models, and so on (Botha, 2008). In this Ph.D. work, the focus is on explicit knowledge and how to improve its management in the professional settings.

2.1.3 Current Challenges of knowledge management in enterprises

Engineers' interactions for knowledge access and sharing

To perform a task, relevant knowledge and information is required to analyze task specifications and requirements. To find and access the required information, engineers employ different ways such as direct collaboration with informed people, through reading documents, and interactions with software-based systems (Williams et al., 2013). However, these communications and interactions with people and searching through documents are time-consuming processes. As evidenced by several studies, engineers may spend about 40-66% of their time for finding the needed input to their assigned tasks (King, 1994). On the other side, finding the relevant knowledge and information is critical, since they might have valuable content to be reused in new tasks.

Information overload and search for information needs

Search for information is reported as one of the four challenges of collaboration among people in enterprises (M. T. Hansen & Nohria, 2004). Finding the required information is not straightforward and people need to collaborate with other experts to get the input for their work. However, sometimes the experts are not accessible to assist those that need their expertise. Increasingly growth of knowledge and information makes search more challenging to find and access the required documents stored in varied data repositories (Ahlers, Mehrpoor, Kristensen, & Krogstie, 2015). Using desktop search engines may not be efficient enough for engineers' information needs and expectations (Ahlers & Mehrpoor, 2015). Moreover, task complexity increases the complexity of information needs and the success of information seeking decreases (Byström & Järvelin, 1995). A solution that improves knowledge access and sharing are required to be able to identify the right knowledge and information that meet engineers' information needs for performing different tasks.

2.1.4 Knowledge and information retrieval

"Information retrieval (IR) is finding material (usually documents) of an unstructured nature (usually text) that satisfies an information need from within large collections (usually stored on computers)." (Manning, Raghavan, & Schütze, 2008).

In this section, different document types are described along with commonly used data management systems and search tools in organizations. Afterwards, the common subjects and terms that have been used in an information storing and retrieval domain are described such as metadata, annotation, and indexing documents.

Structured and unstructured documents

In any database management system used in organizations, knowledge and information is embedded in different document types that might be unstructured, semi-structured, or structured. When a document is structured, it means that some method is used to give the whole, or parts, of the document various structural meanings according to a schema for organizing data. While, in the unstructured documents, information can appear in unexpected places on the document and is not based on a defined template or outline such as a letter or a hand-written note (Van Ittersum & Spalding, 2005).

Document storing and management systems

For storing and retrieving documents, organizations employ data storage and document management systems (DMS). File systems are one of the common systems used for managing data in engineering workplaces. In these systems, groups of data are given a name which makes the information to be easily identified. Each group of data is called a file that is stored in hierarchical directories. Any file needs to have a specific name and path to be used for storing and retrieving data. Many organizations have a standard convention for naming file and path, including the utilization of version numbers, states flag, file creation dates, filenames, and information regarding the person or system used to create a file (Eck & Schaefer, 2011).

In order to retrieve the documents, file systems are explored by search systems known as desktop search tools to search a user's own computer files. Desktop search tools typically collect three types of information about files; file and folder names, metadata, and file content (only for supported types of documents).

Metadata, annotation, and indexing

To find the documents that meet engineers' information needs, firstly, documents need to be searchable. At the beginning, a sort of information needs to be specified for a document to annotate the document based on them which is known as metadata. Metadata means data representing a set of properties of a given type and meaning about the contents of an object and provides extra information about it. Examples of metadata properties in the application software system include names, dates and times, comments, locations, descriptions, sizes, dimensions, exposure data, keywords and phrases, links, ownership, and so on (Horn, 2016). File systems define different metadata that is used while searching for stored documents. A document might have some annotations as well; annotations are actually metadata that was not originally associated with an object, but which is defined or specified by either the user or the system for organizational purposes (Horn, 2016).

After specifying the required metadata for a document, the document can be indexed based on the specified metadata and becomes available for searching. During the process of indexing, every document is stored and organized based on the specified metadata and it optimizes the speed and performance in finding relevant documents and make the documents more accessible.

2.1.5 Users' information seeking in work environments, task-based search, and context-driven search

As data and information are rapidly growing, the need of searching increases in enterprises as well as other environments. Finding a piece of useful information that can be applied in a problem or task is challenging among a large number of documents. There are many factors that need to be considered about the user in a particular work situation to figure out the type of their information needs (Mehrpoor et al., 2014). In recent years, these factors have studied in different research such as in (Freund, 2008) that investigates the role of contextual factors in determining how professionals in workplaces search and select information and affect their search behavior. In this work, among identifying the contextual factors, work tasks and information tasks were found to be significantly associated with document genres (Freund, 2008).

Task types have been shown to influence search behaviors. As another research work, user behaviors associated with different task types have been investigated. An investigation is performed on a group of users and was asked to search on four tasks from four different dimensions of complexity, task product, task goal, and task level (Jingjing Liu et al., 2010). The results report regular differences in different task characteristics in several search behaviors. These behaviors can be used as implicit indicators of the user's task type. For predicting potentially useful documents based on the type of tasks, multiple use behavioral measures were modeled in (C. Liu, Liu, Cole, Belkin, & Zhang, 2012) as evidence for implicit relevance feedback. The results show that combining multiple behaviors on content pages and search results can improve the prediction of useful documents.

Another research on investigating people's behavior in information search tasks reports on relationships between tasks and individual reading behavior at task level (Cole et al., 2011). Users' information seeking behaviors and their task types have been regarded to be very effective in identifying their information needs and providing the documents that can be relevant to users' tasks (P. Hansen, 2011).

2.2 Semantic web technologies

"The semantic web is not a separate Web but an extension of the current one, in which information is given well-defined meaning, better enabling computers and people to work in cooperation" (Berners-Lee, Hendler, & Lassila, 2001). The aim of the semantic web is to provide meaning for representing information and enabling it to be processed by computers. This section describes the concept ontology which is used as a tool in semantic web technologies. The methodologies for constructing ontologies are described and the role of ontologies in knowledge management is discussed.

2.2.1 Ontology definition and its components

According to a definition by Gruber, "An ontology is a formal, explicit specification of a shared conceptualization." (Gruber, 1993). To describe the description in more detail, the specified attributes are denoted in (Domingue, Fensel, & Hendler, 2011) that are summarized below:

- **Formal**: It refers to representing the specification in a formal language such as RDF in the semantic web which can be processed by machines.
- **Explicit**: It refers to how much of a domain should be contained in a specification: the level of granularity and the level of genericity
- **Specification**: It refers to an ontology which is a description of the artifact and is independent of the entity described.
- **Shared**: It refers to an ontology if it is shared by a community of use. The purpose of ontologies is that they support interoperability between the designer and the user.
- **Conceptualization**: It refers to an abstract simplified view of a domain of interest which is required for some task or purpose.

In order to define the relationships between entities in a domain, ontologies use different components. Common components are listed in the following:

- **Classes or concepts**: a class represents a group of individuals that share common characteristics and it is known as the core component of ontologies. For example, Engineer is a class that refers to any individual that has the characteristics that define an engineer.
- **Individuals or instances**: They are the objects that the ontology describes. For example, John is an individual for the class Engineer.
- **Relations**: It describes the ways in which classes and individuals can be related to each other. For example, a relation between two individuals: John has role Analyst.
- **Attributes**: It refers to the aspects, properties, characteristics, or parameters that a class can have. For example, an engineer has personal information such as name and age, or professional information such as role, skills, tasks, and so on.

Figure 2.1 represents a simple ontology regarding the mentioned example on Engineer and Role classes.

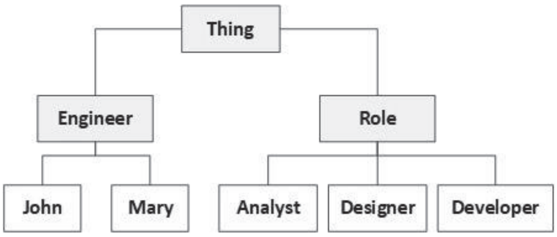


Figure 2.1. An example of a simple ontology to represent Engineer and Role concepts

2.2.2 Methodologies for building ontologies

In order to build ontologies, varied methodologies have been used in different areas namely METHONTOLOGY[1], On-To-Knowledge[2], DILIGENT[3], and NeOn[4] methodology that contain guidelines for single ontology construction from the early stages of ontology specification to the late stages of ontology implementation. In this Ph.D. work, NeOn methodology was applied for building the ontology (Suárez-Figueroa, Gómez-Pérez, & Fernández-López, 2012). In contrast with other methodologies and approaches, NeOn methodology suggests a variety of pathways and does not prescribe a rigid workflow for building ontologies. In addition, this methodology supports the collaborative aspects of ontology development and re-use, as well as the dynamic evolution of ontology networks.

The NeOn methodology contains nine flexible scenarios for building ontologies and ontology networks with the emphasis on reuse of ontological and non-ontological resources, re-engineering and merging, and considering collaboration and dynamism. Each scenario is decomposed into different processes and activities as represented in Figure 2.2. Note that, these scenarios can be combined in flexible ways including the first scenario which contains the core activities and is required for any building process.

In this research, there was not any ontology available that describes the knowledge domain of our case study and the ontology needed to be created from scratch through studying the domain and collecting the involved concepts. Thus, the first scenario is used. This scenario is called “From specification to implementation” and is used when the ontology is developed from scratch without reusing any knowledge resources. Following this scenario, firstly the requirements that the ontology should fulfill is specified through ontology requirement specification activity. To perform this activity, NeOn methodology proposes a filling card which is shown in Figure 2.3.

[1] . www.semanticweb.org/wiki/METHONTOLOGY

[2] . www.ontotext.com/research/otk

[3] . www.semanticweb.org/wiki/DILIGENT

[4] . www.neon-project.org/

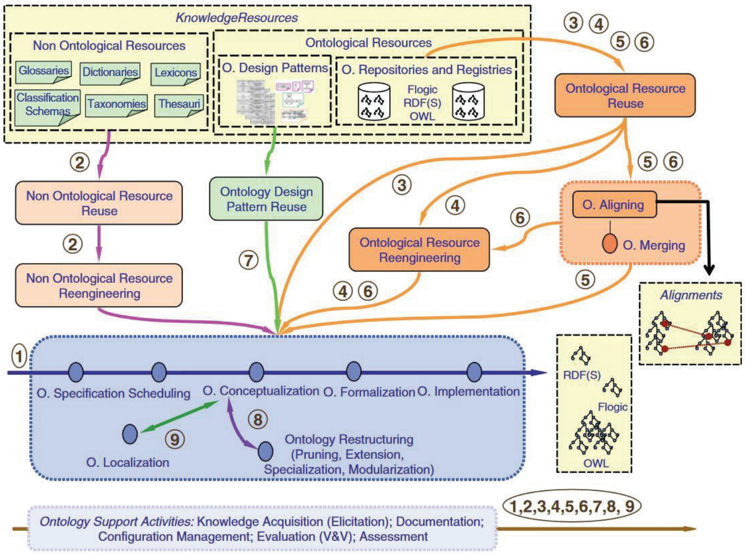
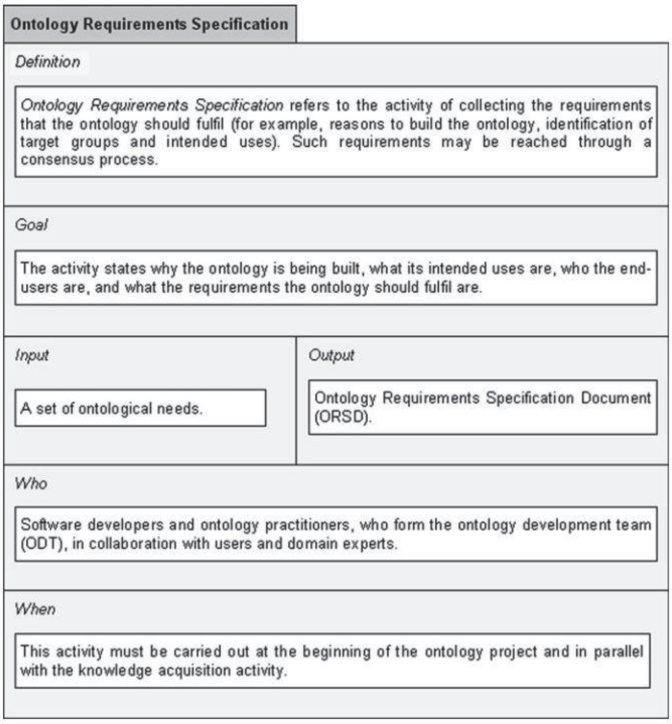

KnowledgeResources
Non Ontological Resources
Glossaries
Dictionaries
Lexicons
Classification Schemas
Taxonomies
Thesauri
Ontological Resources
O. Design Patterns
O. Repositories and Registries
Flogic
RDF(S)
OWL
③ ④
⑤ ⑥
Ontological Resource Reuse
②
Non Ontological Resource Reuse
②
Non Ontological Resource Reengineering
Ontology Design Pattern Reuse
⑦
③
④
⑥
Ontological Resource Reengineering
④ ⑥
⑤ ⑥
O. Aligning
O. Merging
⑤
Alignments
①
O. Specification Scheduling
O. Conceptualization
O. Formalization
O. Implementation
RDF(S)
Flogic
OWL
⑧
⑨
O. Localization
Ontology Restructuring (Pruning, Extension, Specialization, Modularization)
1,2,3,4,5,6,7,8, 9
Ontology Support Activities: Knowledge Acquisition (Elicitation); Documentation; Configuration Management; Evaluation (V&V); Assessment

Figure 2.2. Scenarios for building ontologies by NeOn methodology (Suárez-Figueroa et al., 2012)

This activity contains 8 tasks for building an ontology that is represented in Figure 2.4. The output of this activity is collected as Ontology Requirement Specification Document (ORSD). After performing ORS activity, candidate knowledge resources should be identified for reuse intention. Then, ontology building process is scheduled and eventually, three phases are carried out by ontology developers. Firstly, "ontology conceptualization activity" where knowledge is organized and structured into meaningful models at the knowledge level. Secondly, "ontology formalization activity" where the conceptual model is transformed into a semi-computable model. And thirdly, "ontology implementation activity" where the computable model is implemented by an ontology language and gets generated (Suárez-Figueroa et al., 2012).



Ontology Requirements Specification

Definition

Ontology Requirements Specification refers to the activity of collecting the requirements that the ontology should fulfil (for example, reasons to build the ontology, identification of target groups and intended uses). Such requirements may be reached through a consensus process.

Goal

The activity states why the ontology is being built, what its intended uses are, who the end-users are, and what the requirements the ontology should fulfil are.

Input	*Output*
A set of ontological needs.	Ontology Requirements Specification Document (ORSD).

Who

Software developers and ontology practitioners, who form the ontology development team (ODT), in collaboration with users and domain experts.

When

This activity must be carried out at the beginning of the ontology project and in parallel with the knowledge acquisition activity.

Figure 2.3. Ontology requirements specification filling card (Suárez-Figueroa et al., 2012)

2.2.3 Ontologies for knowledge management in organizations

Usage and benefits of ontologies

It is reported by major knowledge management applications that ontologies are mainly used for three general purposes (Abecker & van Elst, 2009; Davies, Fensel, & Van Harmelen, 2003):

1. *Ontologies support knowledge search, retrieval, and personalization*
2. *Ontologies serve as a basis for information gathering, integration, and organization*
3. *Ontologies support knowledge visualization*

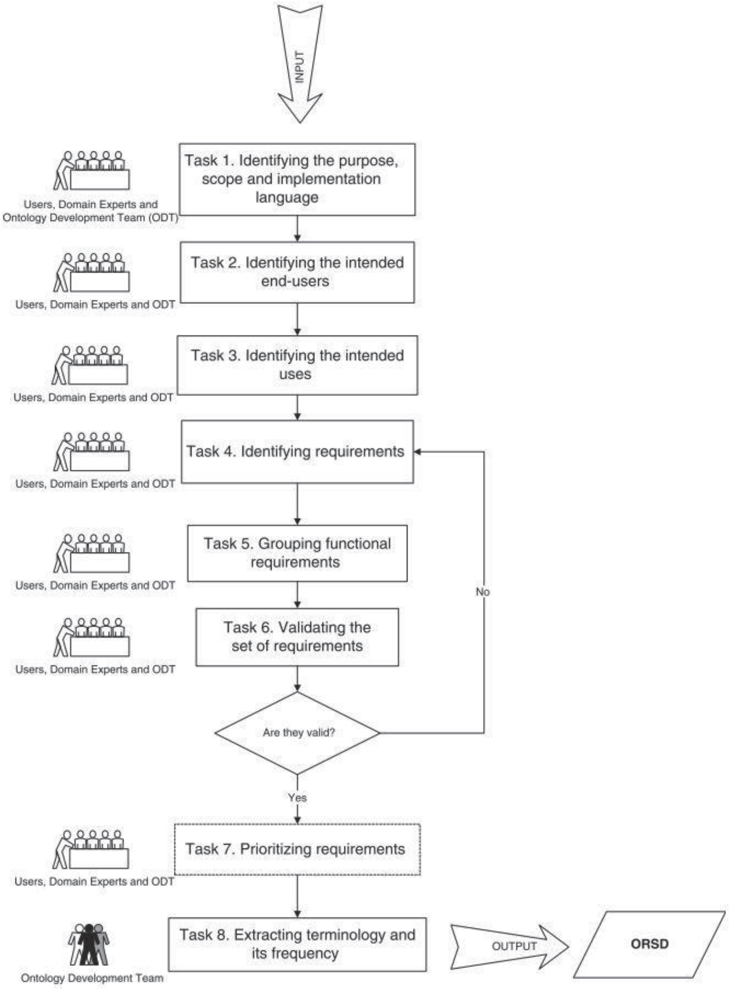

INPUT
Task 1. Identifying the purpose, scope and implementation language
Users, Domain Experts and Ontology Development Team (ODT)
Task 2. Identifying the intended end-users
Users, Domain Experts and ODT
Task 3. Identifying the intended uses
Users, Domain Experts and ODT
Task 4. Identifying requirements
Users, Domain Experts and ODT
Task 5. Grouping functional requirements
Users, Domain Experts and ODT
Task 6. Validating the set of requirements
Users, Domain Experts and ODT
Are they valid?
No
Yes
Task 7. Prioritizing requirements
Users, Domain Experts and ODT
Task 8. Extracting terminology and its frequency
Ontology Development Team
OUTPUT
ORSD

Figure 2.4 Tasks for ontology requirements specification (Suárez-Figueroa et al., 2012)

Regarding the first usage of ontologies, an ontology-based tool is built for processing loads of heterogeneous, distributed, and semi-structured documents. This system exploits the power of ontologies to provide automated support for acquiring, maintaining, and accessing weakly structured information sources (Fensel, 2002). In (McGuinness, 1998), taxonomies are proposed to improve information retrieval during browsing and querying. In case of having not sufficient input for querying, taxonomic knowledge is used for extending the query by super-concepts and sub-concepts.

For improving design tasks in an engineering context, a document model is proposed with aim of semantic tags for annotation. Regarding their case study which is digital camera families, they illustrate how the faceted search and retrieval of product information can be accomplished based on the semantically annotated camera family ontology (Lim, Liu, & Lee, 2009). In order to improve engineering information retrieval during product lifecycle, a computational framework is proposed in

(Li, Raskin, & Ramani, 2007) and an engineering ontology, EO, is developed for representing established design and manufacturing knowledge. The proposed framework for EO-based search system outperforms the keyword-based search in retrieving unstructured engineering documents. Its experimental results report that the system understands users' query at concept level when exact query terms are not available and therefore improves engineering information retrieval that is not properly handled by traditional information retrieval systems.

Regarding the second usage of ontologies, having a more formal way of gathering and organizing information leads to better inferencing and deriving new knowledge. Ontologies are a tailored basis for such inferencing and structuring informal knowledge resources. In (Alani et al., 2003), to improve knowledge extraction from unstructured text on the web using a knowledge extraction tool, a domain-specific ontology is used to determine in detail what type of knowledge needs to be harvested and it is linked to the extraction tool. The ontology uses concepts and relations to classify domain knowledge and then used in the knowledge extraction process to match extracting knowledge to the classification structure.

Referring to an example in enterprises, an ontology-based workflow system is constructed for accumulating and integrating knowledge during the business process and the developed prototype system is applied on two cases of an aircraft industry and a barcode management project (Huang & Diao, 2008). In (Chang, Sahin, & Terpenny, 2008) a graphical modeling tool is developed to support designers at the conceptual design stage. The developed modeling tool works with an ontology-based approach for knowledge management and promotes the systematic capture of design knowledge and improves reusing of design knowledge. The proposed ontology-based method improves integrating heterogeneous data resources and provides more accurate and comprehensive data.

Regarding the third usage of ontologies, visualization of content structure is valuable for finding useful knowledge items and their relations. In order to support users' interaction on shared resources, and support information discovery, developing a visualization ontology is proposed by UK national e-science program (Duke, Brodlie, & Duce, 2004). For capturing architectural knowledge, an ontology-driven visualization of architectural design decisions is proposed in (De Boer, Lago, Telea, & Van Vliet, 2009) for helping software product audits in product quality assessment. They combine a tabular information representation with a real-time ontological inference of decision attributes used by auditors. The proposed solution enables auditors in efficient knowledge reuse and also assists them in their decision-making process.

Due to the complexity of patient data which might be heteronomous, found in different formats and structures, and carry different semantics, ontology concept is utilized for visualizing these types of data to assist clinical decision support systems in exploring similar patients. They aim to map patient data onto a relevant fragment of ontologies and inferred ontological structures as a basis for improving patient data visualization, comparison, and analysis (Zillner et al., 2008). Ontologies are adopted to solve the challenge of finding information in the engineering contexts. As an example, ontology is utilized for semantical representation of the content of manuals for mechanics (Ha et al., 2014). A tailored ontology is modeled after analyzing aircraft maintenance processes together with preprocessing of raw data of maintenance manuals to well-formed format. Then, a set of rules are created for mapping the well-formed documents and ontology schema. The proposed solutions enable the mechanics to easily obtain the information to given tasks, reduce their time for searching required information, and understand the information through visualization.

2.3 Recommender systems

In Resnick and Varian's seminal article, a recommender system is defined as: "In a typical recommender system people provide recommendations as inputs, which the system then aggregates and directs to appropriate recipients. In some cases the primary transformation is in the aggregation; in others, the system's value lies in its ability to make good matches between the recommenders and those seeking recommendations." (Resnick & Varian, 1997).

2.3.1 Recommendation approaches

In order to select a proper recommendation approach in a domain, firstly knowledge about the users and the features of the items to be recommended is required. This knowledge is categorized in three groups of social, individual, and content (Burke & Ramezani, 2011) that are illustrated in Figure 2.5. Based on the availability and necessity of knowledge in different groups, an appropriate recommendation approach is followed which is described in the following. The Social group refers to the knowledge about the larger community of users other than the target user; individual group refers to the knowledge about the target user; and content group refers to the knowledge about items being recommended and about their uses.

There are varied recommendation approaches being used in different areas for personalized information retrieval (Jannach, Zanker, Felfernig, & Friedrich, 2010) (Ricci, Rokach, Shapira, & Kantor, 2011). Content-based filtering and collaborative filtering are the two major approaches that are broadly used in different cases. In the content-based filtering approach, on one side the knowledge about users' interests and preferences are collected explicitly or implicitly through asking their opinions by rating or reviews, or through analyzing their behavior, interactions, and history of activities. On the other side, the features and specifications of items are analyzed and annotated to be used in comparing with users' interests. The objective of this approach is to identify the similarities between the items and the given users' preferences and then recommend the items that better match the type of items that the given user would prefer (Pazzani & Billsus, 2007).

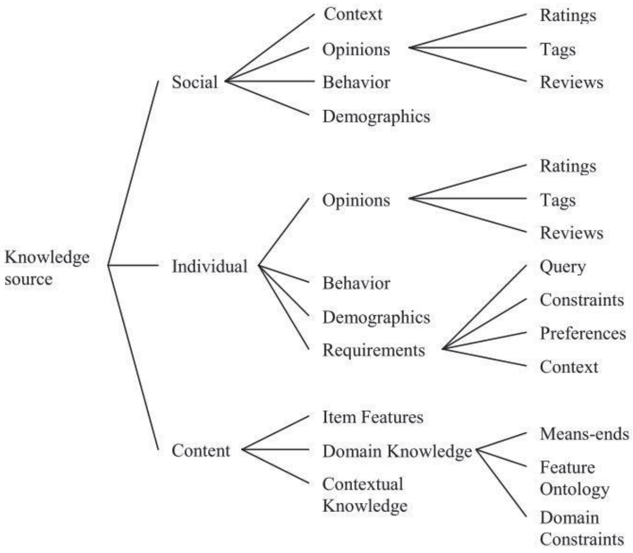


Figure 2.5. Taxonomy of knowledge sources in recommendation (Burke & Ramezani, 2011)

In the Collaborative filtering approach, the items are recommended to the users according to the opinions, interactions, and behaviors of other people. While the number of users is large, having large interconnected communities, this approach is a right choice for filtering the substantial quantities of data and personalize them for the given users (Schafer, Frankowski, Herlocker, & Sen, 2007).

Another recommendation approach that is inspired in this Ph.D. work is the context-aware recommendation. According to the definition in (Dey, 2001), context is defined as "Context is any information that can be used to characterize the situation of an entity. An entity is a person, place, or object that is considered relevant to the interaction between a user and an application, including the user and applications themselves."

Elements to the description of this context information fall into five categories: individuality, activity, location, time, and relations that are represented in Figure 2.6. The individuality category contains properties and attributes which describe the entity. The activity category contains any activities and tasks that the given entity is involved. The time and location categories provide the spatio-temporal coordinates of the given entity. And eventually, relations category refers to the information about any possible relationship that the given entity might have with other entities (Zimmermann, Lorenz, & Oppermann, 2007).

In context-aware recommendation approach, the emphasis is on the situation of a user in a specific domain. Any entity that is involved in the situation of the user is taken into consideration to provide the recommendations that match the respective situation (Adomavicius & Tuzhilin, 2011; Verbert et al., 2012).

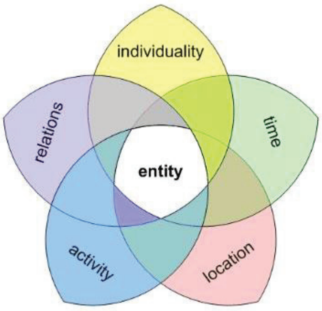


Figure 2.6. Five fundamental categories for context information (Zimmermann et al., 2007)

2.3.2 Content-based recommendation approach

2.3.2.1 Content-based Recommender systems framework

Content-based recommender systems have been vastly used in different areas and varied frameworks have been proposed to provide relevant recommendation for target users through content-based recommendation approach that mainly contains major components such as content analyzer, profile learner, and filtering component (Lops, De Gemmis, & Semeraro, 2011) that are represented in Figure 2.8.

- **Content analyzer**: At the first step, the items to be recommended need to be analyzed and processed to make them structured. As an example, if these items are unstructured documents, they first need to be structured through extracting some information such as metadata for them and organize this information for documents to be used later by next components that are profile learner and filtering component.
- **Profile learner**: The second step is to analyze what the user interests and preferences are, collect and generalize them, and create a user profile for each user. Machine learning techniques (Mitchell, 1997) are mostly used for generalization strategy. Using machine learning enables the system to infer a user model through the items that he or she liked/disliked in the past. This model is then used in recommending the items that are more likely relevant to the respective user.

- **Filtering component**: The task of the third step is to exploit user profiles and match them against the analyzed items and find the right items for recommendation. Some similarity metrics are used to identify the level of relevancy of items which will be explained later in this section. Finally, the results of filtering are ranked and potentially relevant items are represented to the given users.

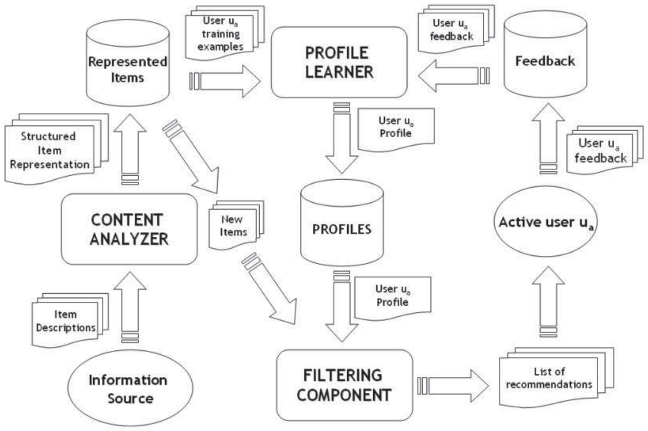


Figure 2.7. High level architecture of content-based recommender (Lops et al., 2011)

2.3.2.2 Item representation in content-based recommender systems

Each item to be recommended to users is associated with a set of properties and features that describe the key points of the respective item. These set of specific properties and attributes make the item structured and which improves finding and identification of relevant items easier during the recommendation process through matching users' interests collected in user profiles against attributes of structured items.

However, it should be considered that item properties are mainly textual and in some cases, they cause complications while learning user profiles and because of possible ambiguities in natural language. As the main challenge of traditional key-word based profiles, they lack semantically comparing users' interests with items since they perform matching process through string matching operation. String matching causes some problems such as polysemy that means the presence of multiple meaning of one word, and synonymy that means multiple words with the same meaning. Due to these problems, some relevant information can be missed since they might contain synonyms that are collected in user profiles, or according to polysemy wrong documents might be considered as relevant (Lops et al., 2011).

In the following, keyword-based approaches for document representation are described and then more advanced systems that apply semantic analysis through using ontologies as domain knowledge are described, respectively.

Key-word based vector space model

Keyword matching and Vector Space Model (VSM) are broadly used in content-based recommender systems as retrieval models. VSM is a geometric representation of text documents where each document is represented by a vector in an n-dimensional space and the dimensions are based on all document terms with associated weights (Turney & Pantel, 2010). In such these systems, any document is represented as a vector of terms. Each term in the vector contains a weight that represents its value and association to the given document. For calculating the weight of associated terms, VSM uses TF-IDF (Term Frequency-Inverse Document Frequency) which is a commonly used term weighting scheme. According to (Salton, 1989), multiple occurrences of a term in a document are not less relevant than single occurrences (TF assumption); rare terms are not less relevant than frequent terms (IDF assumption); and long documents are not preferred to short documents (normalization assumption). Similar to representing documents as weighted term vectors, user profiles are also represented as weighted terms vectors based on collected users' preferences. In order to predict how relevant a document is to a user's interests, cosine similarity is calculated between the document vector and user profile vector; the less value shows the more relevant document to be recommended to the respective user.

Varied keyword-based recommender systems have been developed in different areas of music, movie, e-commerce, new, etc. As the example of these systems, a user interface agent called Letizia is implemented that tracks user behavior and attempts to anticipate items that are interesting for target users. It builds personalized models that contain the keywords that describe users' interests (Lieberman, 1995). Syskill & Webert is another example which is a software agent that learns to rate pages on the web and determines the pages that might be interesting for users. The documents are represented by a number of informative words. The process of building user profiles is through analyzing the pages that have been rated by a user (Pazzani, Muramatsu, & Billsus, 1996). In the field of news recommendation, a personalized news system named YourNews is developed that maintains separate interest profiles main 8 topics and in each user profile, the interesting topics are represented as weighted prototype terms vectors that are extracted from the history of users' activity. It manages both short-term and long-term profiles by limiting the number of recent top-weighted extracted terms from the recent views as a short-term profile and considers all the past extracted terms as a long-term profile from all past views (Ahn, Brusilovsky, Grady, He, & Syn, 2007).

There are many other examples in different fields that can be found in (Lops et al., 2011) about traditional keyword-based recommender systems. The main lessons learned from these systems is that for getting promising recommendations that meet users' needs in this method, usually sufficient number of evidence of user interests should be available and actually they are syntactic pieces of evidence. In other words, a shortcoming with these systems is lack of intelligence (Lops et al., 2011). More advanced methods and strategies are required to augment content-based recommender systems to have semantic intelligence as well to identify the items that are relevant to a user, although they are not associated with explicit keywords that are collected in user profiles but semantically relevant to their interests.

In addition to the above approaches of identifying users' interests for personalization intention, broad analysis has been performed on users' activities on the web to investigate and model their behavior while browsing, searching, and interacting with different web pages. As the examples of this approach, a user modeling framework was developed for a news recommendation context (Twitter) and several advantageous of semantic entity-based, topic-based, and hashtag-based user modeling strategies through this framework (Abel, Gao, Houben, & Tao, 2011). It is reported that further enrichment with semantics extracted from news articles, improves the constructed user profiles and accuracy of recommendations.

In another work in similar context, Google News, users' history were investigated by tracking their past click behavior (J. Liu, Dolan, & Pedersen, 2010). A Bayesian framework was developed for predicting users' current news interests from the users' particular activities. In this work, content-based recommendation mechanism is combined with a collaborative filtering mechanism to generate personalized news recommendation which leads to improvement of the quality of news recommendation and traffic to the site. In another work, users' data on social media is studied to be aggregated in another system with the aim of addressing the cold-start problem. Based on a large aggregated dataset from the social web, distributed form-based and tag-based user profiles were studied. The consistency, completeness, and replication of form-based profiles, which are explicitly created by users in social web systems, were analyzed along with investigating tag-based profiles, which are the results of social tagging activities. Through developing and evaluating the performance of several cross-system user modeling strategies in the context of recommender systems, the evaluation results show that the proposed solution improve the challenge of cold-start problem and provides better recommendation quality (Abel et al., 2011).

Semantic analysis by using ontologies

In this part, the performance of ontologies as a semantic web technology are described and how they can bring semantics to recommender systems and improve their performance are elaborated through some examples of developed applications in different fields.

As the examples of the recommender systems that applied ontologies, an ontology-based method is presented in (Ge, Chen, Peng, & Li, 2012) for personalizing recommendation of knowledge in a heterogeneous environment to minimize repetitive retrieved information for users. A domain ontology is built by integrating multi-source and heterogeneous data and a user's interest ontology is generated through analyzing users' personal preferences and demographic characteristics. The developed recommender system can suggest proper information to the users that meet their interests through matching the results of a domain ontology, interest ontology, and user's query requests.

A recommendation system is proposed in e-commerce for a book recommendation. The architecture of the system which is depicted in Figure 2.9, classifies system tasks in two phases: ontology generation and recommendation. Ontology generation consists of constructing two types of ontologies: the general ontology that contains conceptual relations among documents and preference ontology to represent the weight of preference concepts. In order to recommend the books, the applied web robot collects web documents. Then, the documents are analyzed by the recommender system to define the relations between the concepts and properties that are stored in the general ontology. In the next step, the concept of the documents that are visited by users are analyzed and their weights get updated. The developed system provides relevant documents to users through identifying semantic relations between an ontology that semantically represents documents and user behavior history (Kang & Choi, 2011).

To recommend online academic research papers, two recommender systems called QuickStep and Foxtrot. Through monitoring users' behavior and activities, and relevance feedback, user profiles are created, and they are represented in terms of a research paper topic ontology. In this work, a hybrid recommendation approach is applied that consists of content-based recommendation and collaborative filtering recommendation approach. Research papers are classified by ontological classes and then the papers are recommended to the users who have a similar topic of interest by using collaborative recommendation. The performed experiments show that ontological inferencing improves user profiling. In addition, profile visualization and profile feedback have substantial role in profiling accuracy and recommendation process (Middleton, Shadbolt, & De Roure, 2004).

In addition to ontologies, for structuring and organizing data, another method of publishing structured data has been broadly used known as Linked Data. In this method, data can be interlinked and becomes more useful for semantic querying. These days, enterprise knowledge graph has emerged as known as one of the most useful applications in graph database technology that is based on a linked enterprise data approach for determining the relations between individual nodes of data and graphs. The synergy of semantic technologies with modern developments in artificial intelligence provides better opportunities for enterprises in managing knowledge and information in this context (Aasman, 2017). A knowledge graph is described in (Duan et al., 2017) as a graph that is constructed by representing each item, entity, and user as nodes and linking those nodes that interact with each other via edges and the architecture of knowledge graph has been clarified from data, information, knowledge, and wisdom aspects.

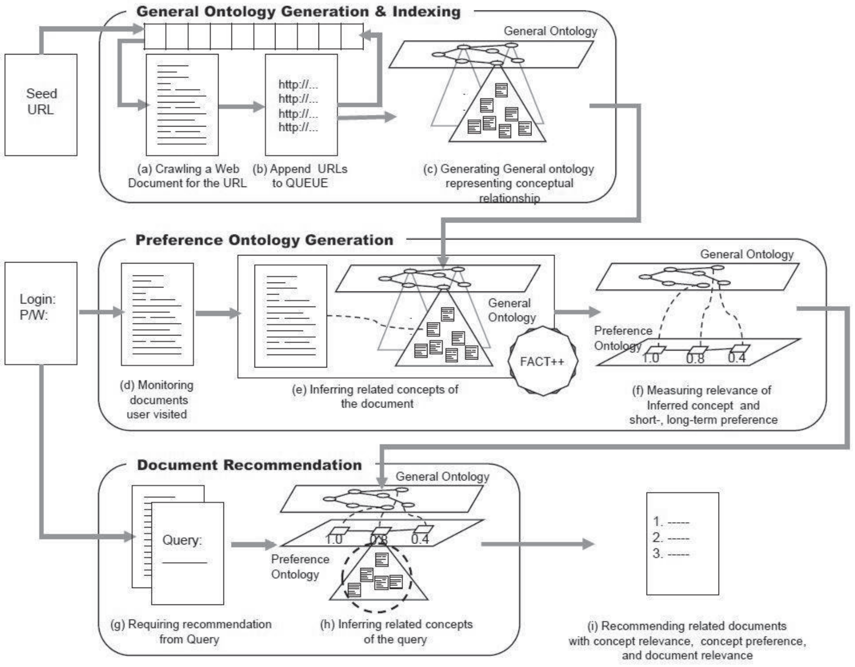

General Ontology Generation & Indexing
Seed URL
General Ontology
http://...
http://...
http://...
http://...
(a) Crawling a Web Document for the URL
(b) Append URLs to QUEUE
(c) Generating General ontology representing conceptual relationship
Preference Ontology Generation
Login: P/W:
General Ontology
General Ontology
Preference Ontology
1.0
0.8
0.4
FACT++
(d) Monitoring documents user visited
(e) Inferring related concepts of the document
(f) Measuring relevance of Inferred concept and short-, long-term preference
Document Recommendation
General Ontology
Query:
1.0
0.8
0.4
Preference Ontology
1. -----
2. -----
3. -----
(g) Requiring recommendation from Query
(h) Inferring related concepts of the query
(i) Recommending related documents with concept relevance, concept preference, and document relevance

Figure 2.8. A proposed architecture of an ontology-driven content-based recommender system (Kang & Choi, 2011)

As an example of using linked data, for identifying and measuring relatedness between resources, an approach is proposed in (Passant, 2010) to compute semantic distance on linked data by considering the existing links between resources. Several algorithms have been discussed for measuring resources and how they can be applied for recommending resources. As another example in recommender systems, linked data has been used to mitigate the challenges of new-user, new-item, and sparsity in collaborative filtering recommender systems (Heitmann & Hayes, 2010). Thorough decreasing data acquisition problem, linked data about object-centered sociality can be used to improve the mentioned challenges. As more examples, a systematic literature review of linked-data based recommender systems has been presented in (Figueroa, Vagliano, Rocha, & Morisio, 2015) describes the use of structured data published as linked data in recommender systems.

In conclusion, utilizing ontologies in recommender systems provide better and more accurate results compared to traditional recommender systems and improves the quality of recommendations.

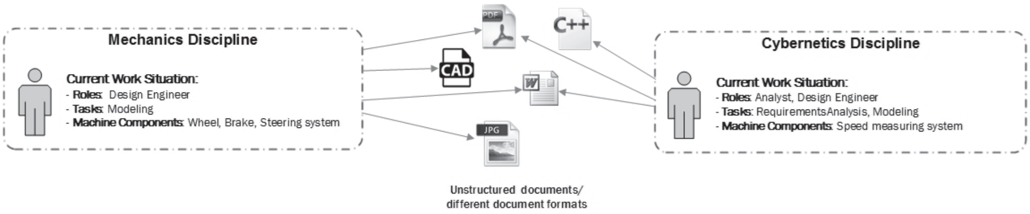
3 Research design

This chapter presents the research method during this Ph.D. work. At first, the use-case in the engineering domain is described and afterwards, the design research approach is elaborated to address the research questions. At each stage of the work, particular approaches are implemented to cover the research questions from the early phases of analysis to the late phases of development and evaluation.

3.1 Case study: DNV GL fuel fighter

The applied use-case is a multi-disciplinary engineering team at Norwegian University of Science and Technology, NTNU, that collaborate to design and develop energy-efficient vehicles. This use-case has many similarities with the challenges that engineers deal with in large-scale commercial projects such as heterogeneous data, unknown file structures, large amounts of data, personnel turnover, the need of reusing knowledge and information, knowledge access and sharing, and so on. The objectives of this project come from Shell Eco Marathon[1] that organizes an annual competition and encourages students around the world to design and develop ultra-energy-efficient vehicles. Since 2007, a group of students of engineering fields at NTNU has participated to this competition and work on a machine called DNV-GL fuel fighter[2] (Bøvre, Kyllo, Aalberg, Nordal, & Imaz Abal, 2014) (Buodd & Halsøy, 2015). The students are from different engineering backgrounds of mechanics, electronics, cybernetics, and others involved from design planning, materials, aerodynamics, and safety. As shown in Figure 3.1, each user could have one or more work context in the project and each context has specific information needs that are stored in different unstructured document formats.

Every year, a large number of documents are created and used containing knowledge and information about the process of developing the machine in different stages of the project, from analysis to test and evaluation. Transferring the existing knowledge from the past teams to the next teams has been a challenge and the new teams find it difficult to find, reuse, and share these valuable resources. It is very important to learn from the past teams in order to improve their work and make more innovative plans to provide better results. Having a knowledge access system is required to improve knowledge access, reuse, and share in such a context, save time spent on searching, and reduce the need to start from scratch. Figure 3.2 represents the designed vehicles in two racing and urban concept classes.



Mechanics Discipline

Current Work Situation:
- Roles: Design Engineer
- Tasks: Modeling
- Machine Components: Wheel, Brake, Steering system

Cybernetics Discipline

Current Work Situation:
- Roles: Analyst, Design Engineer
- Tasks: RequirementsAnalysis, Modeling
- Machine Components: Speed measuring system

Figure 3.1. Engineers in particular work contexts and their required documents

[1] . www.shell.com/energy-and-innovation/shell-ecomarathon.html

[2] . http://www.fuelfighter.no

Figure 3.2. DNV GL fuel fighter prototypes. The left side: Racing machine designed in 2014; The right side: Urban concept machine designed in 2014

3.2 Research Method

In order to conduct the research work, Design Science Research Methodology proposed by (Peffers et al., 2007) in information systems development is followed since it incorporates principles, practices, and procedures required to carry out such research. In addition, this methodology suggests three objectives: it is consistent with previous literature, it provides a nominal process model for doing design science research and it provides a mental model for presenting and evaluating design science research in information systems. Depending on the type of problem in a research, the research process can be sequential, or it can be started from different activities. Four possible research entry points are suggested in (Peffers et al., 2007) which is represented in Figure 3.2.

A problem-centered approach can be adopted if the idea for the research derived from observation of a problem. An objective-centered approach can be adopted if it is triggered by an industry or a research need that can be addressed by developing an artifact. A design and development-centered approach can be adopted if an existing artifact is not realized as a solution yet for an explicit problem domain. And another approach called client/context initiated solution that can base on observing a practical solution that worked and it can start with activity four.

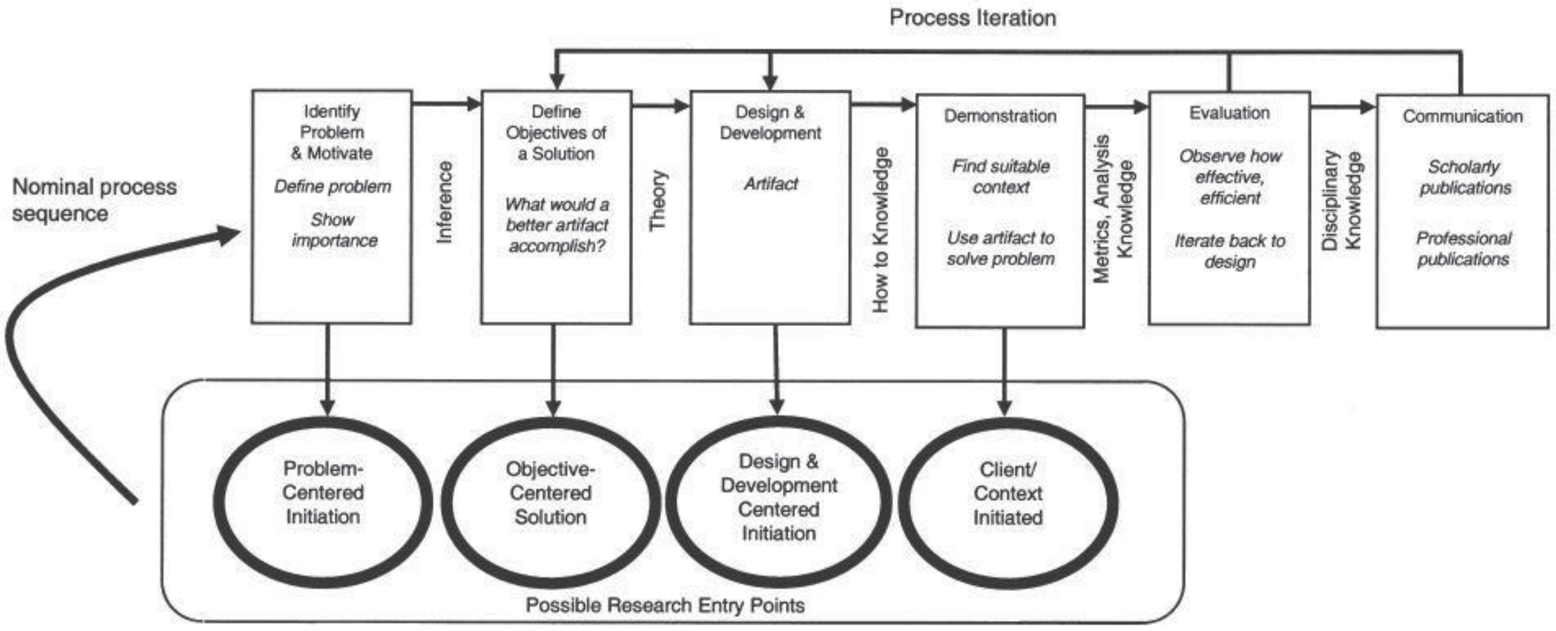

Process Iteration
Nominal process sequence
Identify Problem & Motivate
Define problem
Show importance
Inference
Define Objectives of a Solution
What would a better artifact accomplish?
Theory
Design & Development
Artifact
How to Knowledge
Demonstration
Find suitable context
Use artifact to solve problem
Metrics, Analysis Knowledge
Evaluation
Observe how effective, efficient
Iterate back to design
Disciplinary Knowledge
Communication
Scholarly publications
Professional publications
Problem-Centered Initiation
Objective-Centered Solution
Design & Development Centered Initiation
Client/ Context Initiated
Possible Research Entry Points

Figure 3.3. Design science research methodology (Peffers, Tuunanen, Rothenberger, & Chatterjee, 2007)

Since the focus of this Ph.D. research is on observing the case study and prototyping the proposed system, a problem-centered approach is selected. The suggested research process which is nominal sequence is followed and described in the activities in the following. These activities are also visualized in the graph represented in Figure 3.4 and are adapted to the thesis contributions and steps.

Activity 1: Problem identification and motivation: In this activity, a specific research problem is defined and the value of a solution is justified. The problem statement and the objective in this research are explained in sections 1.1 and 1.2 in more detail which is about the challenges of knowledge management in traditional engineering settings regarding knowledge access, sharing, and reuse. Problem definition is actually a basis for the artifact to be developed and to be efficient for providing a solution to the given problem. Required resources for this activity are literature studies of the state-of-the-art of research topics including knowledge management in enterprises, and related technologies such as recommender systems, ontologies, and information retrieval tools.

Activity 2: Objectives for solution: In this activity inferring the objectives of the solution from a problem definition and knowledge of what is feasible and possible are required. These objectives can be quantitative or qualitative. According to the methodology, they refer to a description of how a new artifact is expected to support solutions to the problem. Required resources include knowledge of the state of problems and current solutions if any. In this research, the introduced use-case is studied and analyzed in detail from different aspects of involved people, their work context, features, and specifications of existing knowledge sources and applied knowledge repositories through organizing semi-structured and unstructured interviews with the people from different engineering disciplines. Their challenges in knowledge access and sharing and the current solutions are identified.

Activity 3: Design and development: In this activity, the artifact will be created in the form of a model, a construct, a method, or an instantiation. Determining the desired functionalities for the artifact and its architecture are included in this activity.

In this research, the identified applied technologies and tools in similar works are analyzed and their synergies are investigated for structuring the target solution. Then, the architecture of the tailored system is proposed including the essential required components.

Activity 4: Demonstration: In this activity, the use of the artifact is demonstrated to solve one or more instances of a problem that could involve experimentation, simulation, case study, proof, or other related activities. In this research, the archived unstructured documents are indexed and became searchable; the ontology as domain knowledge of the recommender system is constructed through analyzing people's work context and the engineering domain. Then, the built ontology is validated against indexed documents to validate retrievability of documents by ontology concepts. The recommender system prototype is developed in two phases of initial system and revised system by integrating varied approaches of recommendation, explicit feedback functionality, the constructed domain-specific ontology, and advanced information retrieval tools. These activities are described in detail in the journal publications.

Activity 5: Evaluation: In this activity, the artifact is evaluated and its performance is measured to see how well it supports a solution to the problem. The objectives of the solutions are compared with the actual observed results. Here, knowledge of relevant metrics and analysis techniques is required. This process can be iterative to improve the artifact and is dependent to the nature of the artifact. In this research, the developed system is evaluated in two phases of the initial and revised system through testing its performance by a group of involved people, students as engineers, in the use-case. Planned qualitative and quantitative evaluations are done by using questionnaires and analysis of logged data of users' explicit feedback respectively.

Activity 6: Communication: Communicate the problem and its importance, the artifact, its utility and novelty, and its effectiveness to researchers and other relevant audiences. Regarding this activity, 4 articles have been published and submitted in relevant conferences and journals with the focus of knowledge management and information systems. A list of papers is represented in chapter 1 along with the supplementary published articles.

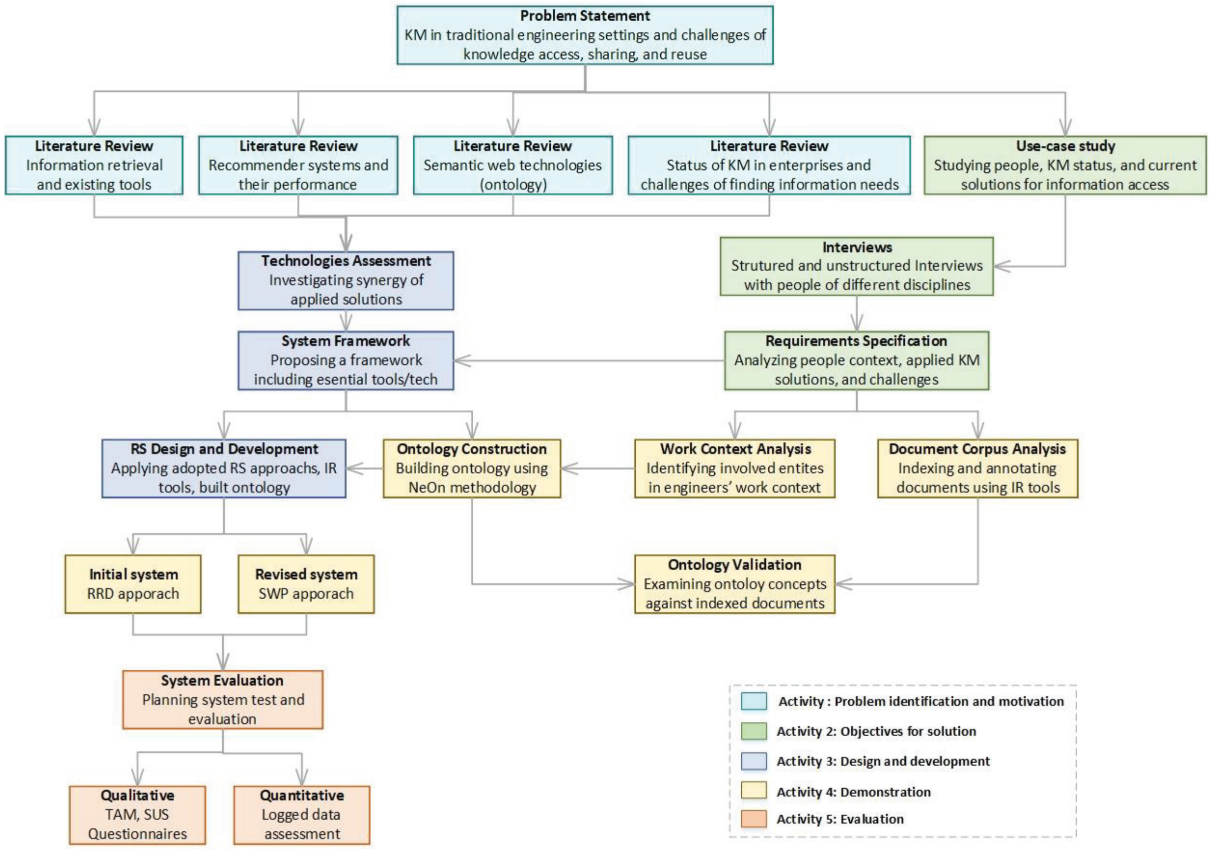

Problem Statement
KM in traditional engineering settings and challenges of knowledge access, sharing, and reuse
Literature Review
Information retrieval and existing tools
Literature Review
Recommender systems and their performance
Literature Review
Semantic web technologies (ontology)
Literature Review
Status of KM in enterprises and challenges of finding information needs
Use-case study
Studying people, KM status, and current solutions for information access
Technologies Assessment
Investigating synergy of applied solutions
Interviews
Strutured and unstructured Interviews with people of different disciplines
System Framework
Proposing a framework including esential tools/tech
Requirements Specification
Analyzing people context, applied KM solutions, and challenges
RS Design and Development
Applying adopted RS approachs, IR tools, built ontology
Ontology Construction
Building ontology using NeOn methodology
Work Context Analysis
Identifying involved entites in engineers' work context
Document Corpus Analysis
Indexing and annotating documents using IR tools
Initial system
RRD apporach
Revised system
SWP apporach
Ontology Validation
Examining ontoloy concepts against indexed documents
System Evaluation
Planning system test and evaluation
Qualitative
TAM, SUS Questionnaires
Quantitative
Logged data assessment
Activity : Problem identification and motivation
Activity 2: Objectives for solution
Activity 3: Design and development
Activity 4: Demonstration
Activity 5: Evaluation

Figure 3.4. A high-level research method graph of the PhD work

4 Approach and results

This chapter gives an overview of the results of the research conducted in this thesis. For each research paper, its relevance to the thesis, the contribution, approach, and results are briefly described and addressed in the following. The actual papers can be found in the Appendices.

4.1 Paper I: Intelligent Services: A Semantic Recommender System for Knowledge Representation in Industry

Abstract: *Intense competition in industrial area pushes companies to increase pace and efficiency of development. During process plant design projects, large amount of information causes a challenge in stakeholders' collaboration in decision making and it breeds lower development speed. An intelligent service is required to improve knowledge and information accessibility by personalizing the knowledge and information based on the stakeholder's situation in their working life which is known as a recommender system. This paper describes the early phases of a PhD project that explores the idea of applying a semantic recommender system in process plant design. To achieve this goal we aim to employ various recommendation approaches, data analysis and ontology engineering. The resource of data is provided by an industrial partner, Aker Solutions. The results of discussion show that similar to the way recommender systems personalize information in Web search, it is also feasible to develop an ontology-based recommender system for industry to explore the most relevant explicit and implicit knowledge and information for a given stakeholder.*

Relevance to the thesis: This paper discusses the problem of knowledge and information access and sharing in enterprises and gives a state of the art research on the applied technologies for improving the discussed challenges. This paper addresses RQ1 and RQ2 and proposes the initial idea of applying recommender systems in a professional context for knowledge access improvement.

Approach and contributions: A broad literature review was conducted on the usage of recommender systems and how they are applied in different fields such as e-commerce and web search to improve relevant recommendation and personalization to the given users of varied domains. The objective is to investigate the performance of recommender systems and how applicable they are in a professional context such as enterprises and manufacturing workplaces to improve the challenge of information overload. Since ontologies have been utilized in recommender systems as an improvement tool for identifying relevant items, a literature view is conducted on the ontology concept and their specifications and contribution in recent recommender systems. Besides commonly used recommendation approaches, context-driven recommendation approach is considered to investigate the involved entities that characterize the professional situation of people in enterprises.

Results: This paper presents the understanding of applied tools and technologies for improving knowledge and proposes the general idea of how applying a recommender system along with a domain-specific ontology can be beneficial in providing relevant documents to the knowledge workers. In addition, a broad literature review has been performed on key aspects that define the context of people in a professional setting. People's interactions for accessing their information needs in an engineering context are investigated and analysis results are collected from different views of historical and personal information, collaborative, and environmental situation presented as a domain model. Each view indicates a group of entities and informative relations between them. In addition, the components of the planned prototype system have been mentioned in general as future work.

4.2 Paper II: Using Process Ontologies to Contextualize Recommender Systems in Engineering Projects for Knowledge Access Improvement

Abstract: *Knowledge and information are highly important resources in today's knowledge economy and vital in achieving organization's goals. Particularly in engineering projects, users' tasks are highly dependent on accessing, using, and reusing these resources and users already spend a lot of time searching for relevant knowledge. As the total volume of documents across different sources and repositories increases, users face additional overhead related to search and retrieval. Knowledge workers across multiple disciplines experience fierce competition and a persistent pressure to deliver value-added contributions in a competitive global business environment with complex, multidisciplinary problems. Simple search engines are often not sufficient since they are not designed to retrieve those relevant documents that match the user's current work situation and information need. Therefore, the need for a semantics-based solution has been identified. This paper describes the early stages of a PhD project that proposes a tailored recommender system for improving knowledge accessibility in an engineering setting. The recommender system will be developed for and validated in a multidisciplinary engineering project as a case study. We take advantage of content-based filtering and collaborative filtering along with semantic technologies to provide relevant and accurate recommendations. In order to contextualize users' work situations during a project development process, the recommender system utilizes a tailored process ontology to be able to explore different dimensions of users' situations. By merging the concepts derived from the ontology, the current work situation of a given user is identified and varied fine-grained user profiles will be created at real time, so-called dynamic user profiles. Therefore, the recommender system is able to set the scope of users' interests to the exact level that a user desires in the current situation. To classify the identified relevant documents, we propose creating concept profiles that originate from process ontology concepts for further recommendation according to a collaborative approach. This paper describes the recommender system components, proposes a framework for the target recommender system, and discusses how its components are integrated and interact in order to improve information access in engineering projects.*

Relevance to the thesis: In this paper, a high-level framework is presented for the proposed system following the initial idea proposed in P1, and the essential components for creating the proposed system have been described. This paper briefly addresses RQ4 and contributes towards C1 and C2. In paper 3, the research question and contributions are thoroughly addressed in detail.

Approach and contributions: This paper investigates the shortcomings of the current solutions for finding relevant documents in a professional context in more detail and starts to structure a system that aims to improve these challenges. In particular, the use-case is briefly described in terms of the status of engineers, knowledge resources, and current ways that engineers apply for accessing knowledge to assign their tasks. According to the specifications of these concepts and the results of studies, the framework of a tailored recommender system is presented along with the essential components that need to be considered in the target system.

Results: In this paper, a structured framework is proposed for creating the planned recommender system. Essential components that are required for the target system are described that are the built ontology as knowledge domain for contextualizing users' work situations, static and dynamic user profiles, recommendation approaches, concept profiles derived from the constructed ontology, indexer to manage documents for search, and feedback collection approach.

Regarding the profiling approach, the creation of varied fine-grained user profiles has been proposed by combining ontology concepts that characterize users' information needs. The aim of this functionality is to enable users to narrow down their scope of information needs to find those documents that are closer to their current work situation. This may lead to spending less time in the information retrieval process and more accurate results.

4.3 Paper III: Investigating contextual ontologies and document corpus characteristics for information access in engineering settings

Abstract: *Knowledge and information resources play a pivotal role in enterprises and are valuable for solution reuse and learning through information access. However, identifying relevant information from a rapidly growing number of unstructured resources is challenging for users. We discuss a personalized information access tool for professional workplaces based on the recommender systems to provide relevant documents for users in specific work contexts based on domain-specific ontologies. Our use case is a multidisciplinary engineering project building an energy-efficient vehicle. We provide an in-depth analysis of document corpus characteristics of this real-life shared engineering workspace to understand the content and context of documents using information retrieval methods and semantic annotations. Upon this, we build a contextual ontology as our knowledge domain for the recommender system. We validate our ontology-based content matching approach by evaluating the level of retrievability and coverage of the ontology against the indexed document corpus through experiments on the corpus and ontology. Our results provide insight into engineers' document workspaces and show that even a simple domain ontology is able to match a majority of documents from a domain-oriented corpus. The findings support our approach of using ontology-based recommendation for domain-specific workspaces.*

Relevance to the thesis: The focus of this paper is detailed analysis of document corpus, construction of ontology, and evaluation of retrievability and coverage of the ontology against the document corpus. In this paper, RQ3 is completely addressed, along with C1 and C3.

Approach and contributions: This paper is a detailed version of the second publication. This paper consists of three main contributions. First, the live, in-the-field document corpus has been analyzed in depth from real projects to understand the structure of existing documents in the engineering context. Second, the requirements and specifications of users have been investigated through semi-structured interviews with engineers of different disciplines. The focus of the interviews was to identify users' regular tasks, their information needs and information seeking challenges. Third, an ontology as the knowledge domain has been constructed, using NeOn methodology, and its use and suitability have been verified for the engineering domain by validating it against the document corpus. In order to assess the retrievability and coverage of ontology concepts against the documents, the ontology-based content matching approach has been examined with an advanced search tool to measure the performance of the created knowledge domain.

Results: The contributions described above provided the following results. In-depth analysis of document corpus provided a deep insight on the existing documents and their specifications. Documents have been grouped in two groups of textual and non-textual based on their formats. The number of non-textual documents has been significantly more than textual documents and images have a large proportion in the non-textual group. Regarding retrievability and coverage of the constructed ontology against the document corpus, the results show that the provided knowledge domain could cover a majority of documents, with up to 95% for textual and still 43% for non-textual files. Regarding non-textual documents, additional metadata and annotations could be a complementary solution to the proposed approach to improve information access and retrievability. In addition, the results indicate a cold-start problem for some concepts that are either too wide or too narrow according to the level of matching with the document corpus. This causes retrieving too many or too few documents. However, it proves valuable insight to refine the matching towards better specificity for certain concepts based on the document corpus, especially for very specific concepts.

In addition, it is examined and validated that the approach of building a domain-specific ontology for the engineering domain for a semantically improved recommender system is suitable for a typical scenario that collects a huge number of varied documents in a shared file system for an engineering project; and this approach forms a suitable basis for the proposed knowledge access system for the engineering case.

4.4 Paper IV: Development and Evaluation of a Knowledge Access System for Engineering Workspaces Based on Recommendation and Filtering

Abstract: *The research presented in this paper is a follow-up of our prior works involving evaluation of the developed knowledge access system prototype, named ProRecSys. The aim is to examine ProRecSys performance in terms of recommendation and filtering in our engineering case-study to provide relevant documents for users in specific work contexts using domain-specific ontologies. We start with describing the process-driven system architecture and the logic behind interaction of processes and information retrieval strategies. We follow a content-based recommendation approach to identify relevant documents to users' work context by proposing a novel approach for profiling at ontology concepts level. The concept of ProRecSys has been validated in an engineering context by evaluating the system in two stages. Firstly, evaluating the initial system that filters documents based on explicit terms of work context derived from our ontology and secondly, the revised system*

that works with the collected content of concept profiles that are initialized with a sort of significant terms and strengthened with the results of users' explicit feedback. In general, the results indicate that ProRecSys is potentially an appropriate tool for enhancing knowledge sharing and reuse in traditional engineering projects. We present promising experimental results by comparing the performance of two approaches in initial and revised systems. The results show that collecting semantic terms in concept profiles could provide recommendations with higher quality that better fit users' information needs. At the end, future research directions for improving ProRecSys performance are suggested.

Relevance to the thesis: This paper represents the process of developing and evaluating the proposed system following the results gained in the former publications. RQ5 and RQ6 are addressed, and it contributes towards C4.

Approach and contributions: This paper follows the finding and achieved results during the process of developing the target recommender system. System development is elaborated from a technical aspect and describes in detail the two phases of developing ProRecSys, the initial system, and revised system. Both qualitative and quantitative approaches of evaluation are applied to examine and evaluate the performance of ProRecSys in the engineering use-case.

Results: The qualitative and quantitative evaluation on the performance of two approaches in initial and revised systems reports that the recommendations derived from SWP approach could provide more relevant results in earlier hits for users in their specific work contexts compared to RRD approach. This concludes that the content of concept profiles have been more successful in scoping users' information needs and identifying the documents that meet their preferences.

Applying a system with ProRecSys capabilities is valuable for users with different levels of expertise since it collects all the existing documents, organizes and analyzes them through indexing and make a huge number of documents structured and accessible through different approaches. Note that the role of such a system is stronger at the beginning of a project when users want to start working on a task and need to collect some information to learn about the task, finding some facts and instructions, finding examples to reuse (best practices) and similar past works. Therefore, their search scope is wider for investigating retrieved information and recommendations are helpful to reach them. As the project proceeds, the users need more specific information and tend to try a direct search for specific pieces of information using a specified search system rather than a recommender system. Mainly, recommender systems are more useful for novices and less for experienced engineers since they better know what they want and where to find it in a specific domain.

5 Conclusion

In this chapter, the concept of a recommender system in workplaces is discussed in more detail together with the features and specifications that need to be considered for developing such a system. The future work to improve this research work is also suggested in the following.

5.1 Discussion

In this Ph.D. research, the main objective is to study the challenge of information overload in enterprises and the difficulties that engineers are faced with for accessing to their required knowledge and information to perform their assigned tasks. Similar works both in e-commerce and enterprises, applied approaches, technologies, and tools have acted as sources of inspiration for improving knowledge access, sharing, and reuse in an engineering context. A prototype system has been developed, ProRecSys, to improve identifying relevant items to the right people in a respective domain.

The results of examining the prototype system show that compared to the existing solutions that users apply for finding their information needs in the studied use case, the performance of ProRecSys is promising in exploring documents that are close and relevant to users' work context. The role of such a system is stronger at the beginning of a project (Journal paper submitted to KAIS on 13.02.2018) when users want to start working on a task and need to collect some information to learn about the task, finding some facts and instructions, finding examples to reuse (best practices), and similar past works. Therefore, their search scope is wider for investigating retrieved information, and recommendation results are helpful to reach them. As the project proceeds, the users need more specific information and tend to try a direct search for a specific piece of information using a specified search system rather than a recommender system. Mainly, recommender systems might be more useful for novices and beginners and less for experienced engineers since they better aware what they want and where to find it in a specific domain.

Another interesting point of this research is the usage of an ontology as knowledge domain to compose varied concepts and build dynamic work profiles for identifying relevant recommendations (Mehrpoor et al., 2017; Mehrpoor et al., 2015). It enables the system to be applicable to other domains for exploring relevant documents. The ontology can be extended, modified, or re-designed to match the specifications of the target domain and then ProRecSys works as explained in the case study. Another advantage of using ontology is to use the relations between its entities for identifying new semantically relevant articles that might not be found by recommendation techniques and text-mining. In order to strengthen and improve the collected terms for ontology concepts, the content of concept profiles can be used to link semantically relevant terms to each concept and expand the built ontology for better query results from the preliminary phases of system life cycle for next versions of the system. Note that, over time, the concept profiles might get closer and closer to each other. As a solution, following an approach employed in news recommendation (Gulla, Fidjestøl, Su, & Castejon, 2014), the system should keep the feedback results of the latest documents that were relevant to a work context by versioning the indexes that carry the content of concept profiles to manage and classify relevant document contents in different periods of time. Using such an approach can help investigate the latest ranked documents. In some cases, it might be a better approximation of the user's work context and interests.

Overall, the developed system is a useful tool for improving information access in engineering cases to save spent time for searching the required information. However, system functionality and features can be improved in further iterations and examinations to optimize its performance to provide closer recommendations to work contexts and increase the number of relevant documents in the early hits.

5.2 Generalization and Adaptation

ProRecSys has been developed and examined in a small engineering use case. However, the approach of its performance has been designed and developed in a way to be generalized and adapted to other application domains as well. There are general principles and limitations in the current system that need to be considered for developing such a system. The idea of having a context-driven ontology is useful for scoping users' work context and search for the documents that meet their specified work contexts. The constructed ontology for the prototype recommender system is based on the studied use-case and the identified dimensions for characterizing users' work contexts can be different or more expanded in other domains. Therefore, key dimensions need to be selected if the application domain requires investigating more aspects of users' contexts for building the ontology.

Similar to the limitation for dimensions of work context used in creating the ontology, the specified metadata for the documents can be expanded. Depending on the key information that is found important for target users and document specifications, additional metadata may be selected and used during document indexing process.

The results of evaluating the two approaches of RRD and SWP show that SWP was more successful in identifying relevant documents that meet users' information needs. This means that the collected terms in ontology concept profiles can be used to improve the semantic definition of ontology concepts and consequently improve information retrieval using **ontology concepts.**

Implementation approach and required settings

As an advantage of the proposed system, ProRecSys is designed to be easily extended and applied in other use cases. There are two resources that need to be prepared to make the system work in different contexts.

Firstly, the knowledge domain, which is the contextual ontology, need to be reviewed and customized for the involved concepts. Adding, editing, or removing any concept requires a person who is familiar with RDF framework. In the current version of the ontology, the relations between concepts are simple and limited to e.g. “concept B is sub-class-of concept A” which can be easily applied using a simple text editor when the relations and hierarchy of concepts are determined.

Secondly, the archived documents and ontology concepts need to be indexed using Elasticsearch[1]. The indexing process is embedded in ProRecSys as well. It is only required to specify where the root directory is, and all the existing documents will be indexed under that location. Note that an Elasticsearch instance needs to be running on the system that performs the indexing process.

While the ontology is ready, there is a preparation stage to index documents and concept profiles using administration panel which is available in the system. After document and concept profiles are indexed, in case any initialization of concept profiles is desired for the given use case, it can be performed through the system. In addition, it is required to initialize some values in the configuration file that are mentioned below:

- The path of RDF file which contains the ontology concepts
- The index names for documents and concept profiles

As described, ProRecSys is flexible to be applied and examined in different contexts that apply file systems for managing documents. All the process of identifying users’ work profile and filtering the documents based on that, is the same as what is discussed mostly in (Journal paper submitted to KAIS on 13.02.2018).

5.3 Future work

As future work, extending the functionalities of the developed system is suggested and examining the system in similar professional contexts to better evaluate the system performance and improve it in further iterations. The following improvements are suggested:

- Improving the built ontology using the collected terms in concept profiles for each concept and define semantic and informative relations between ontology concepts to improve recommendation process and include inference and reasoning and consequently identify more diverse recommendations.
- Improving the solution of recommending non-textual documents through collaborative filtering recommendation approach
- Examining the system with an ontology in another domain to evaluate the extendibility of the developed system.
- Examining a multi-lingual ontology is some cases that documents might be created on different languages and not only on a single language.

[1] . www.elastic.co/products/elasticsearch

- Investigating any other aspects that characterize users' work contexts and could be a lead to identify which type of documents can be more relevant to users' information needs.

Overall, the rapid increase of data growth, particularly in professional workplaces, is critical and demands to apply new tools and technologies to control and manage these valuable resources. Broad research and studies on recommender systems have proven that they are very useful tools for identifying pieces of information that are relevant to different task types and therefore can meet users' information needs. The synergy of recommender systems and ontologies improves the performance of recommender systems by including semantics, reasoning, and inferencing to the approach and methodologies that a recommender system follows for identifying the relevant documents out of a large number of documents. Recommender systems have a very important role in early stages of projects where there are many existing resources that are reusable but due to inappropriate management strategies, it is very hard and challenging to explore them. Identifying reusable knowledge and information saves significant time and money that needs to be spent for finding or regenerating them.

6 References

Aasman, J. (2017). Transmuting Information to Knowledge with an Enterprise Knowledge Graph. *IT Professional*(6), 44-51.

Abecker, A., & van Elst, L. (2009). Ontologies for knowledge management *Handbook on ontologies* (pp. 713-734): Springer.

Abel, F., Gao, Q., Houben, G.-J., & Tao, K. (2011). Analyzing user modeling on twitter for personalized news recommendations. *User Modeling, Adaption and Personalization*, 1-12.

Adomavicius, G., & Tuzhilin, A. (2011). Context-aware recommender systems *Recommender systems handbook* (pp. 217-253): Springer.

Ahlers, D., & Mehrpoor, M. (2015). *Everything is Filed under'File': Conceptual Challenges in Applying Semantic Search to Network Shares for Collaborative Work.* Proceedings of the 26th ACM Conference on Hypertext & Social Media. Guzelyurt, Northern Cyprus.

Ahlers, D., Mehrpoor, M., Kristensen, K., & Krogstie, J. (2015). *Challenges for information access in multi-disciplinary product design and engineering settings.* Tenth International Conference on Digital Information Management (ICDIM), 2015. Jeju, South Korea.

Ahn, J.-w., Brusilovsky, P., Grady, J., He, D., & Syn, S. Y. (2007). *Open user profiles for adaptive news systems: help or harm?* Proceedings of the 16th international conference on World Wide Web. New York, USA.

Alani, H., Kim, S., Millard, D. E., Weal, M. J., Hall, W., Lewis, P. H., & Shadbolt, N. R. (2003). Automatic ontology-based knowledge extraction from web documents. *IEEE Intelligent Systems, 18*(1), 14-21.

Berners-Lee, T., Hendler, J., & Lassila, O. (2001). The semantic web. *Scientific american, 284*(5), 28-37.

Botha, A. P. (2008). *Knowledge: living and working with it*: Juta and Company Ltd.

Bøvre, M., Kyllo, M., Aalberg, J., Nordal, I. O. S., & Imaz Abal, L. (2014). DNV GL Fuel Fighter: Development and Construction of Vehicles for Participation in Shell Eco-marathon 2014. Master thesis: NTNU, Trondheim.

Brown, J. S., & Duguid, P. (1998). Organizing knowledge. *California management review, 40*(3), 90-111.

Bruno, G. (2015). *Measuring product semantic similarity by exploiting a manufacturing process ontology Measuring product semantic similarity by exploiting a manufacturing process ontology.* Industrial Engineering and Systems Management (IESM), 2015. Seville, Spain.

Buodd, M., & Halsøy, B. (2015). DNV GL Fuel Fighter towards Shell Eco-marathon 2015. Master thesis: NTNU, Trondheim.

Burke, R. D., & Ramezani, M. (2011). Matching Recommendation Technologies and Domains chapter 11. *Recommender systems handbook, 1*, 367.

Byström, K., & Järvelin, K. (1995). Task complexity affects information seeking and use. *Information Processing & Management, 31*(2), 191-213.

Chang, X., Sahin, A., & Terpenny, J. (2008). An ontology-based support for product conceptual design. *Robotics and Computer-Integrated Manufacturing, 24*(6), 755-762.

Cole, M. J., Gwizdka, J., Liu, C., Bierig, R., Belkin, N. J., & Zhang, X. (2011). Task and user effects on reading patterns in information search. *Interacting with Computers, 23*(4), 346-362.

Davenport, T. H., & Prusak, L. (1998). *Working knowledge: How organizations manage what they know*: Harvard Business Press.

Davies, J., Fensel, D., & Van Harmelen, F. (2003). *Towards the semantic web: ontology-driven knowledge management*: John Wiley & Sons.

De Boer, R. C., Lago, P., Telea, A., & Van Vliet, H. (2009). *Ontology-driven visualization of architectural design decisions.* Software Architecture, 2009 & European Conference on Software Architecture. WICSA/ECSA 2009. Joint Working IEEE/IFIP Conference. Cambridge, UK.

Denkena, B., Shpitalni, M., Kowalski, P., Molcho, G., & Zipori, Y. (2007). Knowledge management in process planning. *CIRP Annals-Manufacturing Technology, 56*(1), 175-180.

Dey, A. K. (2001). Understanding and using context. *Personal and ubiquitous computing, 5*(1), 4-7.

Domingue, J., Fensel, D., & Hendler, J. A. (2011). Introduction to the semantic web technologies *Handbook of Semantic Web Technologies* (pp. 1-41): Springer.

Duan, Y., Shao, L., Hu, G., Zhou, Z., Zou, Q., & Lin, Z. (2017). *Specifying architecture of knowledge graph with data graph, information graph, knowledge graph and wisdom graph.* Software Engineering Research, Management and Applications (SERA), 2017 IEEE 15th International Conference. London, UK.

Duke, D. J., Brodlie, K. W., & Duce, D. A. (2004). *Building an ontology of visualization.* Visualization, 2004. IEEE. Austin, TX, USA.

Eck, O., & Schaefer, D. (2011). A semantic file system for integrated product data management. *Advanced engineering informatics, 25*(2), 177-184.

Fensel, D. (2002). Ontology-based knowledge management. *Computer, 35*(11), 56-59.

Figueroa, C., Vagliano, I., Rocha, O. R., & Morisio, M. (2015). A systematic literature review of Linked Data-based recommender systems. *Concurrency and Computation: Practice and Experience, 27*(17), 4659-4684.

Freund, L. S. (2008). *Exploiting Task-Document Relations in Support of Information Retrieval in the Workplace.*

Ge, J., Chen, Z., Peng, J., & Li, T. (2012). *An ontology-based method for personalized recommendation.* Cognitive Informatics & Cognitive Computing (ICCI* CC), 2012 IEEE 11th International Conference. Kyoto, Japan.

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

Gulla, J. A., Fidjestøl, A. D., Su, X., & Castejon, H. (2014). *Implicit User Profiling in News Recommender Systems (sigterm and qualitative eval).* WEBIST (1). Barcelona, Spain.

Ha, I., Oh, K.-J., Hong, M.-D., Lee, Y.-H., Rosli, A. N., & Jo, G.-S. (2014). Ontology-driven visualization system for semantic searching. *Multimedia Tools and Applications, 71*(2), 947-965.

Hansen, M. (2009). Collaboration: How Leaders Avoid the Traps, Create Unity, and Reap Big Results: Harvard Business School Press.

Hansen, M. T., & Nohria, N. (2004). How to build collaborative advantage. *MIT Sloan Management Review, 46*(1), 22.

Hansen, P. (2011). Task-based information Seeking and Retrieval in the Patent Domain. *Processes and Relationships. Academic Dissertation, University of Tampere*.

Heitmann, B., & Hayes, C. (2010). *Using Linked Data to Build Open, Collaborative Recommender Systems.* AAAI spring symposium: linked data meets artificial intelligence.

Hertzum, M., & Pejtersen, A. M. (2000). The information-seeking practices of engineers: searching for documents as well as for people. *Information Processing & Management, 36*(5), 761-778.

Horn, B. L. (2016). Computer system for automatic organization, indexing and viewing of information from multiple sources: Google Patents.

Huang, N., & Diao, S. (2008). Ontology-based enterprise knowledge integration. *Robotics and Computer-Integrated Manufacturing, 24*(4), 562-571.

Jannach, D., Zanker, M., Felfernig, A., & Friedrich, G. (2010). *Recommender systems: an introduction*: Cambridge University Press.

Kang, J., & Choi, J. (2011). *An ontology-based recommendation system using long-term and short-term preferences.* Information Science and Applications (ICISA), 2011 International Conference. Jeju Island, South Korea.

King, D. W. (1994). Communication by Engineers: A Literature Review of Engineers' Information Needs, Seeking Processes, and Use. Sponsor**:** Council on Library Resources, Inc., Washington, DC.

Kristensen, K., Krogstie, J., Ahlers, D., & Mehrpoor, M. (2016). LEAP Collaboration System. *Taking the LEAP: The Methods and Tools of the Linked Engineering and Manufacturing Platform (LEAP)*, 99-124.

Le Duigou, J., Bernard, A., Perry, N., & Delplace, J.-C. (2012). Generic PLM system for SMEs: Application to an equipment manufacturer. *International Journal of Product Lifecycle Management 7, 6*(1), 51-64.

Li, Z., Raskin, V., & Ramani, K. (2007). *Developing ontologies for engineering information retrieval.* ASME 2007 International Design Engineering Technical Conferences and Computers and Information in Engineering Conference. Las Vegas, Nevada, USA

Lieberman, H. (1995). Letizia: An agent that assists web browsing. *IJCAI (1), 1995*, 924-929.

Lim, S. C. J., Liu, Y., & Lee, W. B. (2009). *Faceted search and retrieval based on semantically annotated product family ontology.* Proceedings of the WSDM'09 Workshop on Exploiting Semantic Annotations in Information Retrieval. Barcelona, Spain.

Liu, C., Liu, J., Cole, M., Belkin, N. J., & Zhang, X. (2012). Task difficulty and domain knowledge effects on information search behaviors. *Proceedings of the American Society for Information Science and Technology, 49*(1), 1-10.

Liu, J., Cole, M. J., Liu, C., Bierig, R., Gwizdka, J., Belkin, N. J., Zhang, X. (2010). *Search behaviors in different task types*. Proceedings of the 10th annual joint conference on Digital libraries, Gold Coast, Queensland, Australia.

Liu, J., Dolan, P., & Pedersen, E. R. (2010). *Personalized news recommendation based on click behavior.* Proceedings of the 15th international conference on Intelligent user interfaces. Hong Kong, China.

Lops, P., De Gemmis, M., & Semeraro, G. (2011). Content-based recommender systems: State of the art and trends *Recommender systems handbook* (pp. 73-105): Springer.

Louis-Sidney, L., Cheutet, V., Lamouri, S., Puron, O., & Mezza, A. (2012). A conceptual model for the implementation of an Inter-Knowledge Objects Exchange System (IKOES) in automotive industry. *Engineering Applications of Artificial Intelligence, 25*(5), 1090-1101.

Lowe, A., McMahon, C., & Culley, S. (2004). Information access, storage and use by engineering designers, part 1. *The Journal of the Institution of Engineering Designers, 30*(2), 30-32.

Manning, C. D., Raghavan, P., & Schütze, H. (2008). *Introduction to information retrieval* (Vol. 1): Cambridge university press Cambridge.

McGuinness, D. L. (1998). *Ontological issues for knowledge-enhanced search.* Proceedings of Formal Ontology in Information Systems. p. 302-316.

Mehrpoor, M., Ahlers, D., Gulla, J. A., Kristensen, K., & Sivertsen, O. I. (2017). Investigating contextual ontologies and document corpus characteristics for information access in engineering settings. *Journal of Information Technology Case and Application Research, 19*(1), 10-33.

Mehrpoor, M., Gjarde, A., & Sivertsen, O. I. (2014). *Intelligent services: A semantic recommender system for knowledge representation in industry.* Engineering, Technology and Innovation (ICE), 2014 International ICE Conference. Bergamo, Italy.

Mehrpoor, M., Gulla, J. A., Ahlers, D., Kristensen, K., Ghodrat, S., & Sivertsen, O. I. (2015). *Using process ontologies to contextualize recommender systems in engineering projects for knowledge access improvement.* European Conference on Knowledge Management (ECKM 2015). Udine, Italy.

Mehrpoor, M., Ahlers, D., Gulla, J. A., & Sivertsen, O. I. (2018). *Development and Evaluation of a Knowledge Access System for Engineering Workspaces based on Recommendation and Filtering*. Submitted to the *Journal of Knowledge and Information Systems.*

Middleton, S. E., Shadbolt, N. R., & De Roure, D. C. (2004). Ontological user profiling in recommender systems. *ACM Transactions on Information Systems (TOIS), 22*(1), 54-88.

Mitchell, T. (1997). Machine Learning, McGraw-Hill Higher Education. *New York*.

Nonaka, I., & Takeuchi, H. (1995). The Knowledge Creating. *New York*.

Passant, A. (2010). *Measuring Semantic Distance on Linking Data and Using it for Resources Recommendations.* AAAI spring symposium: linked data meets artificial intelligence.

Pazzani, M. J., & Billsus, D. (2007). Content-based recommendation systems *The adaptive web* (pp. 325-341): Springer.

Pazzani, M. J., Muramatsu, J., & Billsus, D. (1996). *Syskill & Webert: Identifying interesting web sites.* AAAI/IAAI, Vol. 1.

Peffers, K., Tuunanen, T., Rothenberger, M. A., & Chatterjee, S. (2007). A design science research methodology for information systems research. *Journal of management information systems, 24*(3), 45-77.

Reinsel, D., Gantz, J., & Rydning, J. (2017). Data age 2025: The evolution of data to life-critical don't focus on big data. *Focus on the Data That's Big Sponsored by Seagate The Evolution of Data to Life-Critical Don't Focus on Big Data*.

Resnick, P., & Varian, H. R. (1997). Recommender systems. *Communications of the ACM, 40*(3), 56-58.

Ricci, F., Rokach, L., Shapira, B., & Kantor, P. B. (2011). *Recommender systems handbook* (Vol. 1): Springer.

Salton, G. (1989). Automatic text processing: The transformation, analysis, and retrieval of. *Reading: Addison-Wesley*.

Schafer, J. B., Frankowski, D., Herlocker, J., & Sen, S. (2007). Collaborative filtering recommender systems *The adaptive web* (pp. 291-324): Springer.

Suárez-Figueroa, M. C., Gómez-Pérez, A., & Fernández-López, M. (2012). The NeOn methodology for ontology engineering *Ontology engineering in a networked world* (pp. 9-34): Springer.

Thierauf, R. J. (1999). *Knowledge management systems for business*: Greenwood Publishing Group.

Turney, P. D., & Pantel, P. (2010). From frequency to meaning: Vector space models of semantics. *Journal of artificial intelligence research, 37*, 141-188.

Van Ittersum, R., & Spalding, E. (2005). Understanding the Difference Between Structured and Unstructured Documents".

Verbert, K., Manouselis, N., Ochoa, X., Wolpers, M., Drachsler, H., Bosnic, I., & Duval, E. (2012). Context-aware recommender systems for learning: a survey and future challenges. *IEEE Transactions on Learning Technologies, 5*(4), 318-335.

Wellman, J. (2009). *Organizational learning: How companies and institutions manage and apply knowledge*: Springer.

Williams, B., Figueiredo, J., & Trevelyan, J. (2013). Finding workable solutions: Portuguese engineering experience. *Engineering Practice in a Global Context: Understanding the Technical and the Social*.

Zillner, S., Hauer, T., Rogulin, D., Tsymbal, A., Huber, M., & Solomonides, T. (2008). *Semantic visualization of patient information.* Computer-Based Medical Systems, 2008. CBMS'08. 21st IEEE International Symposium. Jyvaskyla, Finland.

Zimmermann, A., Lorenz, A., & Oppermann, R. (2007). An operational definition of context. *Context, 7*, 558-571.

Appendix A

List of publications in the main body

Paper I

Intelligent Services: A Semantic Recommender System for Knowledge Representation in Industry

Paper II

Using Process Ontologies to Contextualize Recommender Systems in
Engineering Projects for Knowledge Access Improvement

Using Process Ontologies to Contextualize Recommender Systems in Engineering Projects for Knowledge Access Improvement

Mahsa Mehrpoor[1], Jon Atle Gulla[2], Dirk Ahlers[2], Kjetil Kristensen[1], Soroush Ghodrat[1], Ole Ivar Sivertsen[1]

[1] Department of Engineering Design and Materials, Faculty of Engineering Science and Technology, NTNU, Trondheim, Norway
[2] Department of Computer and Information Science, Faculty of Information Technology, Mathematics and Electrical Engineering, NTNU, Trondheim, Norway
mahsa.mehrpoor@ntnu.no
jon.atle.gulla@idi.ntnu.no
dirk.ahlers@idi.ntnu.no
kjetil.kristensen@kristensenconsulting.com
s.ghodrat@gmail.com
ole.ivar.sivertsen@ntnu.no

Abstract: Knowledge and information are highly important resources in today's knowledge economy and vital in achieving organization's goals. Particularly in engineering projects, users' tasks are highly dependent on accessing, using, and reusing these resources and users already spend a lot of time searching for relevant knowledge. As the total volume of documents across different sources and repositories increases, users face additional overhead related to search and retrieval. Knowledge workers across multiple disciplines experience fierce competition and a persistent pressure to deliver value-added contributions in a competitive global business environment with complex, multidisciplinary problems. Simple search engines are often not sufficient since they are not designed to retrieve those relevant documents that match the user's current work situation and information need. Therefore, the need for a semantics-based solution has been identified. This paper describes the early stages of a PhD project that proposes a tailored recommender system for improving knowledge accessibility in an engineering setting. The recommender system will be developed for and validated in a multidisciplinary engineering project as a case study. We take advantage of content-based filtering and collaborative filtering along with semantic technologies to provide relevant and accurate recommendations. In order to contextualize users' work situations during a project development process, the recommender system utilizes a tailored process ontology to be able to explore different dimensions of users' situations. By merging the concepts derived from the ontology, the current work situation of a given user is identified and varied fine-grained user profiles will be created at real time, so-called dynamic user profiles. Therefore, the recommender system is able to set the scope of users' interests to the exact level that a user desires in the current situation. To classify the identified relevant documents, we propose creating concept profiles that originate from process ontology concepts for further recommendation according to a collaborative approach. This paper describes the recommender system components, proposes a framework for the target recommender system, and discusses how its components are integrated and interact in order to improve information access in engineering projects.

Keywords: Work situation; Recommender Systems; Semantic recommendation; Process ontology; Dynamic User profile; Information Access

1. Introduction

Knowledge and information are highly important resources in today's knowledge economy and are vital in achieving organization's goals. One of the issues that organizations are concerned about is how to manage these important resources in order to be able to align them with their objectives and goal achievement strategies. The more complicated the corpus of organization and knowledge resources is, the more effort they need to spend to come up with a tailored solution of knowledge management. Users' tasks are highly dependent on accessing, using, and reusing the knowledge resources. There are many factors that need to be considered about the user in a particular work situation (Mehrpoor et al. 2014) since the way the users interact with the system for exploring their required information influences the management of knowledge resources (Freund 2008). Common search engines have been used for information retrieval but their results are based on the search keywords entered by the user. Search engines are not designed to filter search results according to user's work situation. Finding semantic relations between user's interest and knowledge resource content might be helpful to retrieve more relevant results. Semantic search engines (Jayavel et al. 2013) consider contextual meaning of the search query and provide more relevant results. However, the search keywords might lack the essential concepts that lead to a proper context identification. So, the approach that we have chosen for relevant information retrieval is recommender systems (Ricci et al. 2011) to improve context identification (Ahlers and Mehrpoor 2014) of semantic search engines. Recommender systems can assist knowledge managers to provide the right information for the right users and also reduce search costs.

We propose a tailored recommender system for knowledge management improvement in an engineering setting. This paper describes the early stages of work and is organized as follows: Section 2 provides an overview of related work in the area of recommender systems and using ontologies and user profiles for recommendation improvements. Section 3 elaborates the case study and formulates the challenges of knowledge access in the engineering settings. Section 4 describes the proposed recommender system to address these challenges. Section 5 is about the future research and Section 6 concludes the paper.

2. Background

This section discusses the background for the proposed solution of knowledge access improvement in the engineering projects, namely recommender systems, ontologies as semantic technologies, vector models, information extraction and indexing tools, and user profiling in recommender systems.

The two fundamental approaches of recommender systems are content-based filtering and collaborative filtering. The content-based approach focuses on the similarities between the items and combines it with users' preferences (Lops et al. 2011). The collaborative filtering approach focuses on the similarity between the preferences of the target user with other users (Schafer et al. 2007). The preferences of the user are derived from different explicit and implicit methods such as the history of users' activities in the system as user behavior or by explicitly asking users' interests and storing them in user profiles (Pazzani and Billsus 2007). User profiles might be created statically or dynamically (Hong et al. 2013). Static user profiles contain the information manually added by user; dynamic user profiles are derived from users' behavior, history of their activities and so on. To improve performance, both approaches can be combined into hybrid recommender systems (Garcin et al. 2012). The level of relevance for documents is most commonly calculated using the vector space model in content-based filtering approaches (Werner and Cruz 2013). In addition, the context-aware approach focuses on the contextual information (Dey 2001) and how the main contextual information of the user play a role in identifying the relevant information which fits the user's interests (Adomavicius and Tuzhilin 2011, White et al. 2009, Bouneffouf 2013). Especially when there are only a few ratings available from the users, contextual information becomes more important to be taken into account and analyzed to know the users in more detail (Ma et al. 2011).

In order to improve the recommendation solutions, semantic technologies are applied. Ontologies are a main semantic tool that have been used for knowledge management and constructing semantic models for the concepts of a domain in order to provide more semantic relations between the concepts and avoid the limitations in text-mining techniques (He 2013). Ontologies have been used in different recommender systems: (Paiva et al. 2013) proposes a common hierarchical architecture for ontology-based recommender systems that consists of four layers of context, discovery, recommendation and ontology to provide the relevant recommendation for a given user. For presenting semantic description of both items and user profile, (Werner et al. 2013) proposed an ontology-based recommender system for recommending economic articles. In addition to the usage of ontology for knowledge modeling, they have been used as a fundamental tool for indexing and annotating articles, which makes a system less dependent to a specific area. (Yu et al. 2007) used ontology to contextualize the user, content, and domain in three individual ontologies. Then, a recommendation method with four steps of semantic relevance calculation, recommendation refining, learning path generation, and recommendation augmentation, is proposed for providing relevant recommendations. Recommender systems should be able to extract the information in order to analyze it and identify similarities between the objects. There are available libraries for information extraction and retrieval like Elasticsearch (Banon 2012), which is used in this work. It supports different document formats and will enable the recommender system to search and analyze in real time.

3. Case study: Knowledge access in a multidisciplinary engineering project and its challenges

Shell Eco-Marathon is an annual competition that challenges student teams to design and build ultra energy-efficient vehicles. Every year a team of master students at NTNU participates in this competition and designs the DNV GL fuel fighter vehicle (Bøvre et al. 2014). The team consists of different engineering disciplines such as mechanics, electronics, materials, cybernetics, aero-dynamics and other sub-disciplines. They are working together in a multidisciplinary work environment over one year.

One of their challenges is knowledge access and sharing. During the project development, engineers would often benefit from reusing archived knowledge and information from previous competitions, but having loads of unstructured documents in different formats makes it difficult for them to search and to get access to all the relevant information for reuse, and it often causes them to start from scratch, failing to make full use of available and potentially valuable resources. The current solution of the team for knowledge access is to communicate with other students of past years' competitions to ask for the required knowledge and how to explore the relevant archived knowledge resources. They organize meetings with the past team members or send emails, which can be time-consuming and inefficient. This mirrors a standard situation in industrial companies, where new people always come into the company or new project teams are formed, who then have to learn their way around knowledge management and document storage systems.

An appropriate solution for knowledge access improvement in the engineering project requires analysis of both users and knowledge resources. We need to explore more information to figure out how well it is possible to tailor the documents to the users' assigned tasks. These issues are addressed in the following.

3.1 User information needs

Engineers with different proficiencies and different levels of expertise are involved in the project. Particular responsibilities are assigned to engineers in different phases of the project development from requirements analysis to design, implementation, test and evaluation. Some of the engineers are more involved to the early stages such as mechanical engineers, some of them contribute more in the late stages such as electrical engineers, and some people like project managers have constant responsibilities in the whole project development process. Their responsibilities contain certain role(s) and task(s) and they work with specific machine component(s). During their task performance, the engineers need to interact with each other. For instance, the design engineer needs to interact with the aero-dynamics engineer to come up with an aero-dynamics form of the vehicle. Also, electronics and cybernetics work together in some phases to make a component of the machine. In addition, there are some inter-relations between machine components and they themselves break into some sub-components, which makes them dependent for design concerns.

As an example, the mechanical engineer may have one or more roles such as team leader with some assigned tasks such as requirements analysis and designing of specific components of the machine such as steering system, brake and wheels. In order to find the relevant information, the mechanical engineer looks for those directories in the document storage that look related to his assigned tasks but since the document storage is not well-structured and not all the information that he needs is stored in the place that he expects, it makes knowledge exploration challenging and he has to spend more time on knowledge and information seeking rather than on efficient task performance. Therefore, a tool that helps him in finding his desired knowledge and information is missing.

3.2 Document storage analysis

Available archived documents are stored in a shared file system. According to the analysis of the document storages of the last three years of the competition, the type of stored documents are quite varied. Many different unstructured documents have been identified in different formats such as text-based, multimedia, modeling, programming formats, html files and other types that are created along with the output of the specified applications. In the case of the mechanical engineer, he is interested in modeling documents, a group of textual documents and a sort of modeling images. A proper solution for the identification of relevant documents for both textual and non-textual document types is required. Figure 1 depicts engineers in two disciplines of mechanics and cybernetics and their particular contexts.

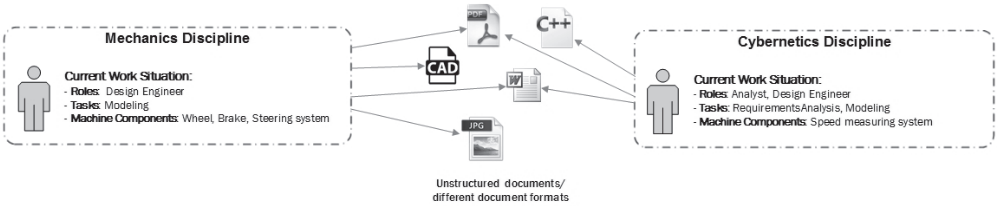

Mechanics Discipline
Current Work Situation:
- Roles: Design Engineer
- Tasks: Modeling
- Machine Components: Wheel, Brake, Steering system
PDF
C++
CAD
JPG
Cybernetics Discipline
Current Work Situation:
- Roles: Analyst, Design Engineer
- Tasks: RequirementsAnalysis, Modeling
- Machine Components: Speed measuring system
Unstructured documents/
different document formats

Figure 1. Engineers in particular work situations and their required documents

4. Tailored recommender system for knowledge access improvement

This section describes the essentials for the target recommender system. Each sub-section introduces a component of the recommender system and discusses why it is used and how it will be employed. The proposed framework consisting of all the introduced components is presented and the integration and interaction of the components is described.

4.1 Recommendation approaches

Recommendation techniques follow varied approaches according to the conducted literature review. Depending on the features and specifications of the use case, one or more of these approaches are used to provide the potentially right knowledge and information for the right users. In our use case, from one side there are users in certain work situations and from the other side there are documents that are classified in two groups of textual and non-textual documents.

Textual document formats are mostly office documents and PDFs. Non-textual document formats are mostly binary files such as multimedia documents, CAD files and so on. In order to identify similar documents to the given user's preferences, the content of the documents needs to be analyzed. Therefore, we follow the content-based recommendation approach. To explore the content of documents precisely, one of the available techniques is to index the documents and make them searchable by using textual search engine libraries (Hatcher and Gospodnetic 2004). In addition, meaningful meta-data included in the document itself might be helpful in document's content identification. Information extraction libraries are used to annotate and index the content of documents. According to the background of our work, we use Elasticsearch (Banon 2012) for information extraction and indexing the documents. In our example of the mechanical engineer, mainly those documents that have the most information pieces about mechanics, designing and modeling of wheels, brake and steering system are likely to be relevant for him. However, it should be taken into consideration that the work situation of the user is not constant during the project development and at each stage his required information narrows down to a specific scope; for instance, in the early phases, he may work on modeling the wheels. So, information about another component is not his interest in this phase.

In the two document groups, the results of content analysis would be different. In textual documents group, since they are text-based, the probability of exploring similar information pieces that fit user's interests are higher than non-textual documents. Since there is not enough meaningful text-based content embedded in non-textual documents, it limits the information extraction and indexing process for annotating these types of documents. So, another recommendation approach should also be applied. For this part, we need to investigate the users' side more and focus on the information that we can gain from their behavior. Therefore, we follow collaborative filtering approach along with content-based approach to study user's side more and come up with better recommendation results. If the mechanical engineer uses an image and rates it, specifies it as relevant, then it is inferred that this image is included in such a situation. Consequently, for another user in a similar situation, the identified document is more likely to be the right candidate for recommending to this user. However, it remains uncertain how accurate the explicit feedback of the first user is in general. Explicit feedback is the feedback given directly by the user like scoring the document. The users do not always tend to rate the documents and not all the time their rating results are reliable. Implicit feedback is also considered which is inferred from the user's behavior indirectly like the amount of time that users spend for studying a document or whether users open the document at all. Another challenge is the number of available users, new users and new documents which causes the cold-start problem where the relevance of new or unrated items is unknown (Gunawardana and Meek 2009). Since at the early phases of this approach very few documents are viewed by the users, the numbers of identified relevant documents are fewer and not sufficient for a collaborative filtering recommendation process.

Identifying the documents that are similar to user's interest only provides potentially relevant documents. However, the question still remains how we can identify the level of relevancy of each potentially relevant document or in other words, how we can rank the documents. We need to identify how close each document is to the given user's situation. To do this, we score the documents with techniques like VSM, Vector Space Model (Werner and Cruz 2013). In addition to this method, we also rank the documents using the explicit and implicit feedback.

To take more advantage of the two discussed recommendation approaches, the documents that are viewed by users can be classified according to the situation of their viewers. For classification of the documents, we propose creating profiles that match users' contextual features, which we elaborate in section 4.3.

4.2 Process ontology and factors of user's context identification

In this research, we aim to apply ontologies as one of the knowledge management tools with the inspiration from OBIE systems, Ontology-Based Information Extraction (Wimalasuriya and Dou 2010). In our recommender system, ontology is used as the fundamental resource of contextual knowledge to help us in identifying the situation of the user. For building the ontology, a scenario-based methodology called NeOn methodology is used. The NeOn methodology supports a knowledge reuse approach (Suárez-Figueroa et al. 2012).

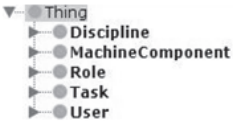


Figure 2. Main concepts of the process ontology

Among nine scenarios defined by this methodology, the first scenario is selected which is “From specification to implementation” since the development of the ontology is from scratch. In ontology requirements specification activity, ORSA, the requirements for the ontology have been specified. The ontology environment has been studied by interviewing the project team. In the conceptualization phase, the extracted terms of ORSA are conceptualized. Afterwards, the conceptual model has been transformed into a formal model and implemented using Protégé (Protege 2015). The main concepts are shown in Figure 2 that consists of five dimensions of user, discipline, role, task, and machine component. Every dimension describes a part of the work situation of the user and by combining their leaf concepts, the concepts defined in the last level of the ontology graph, different work situations are described. Figure 3 illustrates some of the leaf concepts of the ontology.

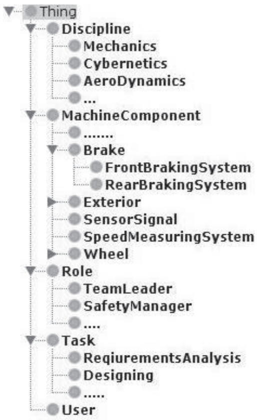


Figure 3. Parts of the hierarchical structure of the concepts of the ontology

4.3 User profiling and Concept profiles

Recommending the relevant documents to the user requires collecting some information about users’ preferences in order to compare documents with users’ preferences to identify similarities. The scope of users’ preferences includes all the essential entities that characterize the situation of the user in his work environment. These entities might be different according to different duties of the user and it causes some modification in users’ preferences scope.

We use process ontology as the generic static user profile that contains different aspects of users’ situations. Utilizing the process ontology also enables the system to create many different fine-grained user profiles which we call dynamic user profiles. Creating dynamic user profiles help the recommender system to set the scope of users’ interest to the exact level that the user desires. Therefore, the recommended items are much closer to the users’ information needs and not too many relevant and irrelevant documents are explored from the document storage. Without considering the dynamic user profiles, a recommender system covers all the preferences of the user in all the project stages and many documents will be recommended that are partially not relevant at this stage of the project. The dynamic user profile is the combination of the individual leaf concept profiles of the ontology and we call them dynamic since they are mutable and are only alive until the end of the lifetime of the current work situation of the given user.

As discussed earlier, identified relevant documents will be classified to some profiles. These profiles match the leaf concepts of the ontology; so we have a profile for mechanics, a profile for electronics, a profile for reporting, and so on. The identified relevant documents will be appended to these profiles for further recommendations to the users who have similar work situations.

In our example, different combinations of the involved aspects to engineer's work situations might cause varied user profiles. Figure 4 (a) illustrates the maximum number of possible user profiles for a particular user. Also, in (b), the maximum number of possible user profiles for the whole system, which can be created by leaf ontology concepts, is depicted. The more roles and tasks the user has, the more situations are identified and consequently the more user profiles could be created in the whole project development process. However, it should also be considered whether there are any constraints on the relations between disciplines, roles, tasks and components. For instance, wheels are related to steering system and also brakes but they are not related to batteries, technically. So the number of possible combination is reduced. These relations are inferred from the built process ontology. This knowledge can also be used to improve the indexing of documents.

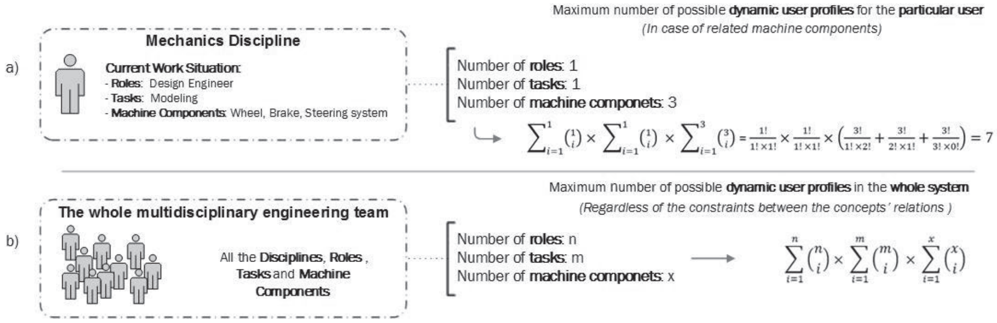

Maximum number of possible dynamic user profiles for the particular user
(In case of related machine components)
a)
Mechanics Discipline
Current Work Situation:
- Roles: Design Engineer
- Tasks: Modeling
- Machine Components: Wheel, Brake, Steering system
Number of roles: 1
Number of tasks: 1
Number of machine componets: 3
$\sum_{i=1}^{1}\binom{1}{i} \times \sum_{i=1}^{1}\binom{1}{i} \times \sum_{i=1}^{3}\binom{3}{i} = \frac{1!}{1!\times1!} \times \frac{1!}{1!\times1!} \times \left(\frac{3!}{1!\times2!} + \frac{3!}{2!\times1!} + \frac{3!}{3!\times0!}\right) = 7$
Maximum number of possible dynamic user profiles in the whole system
(Regardless of the constraints between the concepts' relations)
b)
The whole multidisciplinary engineering team
All the Disciplines, Roles, Tasks and Machine Components
Number of roles: n
Number of tasks: m
Number of machine componets: x
$\sum_{i=1}^{n}\binom{n}{i} \times \sum_{i=1}^{m}\binom{m}{i} \times \sum_{i=1}^{x}\binom{x}{i}$

Figure 4. Users' work situations and possible dynamic user profiles

4.4 Framework of the proposed recommender system

After introducing the essential components of our recommender system, we now describe how these components are integrated and interact with each other as the framework of the recommender system as represented in Figure 5.

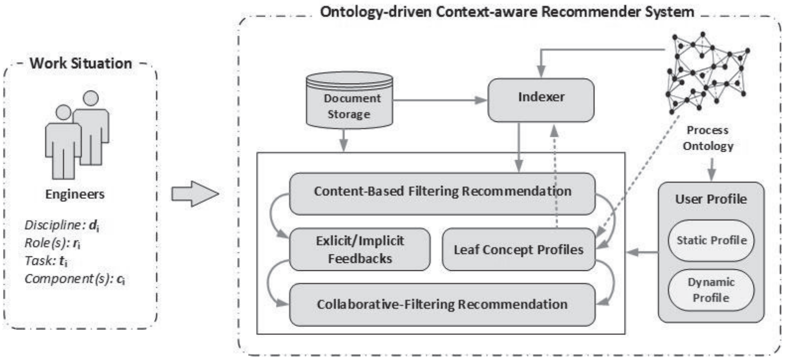


Figure 5. The framework of the ontology-driven context-aware recommender system

The process ontology as the context-based domain knowledge is used in multiple parts of the recommender system. The stored documents in the document storage are posted to Elasticsearch to be indexed and become searchable. Notice that not all the documents may index properly and have only limited metadata available. When one of the users e.g. the mechanical engineer logins to the system, he selects his current work situation derived from the process ontology. At this stage, we consider the ontology as the static user profile. The current identified user's situation is the dynamic user profile. The indexed documents that are similar to the identified user's situation are explored by Elasticsearch according to their level of relevance. Ranked documents are represented to the mechanical engineer and here we use explicit/implicit feedback such as asking the user to score the selected document or considering the time spent for studying the document. So, we capture how well the document matches user's context. At this stage, we need to classify the identified relevant document in concept profiles for later recommendations to all those users' situation that have similarities with the features of this document according to collaborative approach. Here again we take the advantage of the synergy of collaborative filtering and content-based approach. For those documents that are not indexed properly, we aim to use collaborative filtering approach by using some feedback from the users' side and follow the similar approach for textual documents; however, as mentioned earlier, the cold-start problem is still a challenge. Also notice that, the ontology will be developed and improved based on the results that we achieve in the process of documents exploration to improve the ontology concepts and the matching by adapting it to documents inputs.

5. Future work

In our future work, we intend to expand the development of the process ontology. We will investigate how matchable the indexed documents are with the concepts defined in process ontology. We will investigate possible ways to explore all the relevant resources from the document storage. We should be able to match the documents with the ontology and if required involve more aspects to the ontology such as how the documents are classified in the file system and what document formats are specified in each classification group particularly for those documents that are not indexed properly.

To take further advantage of the ontology, we will utilize the relations among concepts of the process ontology to identify more possibilities for recommendation. While the users' contexts are compared, these defined semantic relations help the recommender system to logically infer more items for recommendation since ontology supports inference and reasoning. These inferred recommendable items might not be identified if the approach is only limited to users' behavior and feedback. The task of the recommender system is not only to recommend more, but to recommend more accurate and relevant.

6. Conclusion

The advantages of using the recommender systems in the professional workplaces are discussed in this paper and a structured framework is proposed to create a tailored recommender system. This is a novel approach in knowledge management in the engineering domain by using recommender systems that contextualize users' current work situation by using ontologies. We take advantage of the two fundamental approaches of recommendation, content-based filtering and collaborative filtering, along with studying user's context, inspired from context-aware approaches. Monitoring the users from different aspects helps us to provide more relevant recommendations for the target users.

We propose the solution of creating varied fine-grained user profiles on the fly by utilizing the ontology. These dynamic user profiles help the recommender system to focus on the current scope of user's interest and retrieve those items that are closer to this particular area and do not involve all the user's preferences. This leads to spending less time in the information retrieval process and more accurate results. Any user is able to define a new situation during the project development process at real time and utilize the recommender system for relevant knowledge access.

7. Acknowledgements

A special thank you goes to my supervisors; professor Jon Atle Gulla, professor Ole Ivar Sivertsen and Dr. Kjetil Kristensen for their valuable comments, support and guidance. I also specially thank to Dr. Dirk Ahlers for his constructive criticisms and his very helpful guidance during this research. This work has been funded by Norwegian university of science and technology and the European Commission through the Large-scale Integrating Project

LinkedDesign: Linked Knowledge in Manufacturing, Engineering and Design for Next-Generation Production (FoF-ICT-2011.7.4, Project No: 284613).

8. References

Adomavicius, G. and Tuzhilin, A. (2011) Context-aware recommender systems. Recommender Systems Handbook. Springer.

Ahlers, D. and Mehrpoor, M. (2014) Semantic Social Recommendations in Knowledge-Based Engineering. SP 2014: Workshop on Social Personalisation at Hypertext 2014, CEUR-WS. vol. 1210.

Banon, S. (2012) “Elasticsearch”, [online], http://www.elastic.co/products/elasticsearch

Bouneffouf, D. (2013) Situation-aware approach to improve context-based recommender system. arXiv preprint arXiv:1303.0481.

Bøvre, M., Kyllo, M., Aalberg, J., Nordal, I. O. S. and Imaz Abal, L. (2014) DNV GL Fuel Fighter: Development and Construction of Vehicles for Participation in Shell Eco-marathon 2014. Master Thesis, Norwegian University of Science and Technology.

Dey, A. K. (2001) Understanding and using context. Personal and ubiquitous computing, 5**,** 4-7.

Freund, L. S. (2008) Exploiting Task-Document Relations in Support of Information Retrieval in the Workplace. PhD Thesis, University of Toronto.

Garcin, F., Zhou, K., Faltings, B. and Schickel, V. (2012) Personalized news recommendation based on collaborative filtering. Proceedings of the The 2012 IEEE/WIC/ACM International Joint Conferences on Web Intelligence and Intelligent Agent Technology-Volume 01, IEEE, 437-441.

Gunawardana, A. and Meek, C. (2009) A unified approach to building hybrid recommender systems. Proceedings of the third ACM conference on Recommender systems, ACM, 117-124.

Hatcher, E. and Gospodnetic, O. (2004) Lucene in action, Manning Publications.

He, W. (2013) Examining students’ online interaction in a live video streaming environment using data mining and text mining. Computers in Human Behavior, 29**,** 90-102.

Hong, W., Zheng, S. and Wang, H. (2013) Dynamic user profile-based job recommender system. Computer Science and Education (ICCSE), 8th International Conference on, IEEE, 1499-1503.

Jayavel, S., Anouncia, M. and Kapoor, A. (2013) Semantic Search Engine. International Journal of Recent Contributions from Engineering, Science and IT (iJES), 1**,** pp. 19-21.

Lops, P., De Gemmis, M. and Semeraro, G. (2011) Content-based recommender systems: State of the art and trends. Recommender systems handbook. Springer.

Ma, H., Zhou, T. C., Lyu, M. R. and King, I. (2011) Improving recommender systems by incorporating social contextual information. ACM Transactions on Information Systems (TOIS), 29**,** 9.

Mehrpoor, M., Gjaerde, A. and Sivertsen, O. I. (2014) Intelligent services: A semantic recommender system for knowledge representation in industry. Engineering, Technology and Innovation (ICE), IEEE.

Paiva, F. A. P. D., Costa, J. A. F. and Silva, C. R. M. (2013) A Hierarchical Architecture for Ontology-based Recommender Systems. Computational Intelligence and 11th Brazilian Congress on Computational Intelligence (BRICS-CCI and CBIC), IEEE, 362-367.

Pazzani, M. J. and Billsus, D. (2007) Content-based recommendation systems. The Adaptive Web. Springer.

Protégé. (2015) “Protégé, open-source ontology editor”, [online], http://protege.stanford.edu/

Ricci, F., Rokach, L., Shapira, B. and Kantor, P. B. (2011) Recommender systems handbook, Springer.

Schafer, J. B., Frankowski, D., Herlocker, J. and Sen, S. (2007) Collaborative filtering recommender systems. The adaptive web. Springer.

Suárez-Figueroa, M., Gómez-Pérez, A. and Fernández-López, M. (2012) The NeOn Methodology for Ontology Engineering. In: Suárez-Figueroa, M. C., Gómez-Pérez, A., Motta, E. & Gangemi, A. (eds.) Ontology Engineering in a Networked World. Springer Berlin Heidelberg.

Werner, D. and Cruz, C. (2013) A method to manage the precision difference between items and profiles: In a context of content-based recommender system and vector space model. Signal-Image Technology and Internet-Based Systems (SITIS), IEEE, 337-344.

Werner, D., Cruz, C. and Nicolle, C. (2013) Ontology-based recommender system of economic articles. arXiv preprint arXiv:1301.4781.

White, R. W., Bailey, P. and Chen, L. (2009) Predicting user interests from contextual information. Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval. Boston, MA, USA: ACM.

Wimalasuriya, D. C. and Dou, D. (2010) Ontology-based information extraction: An introduction and a survey of current approaches. Journal of Information Science, 39: 211-224.

Yu, Z., Nakamura, Y., Jang, S., Kajita, S. and Mase, K. (2007) Ontology-based semantic recommendation for context-aware e-learning. Ubiquitous Intelligence and Computing. Springer.

Paper III

Investigating contextual ontologies and document corpus characteristics for information access in engineering settings

Investigating Contextual Ontologies and Document Corpus Characteristics for Information Access in Engineering Settings

Authors: Mahsa Mehrpoor[1], Dirk Ahlers[2], Jon Atle Gulla[2], Kjetil Kristensen[1], Ole Ivar Sivertsen[1]

[1] Department of Engineering Design and Materials, Faculty of Engineering Science and Technology, NTNU, Trondheim, Norway

[2] Department of Computer Science, Faculty of Information Technology and Electrical Engineering, NTNU, Trondheim, Norway

mahsa.mehrpoor@ntnu.no, dirk.ahlers@ntnu.no, jon.atle.gulla@ntnu.no, kjetil.kristensen@ntnu.no, ole.ivar.sivertsen@ntnu.no

Abstract

Knowledge and information are valuable resources in enterprises for solution reuse. However, identifying relevant information from a rapidly growing number of unstructured resources is challenging for users. We discuss a personalized information access tool for professional workplaces based on recommender systems to provide relevant documents for users in specific work contexts based on domain-specific ontologies. Our use case is a multidisciplinary engineering project. We provide an in-depth analysis on the content and context of documents using information retrieval methods and semantic annotations. Upon this, we build a contextual ontology as our knowledge domain for the recommender system and evaluate the level of retrievability and coverage of it against the documents. Our results provide insight into engineers' document workspaces and show that even a simple domain ontology is able to match a majority of documents from a domain-oriented corpus. The findings support our approach of using ontology-based recommendation for domain-specific workspaces.

Keywords: Document corpus analysis, Information Access, Engineering settings, Work context, Ontology, Relevance, Recommender Systems

Introduction and Motivation

Knowledge and information resources have a highly important place in enterprises. In a study of 1998-2005, 70% of all US jobs could be classified primarily as "tacit jobs" - typical knowledge intensive jobs, drawing on deep experience and "tacit knowledge", making complex decisions based on knowledge, judgment, experience, and instinct (Johnson, Manyika, & Yee, 2005). Similar patterns can be observed in other economies and knowledge workers involved in tacit interactions represent the quickest growing segment of workers. As knowledge workers, people in an enterprise – engineers in particular – have different levels of expertise and require specific knowledge and information embedded in different types of knowledge objects stored in internal or external resources. Retrieving effective and efficient “tacit knowledge” remains challenging. Better identification, transfer and management of knowledge helps organizations to retain their resources of knowledge and reuse them in other projects instead of having to recreate it (Owen, Burstein, & Mitchell, 2004).

A survey performed in (Williams, Figueiredo, & Trevelyan, 2013) classifies different types of interactions that engineers have for information access into three groups of face-to-face, reading documents, and interactions with abstract systems and data. Abstract systems comprise conceptual and non-physical systems such as software-based systems. Interaction with systems comprises searching for information in file systems, databases, the Web, and other resources for design, modeling, simulation, and programming. File systems continue to be one of the common systems used in engineering projects for managing data. Many companies try to follow a standard naming convention for a document names and paths to improve accessibility (Eck & Schaefer, 2011). Yet directory hierarchies have challenges such as re-finding or quick browsing of filed away information that could be easily forgotten, since it is out of sight and the folder hierarchy can be rather large and complex (Jones, Phuwanartnurak, Gill, & Bruce, 2005; Ahlers & Mehrpoor, 2015). Other approaches to help groups of users access unstructured knowledge can be bookmarking or tagging systems to support knowledge sharing within organizations (Parise, Guinan, Iyer, Cuomo, & Donaldson, 2009).

Further, in many engineering projects, there is little consistent standard for creation of documents, naming and locating them. Engineers spend a lot of time on browsing and searching directories. Furthermore, the most obvious solution, such as personalized desktop search tools or existing search built into the operating system as tools to search users’ own computer files, may not be sufficient for engineers’ expectations and needs. A solution would have to enable for example refinement along workflow, classifications, topics, and other domain-specific features (Ahlers & Mehrpoor, 2015). These shortcomings make it difficult to assess which documents are available in the first place, where they are located, and how relevant they are to a specific task. The problem keeps increasing in complexity and scope with the constant growth of archived documents.

To enhance engineers’ productivity, available data sources should be efficiently re-useable and re-findable without expensive user annotations to avoid wasting time on searching knowledge that already exists within the organization, contributing to a lean enterprise (Kristensen, Krogstie, Ahlers, & Mehrpoor 2016). We aim for a system that identifies users’ information needs and relevant documents fitting their needs. Such systems are known as search engines and in particular as recommender systems that are useful tools for interacting with large and complex information spaces, as e.g. used in e-commerce (Ricci, Rokach, Shapira, & Kantor, 2011. We adapt existing recommender system approaches for dealing with the information access challenge to filter and prepare information according to engineers’ project work context (Mehrpoor, Gjarde, & Sivertsen, 2014).

The purpose of this research is improving access to relevant existing resources in an enterprise, for engineers in multi-disciplinary engineering projects. In our prior research, we discussed challenges for information access in collaborative engineering workplace settings (Ahlers, Mehrpoor, Kristensen, & Krogstie, 2015), dealing with heterogeneous documents stored in shared networked file systems (Ahlers & Mehrpoor, 2015), and the framework of our proposed system in an engineering use case (Mehrpoor et al., 2015).

In this paper, our contributions are twofold. First, we present a detailed corpus analysis of a real-life shared workspace file system from an engineering setting to better understand the available documents from an indexing and retrievability perspective and understand the characteristics and distribution of files; and second, to evaluate the document corpus against a developed domain ontology for document coverage and retrievability to validate the degree to which an ontology-based content matching approach can annotate this corpus and make the documents accessible for improved information access.

Literature review

This section discusses fundamental issues in the proposed information access system. It includes concepts such as search engines, file systems, information retrieval, recommender systems, ontologies as knowledge management tools in enterprises, and content extraction and analysis.

Information retrieval and search engines in file systems

People use different types of information retrieval systems in their daily life to satisfy their information needs such as web search engines, for example Google or Bing. Search engines are tools that index and search documents for specific keywords and return a list of documents that match query keywords. This is an application of information retrieval technology. “Information retrieval (IR) is finding material (usually documents) of an unstructured nature (usually text) that satisfies an information need from within large collections (usually stored on computers)” (Manning, Raghavan, & Schütze, 2008). Apart from well-known public search engines, there are other more specialized search tools. One example designed to work offline is Desktop Search for searching internal documents stored in a local computer. File systems of local computers or shared servers contain loads of data and information embedded in different files and directories. In order to search and access such files, prototypical search engines have been developed for different use cases. Eureka (Bhagwat & Polyzotis, 2005) is a file system search engine that infers the relations among files for improving the ranking of search results. For inferring semantic links between files, it defines three types of semantic links, content overlap, name overlap, and name reference link. They are automatically identified based on content and metadata of files. Another file system search tool called Connection (Soules & Ganger, 2005) combines traditional content-based search with context information that is collected from user activity. It identifies relations among files by tracing file system calls and uses them for reordering traditional content-based results. As discussed in the introduction, such functionalities of file system search are not powerful enough for relevant information retrieval and especially recommendation in our setting. A next step in information access leads to using additional semantic solutions and hierarchical naming towards attribute-based naming in early work in semantic file systems (Gifford, Jouvelot, Sheldon, & James W. O'Toole, 1991). Thus, analysis and retrieval of engineering documents stored in file system could be augmented through utilizing semantic-based systems such as semantic search engines to provide more relevant results (Esa, Taib, & Thi, 2010; Jayavel, Anouncia, & Kapoor, 2013). This also feeds into the topic of enterprise search. Regarding measurement and evaluation, the effect of a retrieval system on users' ability for information access is investigated and a method to capture the level of retrievability of documents is proposed (Azzopardi & Vinay, 2008).

Yet existing systems either are too general or do not apply domain-specific tools and knowledge to improve information retrieval against stored data and information in file systems with the engineering focus and cannot satisfy engineering needs, which is why we propose a new system matching these requirements. We follow the approach of domain-specific understanding of the documents to allow suitable filtering. In short, we use some existing software, but combine, adapt, and extend it in a novel way to be able to add semantic capabilities through the domain-specific ontology into a search system that we adapt to recommend documents based on semantic user context.

Recommender systems

Recommender systems are tools that suggest relevant items to users according to their interest identification (R. Burke, Felfernig, & Göker, 2011; Mahmood & Ricci, 2009). Recommender systems follow different approaches to match knowledge sources of a given domain. Knowledge sources are classified in three groups of social, individual and content; depending on what sources are available in the domain, the appropriate recommendation approach is selected (R. D. Burke & Ramezani, 2011). The two major approaches are content-based filtering and collaborative filtering. In content-based filtering, an item is recommended based on the similarities between item specifications and the profile of user's interests (Pazzani & Billsus, 2007). When the number of users is large, collaborative filtering can utilize the similarities between the preferences of a given user with other users. Items highly rated by similar users are then ranked higher in the recommendation (Schafer, Frankowski, Herlocker, & Sen, 2007). Context-aware recommendation is another approach that considers contextual information such as time and location along with the two basic entities of users and items in order to provide more personalized recommendations for target users (Adomavicius & Tuzhilin, 2011). Any effective contextual variable of the situation of the user is considered to individualize recommended items that fit the given user in certain circumstances (Verbert et al., 2012). Each of the recommendation approaches has its own strengths and weaknesses. For augmenting their performance, hybrid systems can combine recommendation approaches so that they complement the performance of each other and enable better recommendations.

Ontologies as knowledge management tools in enterprises

An ontology can be defined as a formal description of concepts and their relationships (Staab & Studer, 2013). The role of ontologies is widely examined in knowledge sharing and reuse and they have been accepted as an important tool for managing and integrating knowledge in enterprises. Ontologies have been used for knowledge integration of processes and to support complex workflow systems (Huang & Diao, 2008). In product life-cycle management, PLM, data integrity can be realized through defining a modular extendable reference ontology and integrating data along the whole product life cycle to allow semantic search and knowledge reuse (Bruno, Antonelli, Korf, Lentes, & Zimmermann, 2014). In another studied industrial case, an ontology for knowledge management is developed to support designers in generating design concepts. The developed ontology promotes knowledge reuse and systematic capture of design knowledge and helps the integration of the heterogeneous data sources (Chang, Sahin, & Terpenny, 2008). The use of ontologies for knowledge management in engineering industry is increasing as evidenced by (El Kadiri & Kiritsis, 2015; Rao, Mansingh, & Osei-Bryson, 2012; Zhen, Wang, & Li, 2013).

Furthermore, ontologies have been used in many recommender systems to improve the shortcomings that some recommendation approaches may have in different domains. A domain ontology based on users' interest in a heterogeneous environment for personalizing recommendation can minimize repetitive and tedious retrieved information (Ge, Chen, Peng, & Li, 2012). The recommendation mechanism proposed in (Zhen, Huang, & Jiang, 2010) formalizes an ontology-based context model from both the user and knowledge side and performs a semantic matching between for a more proactive way of recommendation.

Content-based filtering, content analysis and semantic annotation

Recommender systems, as well as search engines, need to build internal representations of documents. There are different item representation techniques from traditional text representation to more advanced techniques such as integrating ontologies for exploring features of the objects to be recommended and allowing to take domain-specific document features into account (Ge et al., 2012; Kang & Choi, 2011). A general concept is that only document features that are extracted and indexed can be used for search, filtering, and recommendation. This poses specific challenges to highly domain-specific systems, but also makes them more powerful than for example general search engines such as Google or Bing or general Desktop Search Tools. A high level architecture described in (Lops, De Gemmis, & Semeraro, 2011), defines three main components for a content-based recommendation process: content analyzer, profile learner and filtering component. The content analyzer component performs information and feature extraction from unstructured or semi structured documents that are machine-readable. Structured information becomes input for the profile learner and filtering components. Our focus in this paper is mainly on the content analyzer stage to study the document corpus and explore structured information and extractable features for our specific domain.

To enable information retrieval, documents need to become searchable and semantically annotated. Key information and features of documents are identified, and meta-data is extracted for documents to assist document retrieval. Meta-data represents a set of properties of the documents (Horn, 2016) and can be understood as semantic annotations for each document. An efficient retrieval process needs a central database (an index) for content and meta-data storing (Eck & Schaefer, 2011).

Case study: A multi-disciplinary engineering context

We choose the engineering scenario of interdisciplinary student groups working at our university towards building an ultra-energy-efficient vehicle. The scenario mirrors many aspects of real-world scenarios in companies (heterogeneous data, unknown file structures, large amounts of data, personnel turnover, aim to reuse knowledge, need for learning, multidisciplinary teams) and has the added benefit that we have easy access to the actual users. The Shell Eco Marathon[1] competition (SEM) is held yearly and encourages student teams around the world to design and build an ultra-energy-efficient vehicle. Since 2007, a new student team of NTNU has participated in SEM each year and worked on a vehicle called DNV GL fuel fighter[2] (Buodd & Halsøy, 2015). The main engineering disciplines involved are mechanics, electronics, and cybernetics with others involved such as design planning, aerodynamics, materials, and safety. During the project, students from different technical backgrounds work together with different levels of expertise. Since the project is repeated annually, the experiences and lessons learned of past teams are important for the current team. Building on the documented experience of past teams, every team can formulate innovative plans and consequently deliver improved results along with saving significant time, while avoiding starting from scratch.

Common ways of handover communication are sending emails or organizing meetings to transfer knowledge to the new team which is time-consuming and not efficient and sustainable. Document sources contain loads of unstructured documents of different file types that do not follow standard convention for naming and distributing in the file system. Therefore, the team still has the challenge of exploring relevant knowledge related to their assigned tasks from early to late project phases. As discussed above, ordinary desktop search tools have shortcomings when searching through a huge amount of domain-specific data. For example, documents in the Windows file system search are indexed based on generic metadata from the file system. Specialized file formats may still contain valuable content that cannot be recognized, indexed, and therefore not searched. Moreover, some documents might be only related to a particular field such as electronics or might be related to both fields of electronics and cybernetics. In such a situation, while a document is located in a folder under the root directory of electronics, it is hard to be found by a cybernetics engineer since he might not know how electronic documents are organized or not all the key terms of electronics are familiar to

[1] www.shell.com/energy-and-innovation/shell-ecomarathon.html
[2] www.fuel-fighter.com/

him to search for the required ones. These valuable and potentially reusable resources should be organized in a way to be more accessible for the team.

Approach

To be able to tailor the development of the engineering-domain recommender system, we start by examining the available document corpus to understand content and context. In this case it is a typical shared file system with content from the engineering domain as discussed above in the case description. Existing resources and their storage are studied in detail. Both qualitative and quantitative analyses are provided. Users' requirements and their expectations are studied during the process of building the specific knowledge domain in the form of an ontology. Furthermore, we perform a user study in an early part of the development process to elicit requirements and context; this is accomplished through semi-structured interviews with a user group from our scenario comprising 15 students from different engineering backgrounds that might have multiple roles and responsibilities due to the limited number of team members. After designing the conceptual knowledge model, the ontology building process is continued to formalize the domain knowledge and finally development phases following an evolutionary method.

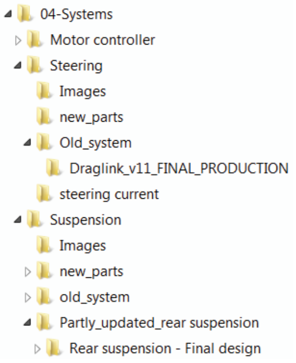
According to the characteristics of information resources, appropriate techniques for extracting information from unstructured documents are identified. Documents are semantically annotated and indexed based on a list of existing metadata from file system and additional metadata extracted from content of documents. Thereafter, the built knowledge domain is examined against indexed documents to evaluate the retrievability of documents through ontology concepts. The goal is to evaluate to what extent the ontology concepts cover existing knowledge and information resources. These individual contributions are described and evaluated in detail in the following sections.

System corpus analysis

Any Information Retrieval or Recommendation system works by matching features of a user query or context to available features within or about the items to be found. Therefore, for specialized systems, the starting point is an in-depth analysis on the available document corpus, to understand the content and context. In this section, we elaborate the results of our analysis of the file storage.

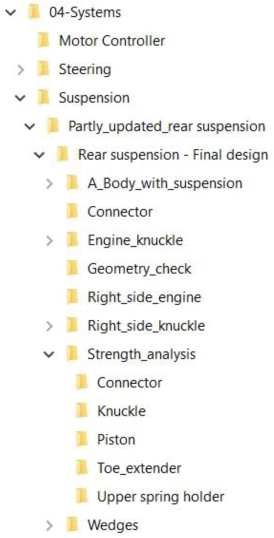
Knowledge and information storage

The data storage used in our case study is a shared networked file system on a university server containing files from the past three years of the competition. A sample of the file system structure for one year is shown in Figure 1, which shows top levels of a hierarchical directory structure. The top levels of the directory structure are similar in the years studied. However, there is not much consistency in organizing deeper levels as demonstrated in Figure 2 in terms of common names per discipline and common methods of arranging them.



```
◢ 04-Systems
  ▷ Motor controller
  ◢ Steering
      Images
      new_parts
    ◢ Old_system
        Draglink_v11_FINAL_PRODUCTION
      steering current
  ◢ Suspension
      Images
    ▷ new_parts
    ▷ old_system
    ◢ Partly_updated_rear suspension
      ▷ Rear suspension - Final design
```

Figure 1. A sample of the file system of the DNV GL project



```
⌄ 04-Systems
    Motor Controller
  › Steering
  ⌄ Suspension
    ⌄ Partly_updated_rear suspension
      ⌄ Rear suspension - Final design
        › A_Body_with_suspension
          Connector
        › Engine_knuckle
          Geometry_check
          Right_side_engine
        › Right_side_knuckle
        ⌄ Strength_analysis
            Connector
            Knuckle
            Piston
            Toe_extender
            Upper spring holder
        › Wedges
```

Figure 2. The same part of the hierarchical directories shown in Figure 1, demonstrated in deeper levels

Knowledge and information resources

Many different document types from the engineering domain are found in the collection. They included domain-specific formats such as computer aided design (CAD), finite element analysis modeling, programming documents etc. as well as common document types such as multimedia and text-based documents, such as office documents, PDFs, or various graphics formats.

To gain a quantitative understanding, a detailed overview of different types of documents is required in terms of their amount, size, format, associated metadata and other important aspects. We first examine metadata features such as document formats and sizes. Then, we examine the textual content of the documents in more detail to understand availability and broadness of keywords and the presence of technical terms.

The initial document type analysis is depicted in Figure 3 for multiple years. Documents are classified initially in two rough groups of textual and non-textual. Textual documents contain Microsoft Office documents, PDFs, texts, code, and other text-based document formats, while non-textual documents contain multimedia documents such as images or photos, as well as CAD models, analysis sheets, and other document types derived from specified engineering applications with no substantially or easily accessible textual content. As shown in Figure 4, a significant number of the non-textual documents are images of different formats and Figure 5 represents size of documents from these groups of documents.

In terms of naming convention, similar to the problem with directory name, there is not any standard for naming documents of different types. Some of the documents have information-rich filenames and give an idea about the content of the file but there are many documents that are named with seemingly arbitrary alphanumerical combinations or are otherwise not informative enough to describe the content. This impedes the search system aiming at keyword extraction.

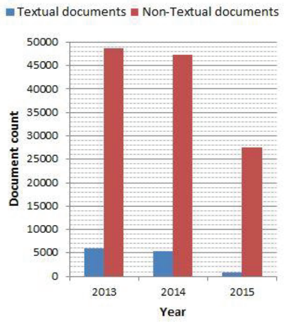


Figure 3. Number of documents in two groups of textual and non-textual in recent years.

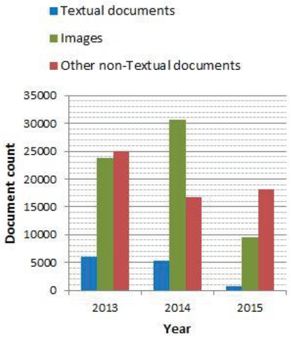


Figure 4. A detailed view of Fig3; non-textual documents and substantial amount of images.

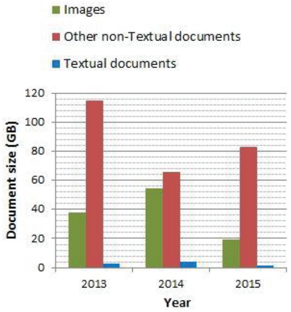


Figure 5. Size of documents from different groups represented in Fig4.

In addition, the documents from both groups are semi- or unstructured which means that documents are not based on a fixed organized model or template. Potentially valuable information is stored in archived documents but there is not any consistency in information structure in the majority of the documents. Thus, classical information retrieval systems extract keywords, but not necessarily strong semantic relations between them. As mentioned earlier, files are linked to associated metadata of the file system that can be used to explore some documents' features. However, there is a lack of metadata about content and semantic information to support browsing documents, which needs to be derived by the search system. In other scenarios where documents are stored in a document management system (DMS), this may change as metadata entry is mandated for users, but many work groups are using the shared file system for ease-of-use. In the following, both groups of documents are analyzed in detail.

Textual documents in the corpus

As shown in Figure 6, archived textual documents of the past three years are analyzed. The result shows that most of the archived knowledge is stored as PDF documents, with Word documents and Excel files following. We also see a decline in overall numbers. This may be due to more information from previous years available and useful, and also possible due to a stronger use of online collaboration tools with their own storage. This needs to be explored in future work and possibly included in our system.

Next, we examine the actual textual content of these documents. To access the textual content, we use application-specific adapters such as Apache Tika[1] for text extraction. In addition, the information retrieval tool, Elasticsearch[2] is applied for document indexing and statistical analysis. The bar charts in Figure 7, 8 and 9 illustrate the numbers of terms (individual words) for each format in recent years.

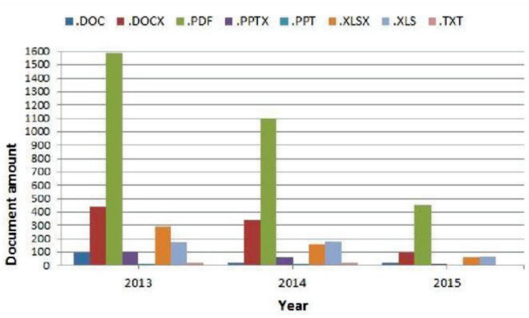


Figure 6. Existing document formats in textual documents group

The results are classified in 5 ranges to study the amount of stored content to understand how much retrievable content exists to later make it searchable and accessible for users. On the low end, a large number of PDFs and Word documents contain no textual content. We noticed that they are actually scanned PDFs or only images copied into documents with no textual content. These documents are very hard to access through textual retrieval and can only be searched by metadata. The range of terms between 0 and 1000 represents documents that contain up to two pages of textual content. The results show that the majority of archived documents are in this range. Documents with more than 1000 terms are mainly PDF and Word documents.

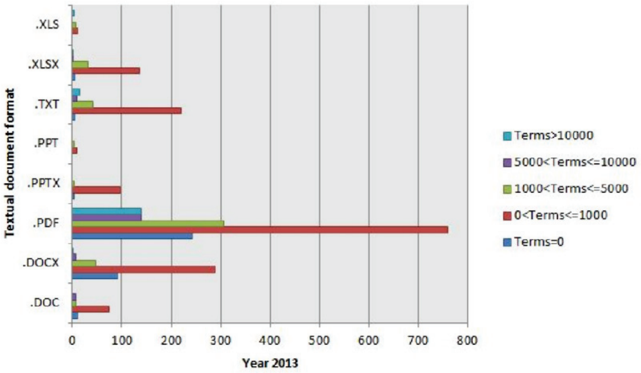


Figure 7. Number of terms per document format in year 2013

[1]. www.tika.apache.org
[2]. www.elastic.co/products/elasticsearch

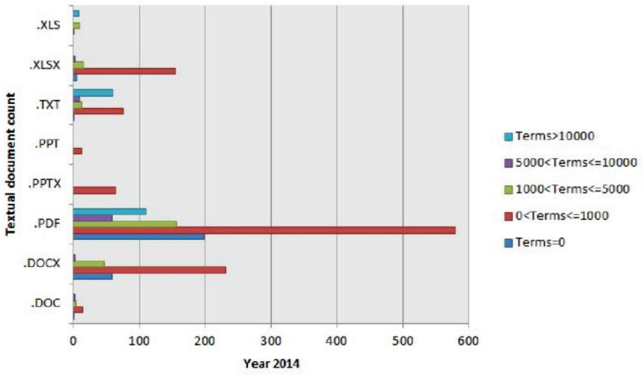


Figure 8. Number of terms per document format in year 2014

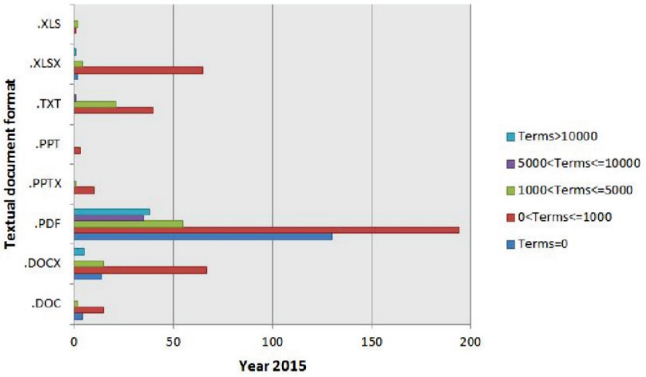


Figure 9. Number of terms per document format in year 2015

Overall, among these documents, the number of documents with no textual content is considerable. Although scanned PDFs might contain valuable information to reuse, the embedded input is hard to extract and there is same challenge for word documents and any other format with no textual content. On the other side, through file system investigation, it is found that embedded images in scanned PDFs are often duplicated individually and stored in other locations which could make them accessible. Also, if these documents are named properly to define their content, they can be explored by taking the file name and path name into account as a feature in the system (Ahlers & Mehrpoor, 2015). Overall, we see that other documents contain a suitable amount of content to make them searchable.

Non-Textual documents in the corpus

We repeat the analysis of the textual documents for the non-textual documents. As shown in Figure 4, images and photos make up the bulk of non-textual documents. Other non-textual document formats are quite varied; average of about 250 varied formats in each year. Figure 10 shows a list of the most frequent formats in this group. The number of .PRT files is significantly higher than other formats. PRT is a common file format in CAD applications for designing 3D components such as NX, SolidWorks and Abaqus. Other formats are also strongly related to specific disciplines such as .C files that refer to programming and cybernetics tasks (which would formally count as textual documents). Others, such as .O or .LOG are output files of processes that will not be useful for recommendation and should therefore be filtered out in the system (Ahlers & Mehrpoor, 2015). More generic formats such as .PRT, .SLDPRT (Solidwork part) and .FEM (Finite element analysis) refer to design and development of mechanical tasks for users from mechanics, body, aerodynamics, or fuel cell disciplines.

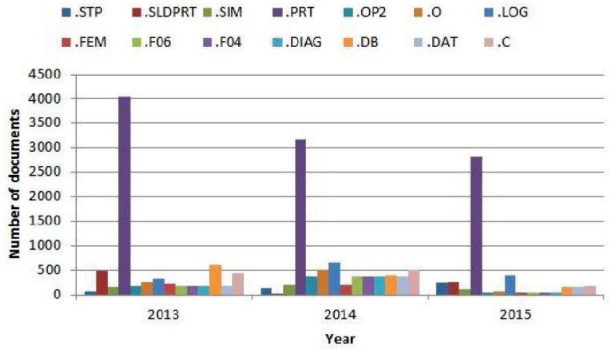


Figure 10. Some of the most frequent technical non-textual document formats

Overall, varied document formats makes it challenging to extract potentially embedded textual content. Several convertor applications are applied for image processing to extract text associated with images and also from more frequent technical document formats. However, the text extraction results were not satisfactory due to incomplete or incorrect spelling of extracted terms which is not sufficiently meaningful to give a description of the document. Similar to what was mentioned for textual documents, document name and path might contain useful content that helps in identifying subject of the given document. In the content analysis section below, we elaborate the use of other available textual metadata.

Contextual ontology as knowledge domain

As mentioned in literature review, ontologies have been utilized by enterprises for knowledge management and sharing in many cases. Our objective is to describe and structure our knowledge domain by developing a tailored ontology and to investigate how the built ontology can improve knowledge sharing and reusability through the recommender system.

For creating the ontology, we chose the NeOn methodology since it supports a knowledge reuse approach (Suárez-Figueroa, Gómez-Pérez, & Fernández-López, 2012). The first scenario among 9 different scenarios by this methodology is suitable for our system which consists of 8 tasks (Mehrpoor et al., 2015). In the early tasks of Ontology Requirements Specification Activity, ORSA, the purpose and scope of creating the ontology is identified by holding a set of semi-structured interviews with users from different disciplines (more particularly with the system engineer) with focus of users' regular tasks in the project, their information needs and information seeking behavior (see Appendix A). In addition, other available resources such as master theses of past years were used to improve the collected information about ontology environment. According to the results of the interviews, the scope of our knowledge domain should cover main concepts that characterize the work context of users during the project life-cycle.

During the middle tasks of ORSA, functional and non-functional requirements of the ontology are identified. As a non-functional requirement, the terminology of the ontology should be able to support users' identified requirements. For identifying functional requirements, a list of Competency Questions (CQ) is prepared and posed to users (See Appendix B). The collected results out of CQs and their answers are categorized in three groups of engineers' discipline, engineers' work tasks, and machine components. Thereafter, generic concepts of the ontology are identified that cover main aspects of engineers' work context and also a pre-glossary of the terms is extracted as specific concepts and assigned to their respective generic concepts. Figure 11 depicts the generic and specific concepts of the ontology developed by Protégé[1]. The conceptualization stage is iterated to check the validity of concepts with leaders of each discipline. Then, validated terminologies are formalized by means of ontology super-classes and their associated sub-classes. A detailed partial view of the final ontology is represented in Appendix C; the whole ontology consists of 134 concepts.

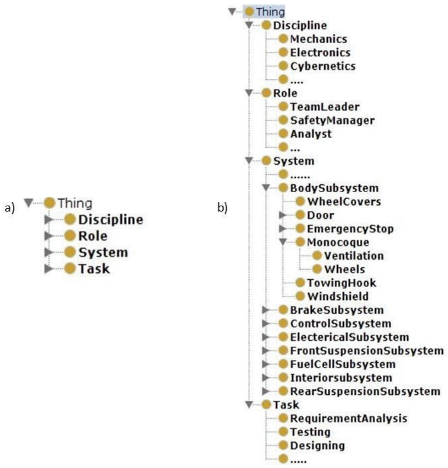


Figure 11. a) Generic concepts of the ontology (The upper ontology section). b) A detailed view of some sub-classes for each concept (The domain-specific ontology section)

[1]. protege.stanford.edu/

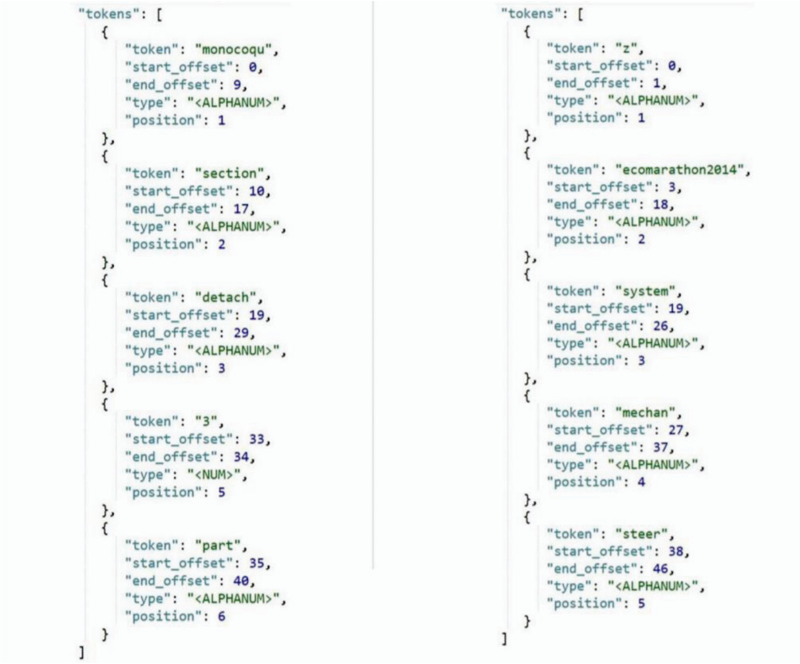
Collecting required metadata for document indexing and annotation

In order to make documents searchable, the first step is to assess what information is extractable from documents if there is any metadata available. A list of key features is collected from different resources and the reasons of choosing them are elaborated as well. Each of the features could be a lead in users' decision making process in identifying relevant information. These features are stored in so-called fields in the indexing process. Some of them are explored from file system metadata such as document name, size, format, year and path. Textual content and possible short description are extracted using Apache Tika. Note that these are only provided for textual documents. For non-textual documents, no content is extracted, leading to a shorter list of features. All the fields for indexing the documents are listed below:

- Document name: it is one of the main properties that users look into while looking for relevant information.
- Document short description: early sentences of a document might be a lead to its subject.
- Document content: it contains substantially more text to match with users' information need and therefore a rich resource to identify relevant information.
- Document format: it is closely related to users' technical background, roles and also their assigned tasks.
- Document size: larger documents identified as more relevant would contain more valuable and reusable inputs. Conversely, although some documents might have quite relevant names, not much information is embedded in them, which could be inferable by their size.
- Year of competition: users are interested to see what is done in a particular year.
- Document path: quite useful because of the hierarchical structure of the file system; it guides users to those resources that are located in deeper levels of hierarchies, and surrounding documents might be interesting for users as well. Moreover, the document path is useful in annotation and classification of documents (Ahlers & Mehrpoor, 2015).

Information extraction and retrieval libraries are applied for document analysis in content-based filtering. In our research work, Elasticsearch is applied to perform this task. To extract specific textual content, specific analyzers are required. For indexing a document, an analyzer tokenizes a given content into individual terms and stores them in an index. It also performs other operations such as removing punctuations and common words, stemming (reducing words to a root form) and any other required specific operations.

In terms of language, the majority of documents are in English and our knowledge domain is also in the same language. Therefore, the Elasticsearch English analyzer is selected since it understands the rules of English grammar and also supports stemming. According to our recommendation approach, the frequency of indexed terms is important for each document. The analyzer enables counting any existing word from the same root. For example, the root term for terms “detachable”, “detached” and “detachment” will be “detach”. This leads to improved term frequency counts and improves the recommendation process which will be discussed in our later works. As using the analyzer changes textual content, certain fields such as document name and document path and metadata are also kept in the raw (original) format for the purpose of presentation in the recommender system interface. Figure 12 illustrates how fields with different textual structures are indexed and how terms are transformed. As shown in part a), a simple text is split into five stemmed tokens and common words in English are removed and not indexed. Part b) shows the analysis result of a document path that is split into five tokens. Finally, all archived documents are indexed and become structured and searchable. Following the recommendation approach, we need to evaluate how well the built ontology is compatible with indexed documents. The results of the evaluation are addressed in the next section.



```
"tokens": [
  {
    "token": "monocoqu",
    "start_offset": 0,
    "end_offset": 9,
    "type": "<ALPHANUM>",
    "position": 1
  },
  {
    "token": "section",
    "start_offset": 10,
    "end_offset": 17,
    "type": "<ALPHANUM>",
    "position": 2
  },
  {
    "token": "detach",
    "start_offset": 19,
    "end_offset": 29,
    "type": "<ALPHANUM>",
    "position": 3
  },
  {
    "token": "3",
    "start_offset": 33,
    "end_offset": 34,
    "type": "<NUM>",
    "position": 5
  },
  {
    "token": "part",
    "start_offset": 35,
    "end_offset": 40,
    "type": "<ALPHANUM>",
    "position": 6
  }
]
```

```
"tokens": [
  {
    "token": "z",
    "start_offset": 0,
    "end_offset": 1,
    "type": "<ALPHANUM>",
    "position": 1
  },
  {
    "token": "ecomarathon2014",
    "start_offset": 3,
    "end_offset": 18,
    "type": "<ALPHANUM>",
    "position": 2
  },
  {
    "token": "system",
    "start_offset": 19,
    "end_offset": 26,
    "type": "<ALPHANUM>",
    "position": 3
  },
  {
    "token": "mechan",
    "start_offset": 27,
    "end_offset": 37,
    "type": "<ALPHANUM>",
    "position": 4
  },
  {
    "token": "steer",
    "start_offset": 38,
    "end_offset": 46,
    "type": "<ALPHANUM>",
    "position": 5
  }
]
```

Figure 12. An English analyzer is applied for different types of textual fields; a) an example of indexing a simple text. Input text: “monocoque section, detachable in 3 parts” b) an example of indexing text with path structure. Input text: “Z:\ecomarathod2014\Systems\Mechanical\Steering\”

Evaluation of the constructed ontology against the corpus

The evaluation is designed to estimate which terms from the ontology can, in raw or processed form, be used to match documents that contain the same or similar keywords. To ease this process, we use an Elasticsearch function called percolator. The normal operation of a search system is to retrieve those documents from a corpus that match a given query. The percolator works in the opposite direction. Generated queries from the ontology are indexed based on the same schema used for documents and then the documents are matched against the indexed queries to see which queries match with documents. Figure 13 depicts the functionality of percolator. This gives us an easy initial measure of the number of documents that can be matched by terms from our developed ontology.

To give a more technical description, a list of queries are generated that each query is pointing to a specific ontology concept that might contain one or several terms. Therefore, the number of generated queries is equal to the number of ontology concepts, 134 queries. The important point in query indexing process is to index queries exactly based on the same schema that is used for indexing documents. Otherwise, documents will not match to the query since the given field might not be compatible. After two indexes are ready, the evaluation process is started by running the query index against indexed documents one by one. This experiment reveals the level of retrievability and coverage of our ontology against the document corpus.

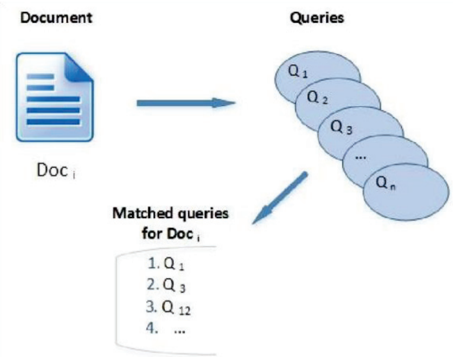


Figure 13. Percolator functionality: Matching a document against a group of queries

Evaluation results for textual documents

We start with statistical results shown in Figure 14 on the numbers of documents that generated queries could catch. Generated percolators represented on the x-axis show all the ontology concepts and the y-axis shows the number of matched documents per built percolator/ontology concept. The results show that more core and generic concepts of the ontology could match more documents. Those ontology

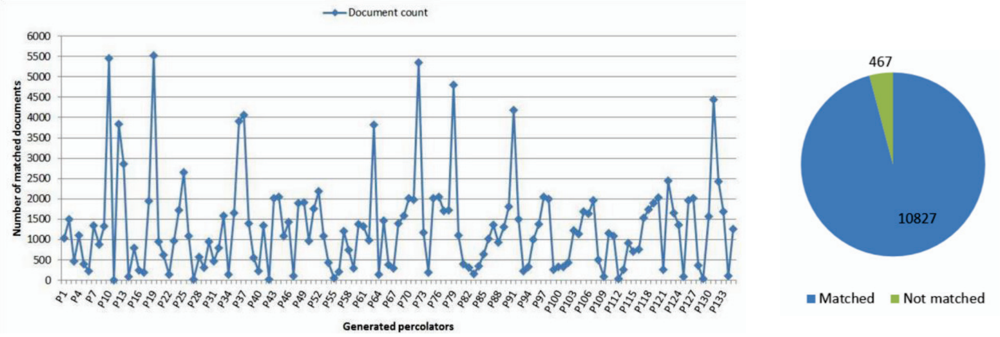


Figure 14. Number of matched textual documents with ontology concepts

concepts that contain any of the terms “control”, “motor” or “system” could catch substantial number of documents as a match, which is not that surprising when comparing to more specialized terms. Note that a document could match with multiple queries. Specific concepts such as components of machine subsystems catch less number of documents compared to generic terms, but will have a higher matching specificity. Certain terms match only a few documents, but all ontology concepts could be matched to documents in the corpus.

Looking at this from the document side, the pie chart in Figure 14 shows that around 95 percent of available textual documents retrievable by direct ontology concepts without any adaptations or relaxations in the matching and only 5 percent of documents are not retrievable, i.e. are not matched by any ontology concept term. This would be an important fraction, so we examine these in detail in terms of their name, format and content. In Figure 15, these 5% of non-matched documents are examined based on their file type. We see that more than half of them are PDF documents, and around two fifth are Word and text documents, with a small number of other document formats.

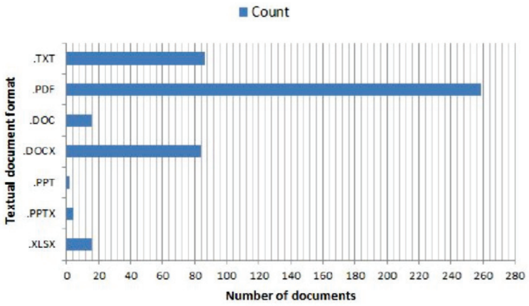


Figure 15. Number of non-matched document from different formats in textual group

In order to assess why these documents could not be matched, we examine their textual content in more detail. Figure 16 represents how many terms could be extracted from non-matched documents. No terms could be extracted from about a third of documents, mostly from PDFs. These documents can be scanned PDFs or documents with image content only as we had discussed above. One fifth of documents contain a maximum of 100 extractable terms and only around one third of documents contain more than 100 terms. These documents mainly belong to administrative, financial and personal directories where non-technical content is stored such as traveling information, expenses etc. In addition, a few documents are in Norwegian language. As the created ontology is in English, this limits the matching process for non-English terms. It shows also that future work could work towards a multilingual ontology.

It is worth mentioning that all the document names and document paths of those non-matched textual documents did not have meaningful technical terms to be matched with ontology concepts. Usually, file name and path are more valuable in the absence of file content, but in this case both main features failed to be useful. It is observed that documents and paths are mostly named with people's name, numbers, combination of letters, joint words e.g., "systemanalysis" or Norwegian terms.

Evaluation results for non-textual documents

We repeat the previous process for non-textual documents where we can match only file name and file path against ontology concepts. With less content to work with, fewer matched documents are expected as represented in Figure 17. The matching results show a drop to around 43% compared to the previous 95%. Among matched documents, there is a similar trend as for textual documents. Generic and core concepts of the ontology could again match with more documents and as terms become more specific, the number of matched documents decreases. Yet, only 6 queries could not find any match, which is surprising, given the usually lower information density in these features. Although no textual content is extracted from documents of this group, fields of document name and path contain useful content to make a group of non-textual documents retrievable by ontology concepts.

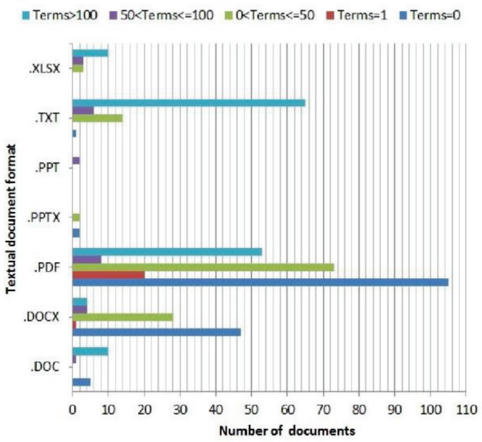


Figure 16. Number of terms embedded in non-matched textual documents

We then investigate the reason of mismatching more than half of the non-textual documents as seen in the pie chart of Figure 17. Choosing a random sample, we observe that we have the same challenge of insufficient naming system for both document name and path. Analytical results reveal that many of the documents that could not match to any of the queries are from the same cluster as before. This cluster covers personal information that refers to team members, many photos captured during project development from team activities per year etc. Furthermore, duplicated documents are observed quite often in both groups of textual and non-textual documents even within the same year.

Overall, mostly those documents that are located in directories with specific and suitable names or having such terms in the file name – often containing machine system structure terms – are retrievable by ontology concepts. Besides the content-based recommendation approach, a complementary approach is required to identify those important non-textual documents that could not be caught by ontology concepts. This would be an argument for extending the approach we outlined in (Ahlers & Mehrpoor, 2015) of how to use clustering and relation analysis to spread semantic labels.

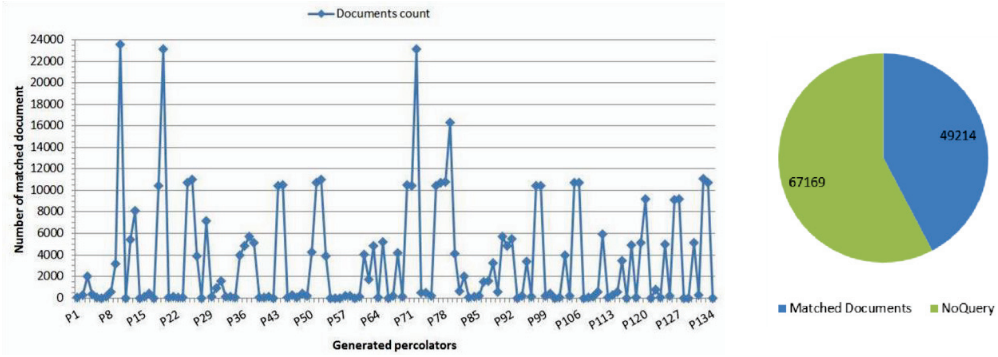


Figure 17. Number of matched non-textual documents with ontology concepts

Constructed ontology and profiling approach

In the professional scope, the context of users describes their interests. Our ontology can model users' preferences since it consists of different dimensions of users' work context. As described in (Mehrpoor et al., 2015), our ontology can be understood as the combination of all possible static user profiles which means that we actually create profiles at ontology level and not as user level in the target recommender system. The content of these profiles will be used later in the developed recommender system to provide more relevant documents for users. Figure 18 depicts the user interface for identifying users' work context from ontology concepts. A user's work context is mutable during project development. By combining different concepts of the ontology, dynamic user profiles are built. Using dynamic profiles enables users to create varied fine-grained user profiles on the fly and narrow down their scope of context. Figure 19 illustrates a preview of a selected work context.

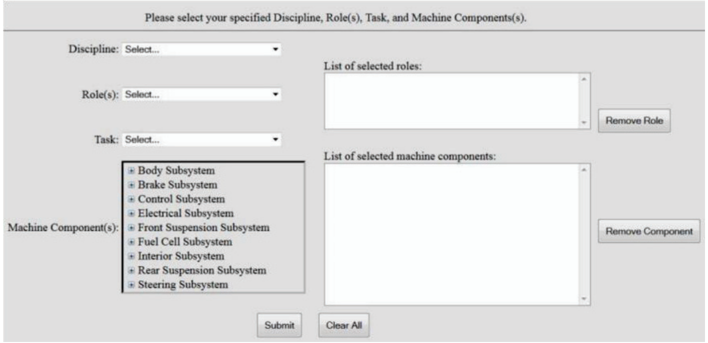
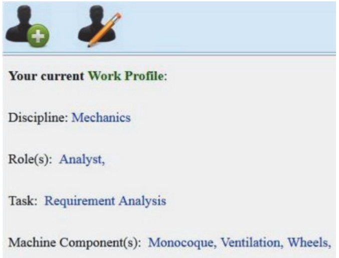


Figure 18. User interface snapshot for selecting work context driven from ontology



Your current Work Profile:

Discipline: Mechanics

Role(s): Analyst,

Task: Requirement Analysis

Machine Component(s): Monocoque, Ventilation, Wheels,

Figure 19. A view of selected work context (Dynamic work-profile)

Future work

In our future work, we aim to further evaluate our proposed system in the engineering context by evaluating the initial and revised systems with a group of users. We aim to investigate the level of accuracy and relevance of search results against different users' work context in the traditional engineering project. Moreover, we investigate our non-personal profiling approach and assess the idea of generating semantic content for ontology concept profiles using the results of users' explicit feedbacks and improving the profile content gradually during the system lifecycle. Through this approach, collected terms in concept profiles could be useful in redesigning the ontology and improving the potential cold-start problem with existing ontology concepts. This would give us an interesting angle to extend and grow the ontology naturally. Overall, in our future work we will evaluate the proposed system from different quality perspectives to be able to deliver more precise recommendation results for this challenging domain.

Discussion and Conclusion

In this work, we have analyzed a document corpus from a multidisciplinary engineering project, to better understand the setting of an engineering-domain recommender system. The objective of such a system is to alleviate the problem of information overload by providing relevant knowledge and information for engineers from different disciplines by focusing on their work context. Our study consists of three main contributions.

First, we analyzed in depth a live, in-the-field document corpus from real projects to understand the structure of existing documents in the engineering context. While this has been only one project, we have seen other very similar systems in our work, even if we could not analyze them in depth. Also, appropriate information extraction and retrieval tools have been applied to analyze valuable input of documents along with annotating the documents semantically.

Second, we investigated the requirements and specifications of users through semi-structured interviews with engineers of different disciplines. The focus of the interviews was to identify users' regular tasks, their information needs and information seeking challenges. The information gained out of the interview results outlined the scope of our knowledge domain to cover the work context of users during the project.

Third, we constructed an ontology as our knowledge domain and verified its use and suitability for the engineering domain by validating it against the document corpus. All the concepts that characterize users' work context are collected and represented by building domain specific ontology. In order to assess the retrievability and coverage of ontology concepts against the documents, we examined our ontology-based content matching approach with advanced search tool to measure the performance of created knowledge domain.

The results show that the provided knowledge domain could cover a majority of documents, with up to 95% for textual and still 43% for non-textual files. Regarding non-textual documents, additional metadata and annotations could be a complementary solution to our approach to improve information access and retrievability. In addition, the results indicate a cold-start problem for some concepts that are either too wide or too narrow according to the level of matching with the document corpus. This causes retrieving too many or too few documents. However, it proves valuable insight to refine the matching towards better specificity for certain concepts based on the document corpus, especially for very specific concepts. Refinements and improvements in the ontology development can then focus on relations between concepts and identifying more semantically related terms per concept to cover more semantically relevant results and improve the cold-start problem. Any combination of ontology concepts could create dynamic work profile for users in different stages of the project. Therefore varied fine-grained work profiles could be created from user side in real-time. This feature enables the recommender system to dynamically narrow or widen the scope of search and focus on current user preferences and thus provide recommendation that is closer to his information needs.

Our findings here and the overall research have relevant implications for academia and industry. We point to a common challenge of information overload and knowledge reuse in multidisciplinary engineering teams. This links to improved engineering processes in all fields to improve efficiency and reduce waste. We expect the research to be generalizable with limited constraints. The results of our experiments on the specifications of knowledge and information resources reveal key aspects that should be considered for analysis of such contexts and domain ontologies. In addition, we examined and validated that the approach of building a domain-specific ontology for the engineering domain for a semantically improved recommender system is suitable for a typical scenario that collects a huge number of varied documents in a shared file system for an engineering project; and this approach forms a suitable basis for our information access system for the engineering case.

References

Adomavicius, G., & Tuzhilin, A. (2011). Context-aware recommender systems *Recommender systems handbook* (pp. 217-253). Springer.

Ahlers, D., & Mehrpoor, M. (2015). *Everything is filed under 'File': Conceptual Challenges in Applying Semantic Search to Network Shares for Collaborative Work.* Proceedings of the 26th ACM Conference on Hypertext & Social Media Hypertext 2015.

Ahlers, D., Mehrpoor, M., Kristensen, K., & Krogstie, J. (2015). *Challenges for information access in multi-disciplinary product design and engineering settings.* Tenth International Conference on Digital Information Management (ICDIM 2015).

Azzopardi, L., & Vinay, V. (2008). *Retrievability: an evaluation measure for higher order information access tasks.* Proceedings of the 17th ACM conference on Information and knowledge management CIKM.

Bhagwat, D., & Polyzotis, N. (2005). Searching a file system using inferred semantic links. Proceedings of the sixteenth ACM conference on Hypertext and hypermedia, Austria.

Bruno, G., Antonelli, D., Korf, R., Lentes, J., & Zimmermann, N. (2014). Exploitation of a semantic platform to store and reuse PLM knowledge. IFIP International Conference on Advances in Production Management Systems, Springer Berlin Heidelberg.

Buodd, M., & Halsøy, B. (2015). DNV GL Fuel Fighter towards Shell Eco-marathon 2015. Master thesis: NTNU, Trondheim.

Burke, R., Felfernig, A., & Göker, M. H. (2011). Recommender systems: An overview. *AI Magazine, 32*(3), 13-18.

Burke, R. D., & Ramezani, M. (2011). Matching Recommendation Technologies and Domains chapter 11. *Recommender systems handbook, 1*, 367.

Chang, X., Sahin, A., & Terpenny, J. (2008). An ontology-based support for product conceptual design. *Robotics and Computer-Integrated Manufacturing, 24*(6), 755-762.

Eck, O., & Schaefer, D. (2011). A semantic file system for integrated product data management. *Advanced engineering informatics, 25*(2), 177-184.

El Kadiri, S., & Kiritsis, D. (2015). Ontologies in the context of product lifecycle management: state of the art literature review. *International Journal of Production Research, 53*(18), 5657-5668.

Esa, A. M., Taib, S. M., & Thi, H. N. (2010). *Prototype of semantic search engine using ontology.* IEEE Conference on Open Systems (ICOS), 109-114.

Ge, J., Chen, Z., Peng, J., & Li, T. (2012). *An ontology-based method for personalized recommendation.* 11th International Conference on Cognitive Informatics & Cognitive Computing (ICCI* CC), 2012 IEEE.

Gifford, D. K., Jouvelot, P., Sheldon, M. A., & James W. O'Toole, J. (1991). Semantic file systems. *SIGOPS Oper. Syst. Rev., 25*(5), 16-25.

Horn, B. L. (2016). Computer System For Automatic Organization, Indexing And Viewing Of Information From Multiple Sources: US Patent 20,160,117,071.

Huang, N., & Diao, S. (2008). Ontology-based enterprise knowledge integration. *Robotics and Computer-Integrated Manufacturing, 24*(4), 562-571.

Jayavel, S., Anouncia, M., & Kapoor, A. (2013). Semantic Search Engine. *International Journal of Recent Contributions from Engineering, Science & IT (iJES), 1*(2), pp. 19-21.

Johnson, B. C., Manyika, J. M., & Yee, L. A. (2005). The next revolution in interactions. *McKinsey Quarterly, 4*, 20-33.

Jones, W., Phuwanartnurak, A. J., Gill, R., & Bruce, H. (2005). *Don't take my folders away!: organizing personal information to get things done.* CHI'05 extended abstracts on Human factors in computing systems.

Kang, J., & Choi, J. (2011). *An ontology-based recommendation system using long-term and short-term preferences.* International Conference on Information Science and Applications (ICISA), 2011.

Kristensen, K., Krogstie, J., Ahlers, D. & Mehrpoor, M. (2016). LEAP Collaboration System. Chapter 5, in: The Methods and Tools of the Linked Engineering and Manufacturing Platform (LEAP), Academic Press.

Lops, P., De Gemmis, M., & Semeraro, G. (2011). Content-based recommender systems: State of the art and trends *Recommender systems handbook* (pp. 73-105): Springer.

Mahmood, T., & Ricci, F. (2009). *Improving recommender systems with adaptive conversational strategies.* Proceedings of the 20th ACM conference on Hypertext and hypermedia.

Manning, C. D., Raghavan, P., & Schütze, H. (2008). *Introduction to information retrieval* (Vol. 1): Cambridge university press Cambridge.

Mehrpoor, M., Gjarde, A., & Sivertsen, O. I. (2014). *Intelligent services: A semantic recommender system for knowledge representation in industry.* International ICE Conference on Engineering, Technology and Innovation (ICE), 2014.

Mehrpoor, M., Gulla, J. A., Ahlers, D., Kristensen, K., Ghodrat, S., & Sivertsen, O. I. (2015). *Using process ontologies to contextualize recommender systems in engineering projects for knowledge access improvement.* ECKM2015.

Owen, J., Burstein, F., & Mitchell, S. (2004). Knowledge Reuse and Transfer in a Project Management Environment. *Journal of Information Technology Case and Application Research, 6*(4), 21-35. doi:10.1080/15228053.2004.10856052

Parise, S., Guinan, P. J., Iyer, B., Cuomo, D. L., & Donaldson, B. (2009). Harnessing Unstructured Knowledge: The Business Value Of Social Bookmarking At Mitre. *Journal of Information Technology Case and Application Research, 11*(2), 51-76.

Pazzani, M. J., & Billsus, D. (2007). Content-based recommendation systems. The adaptive web (pp. 325-341): Springer.

Rao, L., Mansingh, G., & Osei-Bryson, K.-M. (2012). Building ontology based knowledge maps to assist business process re-engineering. Decision Support Systems, 52(3), 577-589.

Ricci, F., Rokach, L., Shapira, B., & Kantor, P. B. (2011). Recommender systems handbook (Vol. 1): Springer.

Schafer, J. B., Frankowski, D., Herlocker, J., & Sen, S. (2007). Collaborative filtering recommender systems *The adaptive web* (pp. 291-324): Springer.

Soules, C. A. N., & Ganger, G. R. (2005). Connections: using context to enhance file search. *SIGOPS Oper. Syst. Rev., 39*(5), 119-132. doi:10.1145/1095809.1095822

Staab, S., & Studer, R. (2013). *Handbook on ontologies*: Springer Science & Business Media.

Suárez-Figueroa, M., Gómez-Pérez, A., & Fernández-López, M. (2012). The NeOn Methodology for Ontology Engineering. In M. C. Suárez-Figueroa, A. Gómez-Pérez, E. Motta, & A. Gangemi (Eds.), *Ontology Engineering in a Networked World* (pp. 9-34): Springer Berlin Heidelberg.

Verbert, K., Manouselis, N., Ochoa, X., Wolpers, M., Drachsler, H., Bosnic, I., & Duval, E. (2012). Context-aware recommender systems for learning: a survey and future challenges. *IEEE Transactions on Learning Technologies, 5*(4), 318-335.

Williams, B., Figueiredo, J., & Trevelyan, J. (2013). Finding workable solutions: Portuguese engineering experience. *Engineering Practice in a Global Context: Understanding the Technical and the Social*.

Zhen, L., Huang, G. Q., & Jiang, Z. (2010). An inner-enterprise knowledge recommender system. *Expert Systems with Applications, 37*(2), 1703-1712.

Zhen, L., Wang, L., & Li, J.-G. (2013). A design of knowledge management tool for supporting product development. *Information Processing & Management, 49*(4), 884-894.

Appendix A

List of questions and issues for interview with users

Introduction:

The interview is started with asking about engineer's/user's work tasks during project development.

Questions:

1. How engineers/users usually work with documents?
    a) How do you decide if a document is the right one for you to take a look?
    b) What is important about the document to take your attention in first overview?
        - Title, Format, Date modified any other information?
2. Which sorts of documents are related to your discipline?
3. How much do you use the archived documents of past projects?
4. Are different disciplines related to each other? If yes, which ones? And on which parts and activities? (Collaboration aim)
5. Are there any overlaps in the responsibilities of different roles with each other?
6. What are the project stages?
    a) Are the project stages something regular each year?
    b) In each project stage which disciplines are involved and in which responsibilities?
7. How much duplication is there in the document storage in each year?
8. How much the predefined technical rules from SEM affect the documents that you want to reuse from past years?
9. What sorts of documents are more likely to be reused from past projects?
10. What sort of information sources associates with each task?
11. How do you prioritize the disciplines according to their level of importance in the project?
12. How do you prioritize documents according to their types in your work area?
    a) Descriptions, modeling, guidelines, rules, etc.
13. What are the challenges that you face with while looking for required information and knowledge?
14. What kinds of tasks are assigned to experts and what kind of tasks are assigned to novices?
15. What document storages do you have in DNV GL? And by which one you usually work with?
16. How disciplined the documents are stored in the folders in the information resources?

Appendix B

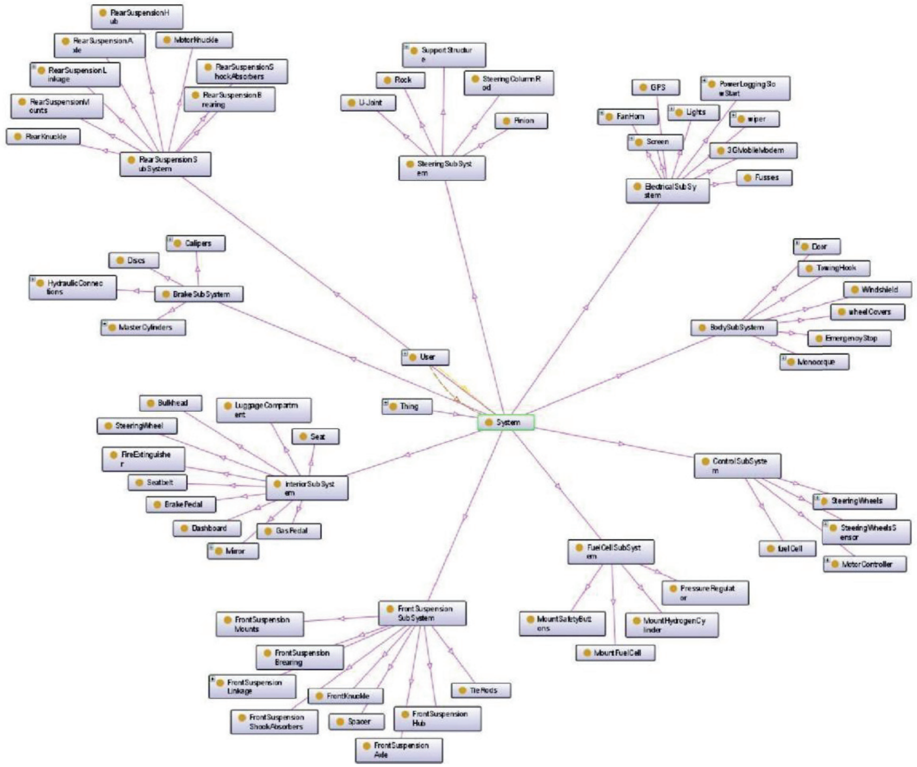
Competency Questions

1. What are different disciplines in DNV GL project?
2. Are different disciplines related to each other? If yes, which ones?
3. What are different roles in each discipline?
4. What are different tasks for each role?
5. Is there any overlapping in the responsibilities?
6. Elaborate each task in each project stage, in each discipline, for each engineering role?
7. Which disciplines are involved in each project stage?
8. What are the main subsystems of the vehicle to be designed?
9. What are the components embedded in each of the subsystems?
10. Which subsystems are related to each other and how?

Appendix C

A Part of implemented ontology

Super classes and subclasses of entity “System”



Development and Evaluation of a Knowledge Access System for Engineering Workspaces Based on Recommendation and Filtering

Paper IV

Appendix B

List of supplementary publications

Challenges for Information Access in Multi-Disciplinary Product Design and Engineering Settings

Paper V

Is not included due to copyright

Everything is Filed under File: Conceptual Challenges in Applying Semantic Search to Network Shares for Collaborative Work

Paper VI

Is not included due to copyright

Book chapter

LEAP Collaboration System

Chapter 5

LEAP Collaboration System

K. Kristensen[1], J. Krogstie[2], D. Ahlers[2], M. Mehrpoor[1]

[1]Department of Engineering Design and Materials, NTNU, 7491 Trondheim, Norway, kjetil.kristensen@ntnu.no, mahsa.mehrpoor@ntnu.no
[2]Department of Computer and Information Science, NTNU, 7491 Trondheim, Norway, john.krogstie@idi.ntnu.no, dirk.ahlers@idi.ntnu.no

Abstract
The LEAP collaboration system is a compilation of models, concepts, elements and technology components that – when combined and structured in a meaningful way to teams of end users – enable companies to execute split location engineering projects in a way that represents a competitive advantage. Written specifically for split-location and multidisciplinary teams, this chapter contains a description of these elements in the context of split location engineering. Furthermore, the chapter describes how to develop and operationalise such a system in a way that it represents a holistic, meaningful and value-adding collaborative working environment that enables engineers and other knowledge workers to make decisions, solve problems and address multidisciplinary issues effectively and efficiently. Based on lean thinking, the LEAP Collaboration System approach identifies ways of reducing waste in collaboration processes. An evaluation of knowledge sources is described, together with approaches supporting knowledge creation in lean engineering environments. The chapter concludes with a toolbox that companies can use to diagnose collaboration problems and challenges, and systematically improve collaboration in their teams.

Keywords. Collaboration system, collaborative engineering, knowledge sources, knowledge creation, collaborative diagnostics, waste in collaboration