Goal-Oriented Know-How Mapping
Mapping process, prototype, and empirical studies
by
Sadra Abrishamkar

A thesis submitted in conformity with the requirements for the degree of Master of Information
Faculty of Information
University of Toronto

© Copyright by Sadra Abrishamkar (2014)
Goal-Oriented Know-How Mapping
Sadra Abrishamkar
Submitted for Degree of Master of Information (2014)

Abstract
We elaborate and evaluate a goal-oriented modelling approach for mapping the means-ends knowledge in technology domains. The means-ends relationship connects a solution to a problem, capturing “know-how” of a domain that can be made explicit. Means-ends knowledge from publications is codified and used to create a know-how map in order to organize domain knowledge in terms of problems and their solutions.

Domain researchers as well as newcomers, can use the map to search for problems and solutions, and to evaluate the strengths and weaknesses of alternative solutions. A know-how map is expected to facilitate discovery of knowledge gaps, thus promoting new research and innovation.

This work further explores goal modelling approach to know-how. The process of creating and using goal-oriented know-how maps is documented. The user evaluation conducted in this work gathered empirical data about the performance of know-how maps in comparison to a written review of the domain.
# Table of Contents

1. **Introduction** ..................................................................................................................... 1
   1.1 Problem Statement ............................................................................................................. 2
   1.2 Research purpose .............................................................................................................. 3
   1.3 Thesis organization .......................................................................................................... 4

2. **Literature review** ............................................................................................................... 5
   2.1 Knowledge classification................................................................................................. 5
   2.2 Knowledge mapping .......................................................................................................... 6
   2.3 Knowledge representation and visualization.................................................................. 9
   2.4 Knowledge management processes................................................................................ 12
   2.5 Goal structuring of knowledge domains ....................................................................... 13
   2.6 Goal-oriented framework and i* modelling ................................................................. 15
   2.7 Scholarly Ontology Project ........................................................................................... 17

3. **Research Methodology** .................................................................................................... 20
   3.1 Information system research: design science ................................................................ 20
       3.1.1 Guideline 1: Design as an artefact .......................................................................... 21
       3.1.2 Guideline 2: Problem relevance .............................................................................. 21
       3.1.3 Guideline 3: Design Evaluation ............................................................................. 22
       3.1.4 Guideline 4: Research Contribution ........................................................................ 22
       3.1.5 Guideline 5: Research Rigour.................................................................................. 23
       3.1.6 Guideline 6: Design as a search process ................................................................. 23
       3.1.7 Guideline 7: Communication of research ............................................................... 23
   3.2 Research methodology steps ............................................................................................ 24
       3.2.1 Mapping web mining domain ............................................................................... 24
       3.2.2 GOKM process documentation .............................................................................. 25
       3.2.3 GOKM prototype .................................................................................................... 25
       3.2.4 Evaluating the GOKM approach ............................................................................ 25

4. **Goal-Oriented Know-how Mapping (GOKM)** ................................................................. 26
   4.1 Web Mining knowledge domain description and Know-how Map .................................. 26
   4.2 GOKM components .......................................................................................................... 29
   4.3 Goal Oriented Know-how Mapping components ............................................................ 30
       Goals (hierarchy and structure) ......................................................................................... 30
1 Introduction

Classification techniques are fundamental for organizing knowledge in a domain. The essential purpose of knowledge classification is to organize and maintain documents for effective retrieval (Mai, 1999). Classification involves an ongoing process of grouping, labelling and relating knowledge components. Knowledge is classified so that it meets the expectations of the user for browsing, retrieval, as well as enabling learning.

Many of classification approaches focus on communicating the knowledge in textual format; however, we believe that the technological knowledge domains can be organized based on the means-ends relationship. The term know-how is often used to describe the kind of knowledge that enables someone to achieve a goal by following a method or solution. Know-how may be in the form of tacit knowledge among humans; however, there is also much explicit means-ends knowledge, such as those published in technical research literature. Means-ends knowledge is a dominant form of knowledge in technological domains since the primary goal of such domains is to develop solutions to technical problems, and to keep seeking better, more effective solutions.

The mapping of know-how knowledge can facilitate the learning process of researchers and newcomers to a domain. Practitioners can use the know-how maps to explore and discover problems and their solutions, and evaluate the alternatives. Know-how maps can also enable researchers to find the gaps between the problems and solutions, which can lead them to narrow the gap by directing more research to that area.
1.1 Problem Statement

The traditional classification and organization of knowledge domains have been always a challenged by the quests for relativity and stability (Mai, 2004). The knowledge domain is a collection of related knowledge, which is classified and bound together on a specific aspect of the knowledge. Although a subject or keyword based classification is able to address concepts such as the document title or discipline, it is not effective when dealing with problem-solution domains where solutions are aimed to address the problems. This is because the problem-solution domains have an additional structure, which is difficult to address via keyword-based classification, and needs different organization and representation technique.

There has been much research on improvements of knowledge organization and its techniques; however, keyword based approaches are limited in areas of 1) ability to explore and browse a knowledge domain, 2) the ability to discover knowledge based on the problem-solution relationship 3) the ability to identify the knowledge gaps, 4) ability to show the relationships of knowledge components (problem-solution) within the knowledge domain. The current limitations of knowledge classification and mapping techniques in technological domains call for innovation and improvement in effectively mapping the problem-solution relationship of the domain.

In this work, we focus on knowledge organization and representation in problem-solution based knowledge domains (i.e. technological domain). Recent works have noted the lack of special attention to the means-ends relationship in most of the existing classification approaches and knowledge mapping techniques (Gross et al., 2013; Nasser, 2012). Knowledge mapping approaches, such as concept maps (Novak & Canas, 2006), and claim-oriented argumentation (Shum et al., 2000; Uren et al., 2006) focus on different aspects and relationships of the
knowledge domain for their mapping approach. In this work, we focus on the means-ends relationship, which is at the core of i* goal-oriented modelling approach. The goal-oriented modelling approach can be used to map the know-how relationships in a domain. The knowledge discovery can be facilitated through better understanding of the know-how structure, and exploration of branches of knowledge.

Nasser (2012) proposed an approach to goal-oriented know-how mapping, which incorporates the concepts of goal-oriented modelling into know-how mapping. This research aims to improve upon that previous work by evaluating and clarifying the features of know-how mapping framework. We also compare the goal oriented know-how mapping with a textual representation of knowledge domains by conducting a user study.

1.2 Research purpose

In this research, we pursue the following goals.

- Organize knowledge domains based on the publication goals and solutions proposed within a technological domain. We apply the concepts developed by Nasser (2012) to a new knowledge domain, and thus evaluate its applicability to different knowledge domains.

- Provide a conceptual definition of knowledge components and process documentation for codification of publications. We provide a detailed definition and characterization of knowledge components. This set of concepts is followed by detailed process documentation on how to codify the know-how components, create Goal-Oriented Know-How Maps (GOKM), and utilize these maps.
• Implement a web based proof-of-concept prototype for GOKM to facilitate knowledge codification and domain research. Until now, GOKM approach has never been attempted to be implemented in any form, and thus lacks the practical evaluation.

• The main purpose of this research is to evaluate the effectiveness of GOKM mapping and compare it to a textual summary of a literature review. With the help of a user evaluation, we examine the extent to which subjects understand and interact with the know-how maps and textual representation of the web-mining domain, and identify the shortcomings and strengths of each approach.

1.3 Thesis organization

In the next chapter, the literature related to the central theme of this work is reviewed. Chapter 3 focuses on the research methodologies that we used in the process of structuring this research. In Chapter 4, we describe and define the components of the know-how maps. In Chapter 5, the process documentation for creating a know-how map are outlined with the demonstration of a web mining domain. Chapter 6 describes the GOKM proof-of-concept prototype. In Chapter 7, we elaborate on the evaluation of goal-oriented know-how maps and in the final chapter, we summarize the research with a conclusion and future works.
2 Literature review

In this chapter, we look at the literature related to the research topic. First, we review the knowledge classification related concepts. Second, we look at the characteristics and quality defining aspects of knowledge mapping systems. After that, knowledge visualization and interactions are described. Next on the review, knowledge management processes are explored in the context of knowledge mapping. Next, the work of Nasser (2012) on goal structuring of knowledge domain is reviewed. We also review the goal-orientation framework and i* modelling notations. Lastly, the ScholOnto project is presented as a rival alternative of knowledge mapping, and we describe how that work has inspired and guided the process of evaluation.

2.1 Knowledge classification

Mai (1999) has observed that the main task of knowledge organization and representation is to organize and represent documents for effective accessibility and retrieval as well as building systems to achieve this goal. The objective of any classification system is to mirror the reality as close as possible to be helpful for the users of the classification. The traditional theories of classification, Mai (1999) argued, are centred around the idea that all knowledge is interwoven into a great web. This requires some pre-established connections and relations between all knowledge, and therefore task of a classification system is to represent the web of knowledge.

Rowley and Farrow (2000) have pointed out that reasons for classifying knowledge and documents are twofold. On one side, classified knowledge provides an overview of the information presented in the document. The other side, knowledge classification facilitates knowledge retrieval. Mai (1999) has distinguished and discussed differences between modern and postmodern classification theories. The main distinction between these two theories is that while the modern classification aims at representing the whole of knowledge (“universe of
postmodern classification seeks to provide a pragmatic domain-specific tool for classification.

Mai (2004) argues that any classification is relative in the sense that no classification can be considered as a true representation of knowledge. He further points out that a classification is merely an explanation of the relationships of a given domain that satisfies the needs of a group of people at a time. Mai (2004) believes that the major challenge of bibliographic classification is that of classing document, which represents ideas and thoughts. The ideas and thoughts exist independent of documents and media that represent them in the sense that ideas and thoughts could have been expressed with different words and in different media (Mai, 2004).

Mai (2004) has emphasized that classifier’s goal is to construct a scheme that lays out a view of the knowledge that makes sense to the knowledge users. In other words, the classifier’s task is to interpret the documents and present them in the scheme that the user can use. The work of Nasser (2012) is an example of an attempt to classify the scientific knowledge with goal-oriented structure. The goal oriented structuring of knowledge is an approach to organizing knowledge domains based on means-ends (goal-solution) relationship. The knowledge domain is considered goal-oriented if the knowledge contains means-end relationship that demonstrates the relationship between the goal and/or problems, and the proposed solutions (Nasser, 2012).

2.2 Knowledge mapping

A knowledge map is an artefact with the intention of guiding users through a complex and changing knowledge domain by representing organized knowledge (Wexler, 2001). These maps are a graphical representation of knowledge components and the relationship between them. In
In this sense, the Goal-Oriented Know-How Map is an extension of knowledge mapping, which is based on the problem and solution the concept represent. In their report on the use of knowledge mapping techniques and technologies, Canas et al. (2003) elaborated on distinguishing characteristics of knowledge mapping as follows.

- **Underlying theory**: Canas et al. (2003) referred to the well-established assimilation theory, which posits that new knowledge can be learned and transferred more effectively if it is related to the existing knowledge. This theory underlines that effective learning and understanding of the knowledge domain depends on how the knowledge pieces are related to each other.

- **Semi-Hierarchical organization**: Building on assimilation theory, the basic motivation to represent the knowledge with hierarchical arrangement is that concepts that are more general subsume more detailed and specific concepts. This theory produces knowledge maps with more general concepts placed on top of the map, with more specific concepts stemming down from it (Canas et al., 2003).

- **Labelled links**: Another characteristic of concept maps is the use of defining phrase to link the knowledge pieces together. Linking the knowledge pieces together with the help of meaningful links provide the necessary context needed for making sense of how the knowledge interact and relate to the knowledge domain.

- **Definition of nodes**: In order to have a functional mapping system, one has to establish a clear definition of its components. The nodes and links need to have a clear definition that further help to reduce the misunderstandings as well as providing a sound framework.

Furthermore, Wexler (2001) set out to identify the conceptual foundations that distinguish the useful knowledge maps from less useful maps. These six key points about usability of the
knowledge map are also at the core of GOKM usability and are used to guide the research in the evaluation process.

- **Communication medium:** Knowledge mapping is a designed communication medium between mapmakers and map users. The use of knowledge maps as a communication medium aims to minimize the miscommunication between the knowledge mapmakers and users. This also implies that map users should be able to provide feedback on how to improve the maps in order to bridge the communication gaps.

- **Syntax and semantics:** Knowledge mapmakers and users must share the same syntax and semantics in the process of creating and using the map. The clarity on the definition of knowledge map components and their relationships is crucial in creating an environment, which fosters productive and effective communication among knowledge mappers and users.

- **Problem-Solution:** Knowledge maps are created to solve a problem. This foundation implies that the knowledge map has to have a goal to address if the mapmakers and users are going to continue using them.

- **Self-correction:** Knowledge maps should encourage self-correcting action and learning. Knowledge maps should act as enablers of envisioning future actions and facilitate organizational learning. In other words, effective knowledge maps foster curiosity to ask questions and challenge the surface of the maps and encourage self-correcting actions that lead to learning that is more effective.

- **Information search:** Knowledge maps must direct the search for information, and not end the information search. Wexler (2001) pointed out that in order to add value, knowledge maps must avoid creating dependency and curtailing of innovative or critical thinking, and instead promote further and deeper search of information. Knowledge maps are brief and logical representation of knowledge, and are not meant to be taken as complete knowledge.
• **Information discovery**: Knowledge maps are dynamic and foster information discovery. Effective knowledge map is dynamic in representation of problems as they progress and lead to the development of new problems and solutions. The dynamism of the knowledge map, Wexler (2001) argued, lead to enhancement of information discovery.

2.3 Knowledge representation and visualization

With the advent of digital communication technologies, an enormous quantity of information can be stored and transferred. The accumulation of information creates complexity that ultimately affects the information quality and transfer. Many users of information technology have often experienced the shortcomings of current information accessing, organizing and navigation. Finding information has become more time consuming and more complex the amount of information grows exponentially. There has been many improvements in information visualization with improvements in computer graphics and power; however, many of the theoretical grounds of knowledge domain visualization lack practical foundation.

“Painting a big picture of scientific knowledge has always been desirable for various reasons” (Borner, et al. 2003, p. 2). This desire is the result of complexity that information users face when navigating through the knowledge domain. As the task of information organization and navigation becomes more complex, the quality of information retrieval will decline. Borner et al. (2003) also noted that creating a high picture of evolving scientific disciplines has almost become impossible since the knowledge domain is rapidly changing its dynamics and landscape.

Knowledge domain visualization is a way of mapping out the information landscape. It is a collection of methods that create more effective and efficient use of limited cognitive resources in order to overcome the complexity of the knowledge domain landscape. The goal of knowledge
domain visualization systems is to present information in more readable and organisable structure, as well as communicating the knowledge that lies beneath the mountain of information. Grinstein and Ward (2002) argued that visualization could provide a qualitative overview of large data sets, summarize data, and help to identify areas of interest based on the parameters. In other words, knowledge domain visualization is the way to harness the perceptual capabilities of the human visual system.

Grinstein and Fayyad (2002) have argued that finding patterns or models in data is fundamentally about data reduction. An effective knowledge visualization system should be able to summarize the amount of information being presented to the user, helping the user avoid information cluttering. The domain knowledge visualization in this research aims to summarize the technological domains into entities, relationships and attributes. Knowledge entities and their attributes are visualized based on the goal of the article and presented with their respective relationships to each other. The goal is to reduce the knowledge domain complexity while losing the least amount of structural information (Grinstein & Fayyad, 2002).

2.3.1 Visualization interactions

The information user can interact with the data in a variety of ways. Grinstein and Ward (2002) identified the following user-information needs. First, users want to get the “big picture” which emphasizes an exploratory aspect of domain visualization. In exploratory visualization, user does not necessarily know what he or she is looking for; rather, they can freely explore what is available in the knowledge domain landscape. The exploratory interactions create a dynamic whereby the user can visualize the trends and paths of knowledge and arrive at a hypothesis. In addition to exploratory visualization, visualization can also be confirmatory, where users use the
visualizations to confirm and test a hypothesis. In addition to these interactions, in goal oriented knowledge structure, we attempt to help the users to identify the knowledge gaps. This can help to direct more research focus on the less known areas of a knowledge domain.

Sampling a knowledge domain is another type of visualization interaction. The goal of sampling is to reduce the size of information so that the user can more easily navigate through the information. This eliminates the cluttered information mappings and reduces the complexity of knowledge domains.

Another form of user interaction with visualized knowledge domain is ad hoc querying where the user can make direct requests for visualization of a special part of a domain. This is an extension to the text based search engine where user finds textual information by making a query. Associative interaction is another form of interaction where users can access the related data. Associative interaction is the essential part of scientific domain visualization as it is in the nature of scientific endeavour to build knowledge incrementally.

Lastly, as part of visualization interaction, we have the user codification process where the user focuses on creating and refining knowledge domain entries. The user codification process has been one of the main issues of theoretical foundation proposed by Nasser (2012) and remains as a crucial part of this research. The users seek to interact with the knowledge visualization system to add new entries, refined the knowledge elements in order to optimize the visual representation of the domain. The knowledge codification process is the first half of the challenges we faced in this research. Knowledge visualization requires a certain set of data formats and relations that has to be captured at the user entry level. Without an adequate user codification process, knowledge visualization would be impossible since the relations and contexts that turn raw data
into knowledge resides with the user. Capturing the context is the essential step of creating meaningful knowledge domain visualization.

### 2.4 Knowledge management processes

Alavi and Leidner (2001) developed a systematic organizational knowledge management process framework that can be helpful for analysis of know-how mapping and its supported processes. This framework holds that organizations as knowledge systems consist of four sets of knowledge processes: (I) creation, (II) storage/retrieval, (III) transfer, and (IV) application. Although these processes are intended to describe the knowledge management processes, in this research we use them to clarify the processes closely related to knowledge mapping. These four processes are described and related to the needs of a mapping approach in the following.

The first stage, creation, involves the development of new content or replaces the existing tacit or explicit knowledge within the organization. This is done through social and collaborative processes of individuals in an organization, which lead to the creation of new knowledge. The scientific domain, just like a universal organization, is constantly going through this process to create new knowledge in different fields of study. In knowledge mapping, new knowledge is not created, rather existing knowledge is used to map the domain.

Knowledge tends to be lost if not captured and stored properly. Thus, it is important for organizations and scientific domains to store their accumulated knowledge to be able to retrieve it later. The storage and retrieval process of a knowledge mapping approach can be facilitated with the development of software solutions. The modern database technologies have provided sophisticated storing and retrieval techniques for knowledge storage and retrieval.
Knowledge transfer is a crucial part of the knowledge management process, where an organization makes sure the knowledge is available to locations where it is needed. In most organizations, knowledge is transferred by means of communication and information flows (Alavi & Leidner, 2001). Knowledge transfer is one of the focal points of knowledge maps. The means and quality of knowledge transfer are at the centre of an evaluation of this research as we aim to evaluate how the knowledge map does transfer the scientific knowledge.

Lastly, Alavi and Leidner (2001) pointed out knowledge application, which emphasizes the importance of application of knowledge in the organization. Scientific knowledge is valuable when it is applied in practice or research. Knowledge maps are one of the means that enable the application of knowledge by facilitating the knowledge discovery. Knowledge maps can also promote innovation and new research by enabling knowledge gap discovery.

In this research, we have used a variation of Alavi and Leidner (2001) framework to guide the evaluation of the GOKM approach. By identifying the knowledge management process, we are able to envision the same set of processes in the context of knowledge mapping. As a result, we are able to focus on the knowledge mapping processes, and evaluate how GOKM can address the needs of each process in technological domains.

2.5 Goal structuring of knowledge domains

In this research, we primarily consider the work of Nasser (2012), in which she introduced a goal-structuring format for the knowledge domains. In her work, Nasser (2012) studied problem-solution based knowledge domains (i.e. technological domains), and argued that the means-ends structure of such domains cannot be fully explored and organized by concept-based classification methodologies and keyword based classification.
Hui et al. (2004) also points out that the concept-based classifications are efficient in addressing document titles, authors’ names and subject areas. The focus of the goal structuring of knowledge domain work was on structuring knowledge domains by showing the means-ends relationships that exist between the contents of the scholarly articles within a problem-solution based domains rather than classifying the available literature in such domains by using a concept-based classification methodology. The advantage of using means-ends relationship is visible in technology domains where a problem is studied and solutions proposed respectively. However, this does not imply that the goal structuring classification can replace concept-based classification as it cannot be applied to empirical based studies where authors intend to prove a point through direct or indirect observation and experience. The goal structuring classification only aims to illustrate the relationship between the problems addressed in a domain.

The hierarchical goal-structuring model that was developed in this work aims to assist newcomers to a domain in educating themselves more effectively with the knowledge domains. It also aims to support researchers in their research by allowing them to visualize the available literature within a domain, and identify the knowledge gaps, thereby helping them in their future research.

The main components of the developed model (Figure 1) include the research goals within a domain, the articles’ goals that correspond to the research goal, sub-goals that are raised by the articles’ goals, solutions that are proposed to them, the effects of the solutions on non-functional goals, the functional problems that a solution introduces, the assumptions that the solutions are based on, and the references to the solutions where the solutions are thoroughly discussed.

Figure 1 shows the meta model developed by Nasser (2012). This meta model has been used as a baseline for the GOKM components. We used this baseline to make some improvements and
simplification throughout this thesis. Some of these improvements include changes in modelling techniques to less clutter the map, as well as fundamental changes in definitions of knowledge components. These improvements and changes are described throughout this thesis.

Figure 1: Merged meta model Nasser (2012)

2.6 Goal-oriented framework and i* modelling

i* is an agent-oriented and goal-oriented modelling framework that can be used for a variety of purposes including: requirements engineering, business process reengineering, organizational
impact analysis, and software process modelling (Lapouchnian, 2005; Yu, 1997). i* supports the modelling activities that take place prior to system requirements are formulated and therefore it can be used for early and late phases of the requirement engineering process. Goal-oriented requirement engineering centres around the problem (goal) rather than on one particular solution (task) (Wieringa, 2004). This approach provides the i* framework with a keen ability for analysis of the systems where the system design alternatives can be comparatively analysed based on the stakeholder goals.

The focus of i* as an agent-orientation approach is on the interdependency of the actors. As Hilts (2011) has observed, each actor has internal goals and carries out tasks and depends on one another for the resources and tasks to achieve their goals. The concept of actor interdependency is well established in i* framework. As we go further in this research, it will become more apparent how the interdependency concepts can be slightly altered to represent the interdependency of knowledge elements based on goals and their respective solutions in a knowledge domain.

The flexibility of i* framework has made it possible to envision a goal-oriented knowledge structuring using its modelling notations and concepts. The use of these notations in structuring knowledge will be explained in more details in later chapters. Nasser (2012) and Hilts (2011) are examples of attempts to integrate the concepts of goal-oriented requirements engineering with knowledge structure representation. Below is the i* framework syntax which will be used for representation of the knowledge domain structure.
2.7 Scholarly Ontology Project

Researchers are benefiting from the worldwide growth of access to digital libraries and electronic journals in recent decades. However, the increased access to information and rapid growth of information bring on its own challenges. In their effort to improve the process of “sense making”, in which individuals and groups construct meaning out of complex information, Shum et al. (2000) seeks to provide more effective discovery of relevant documents, as well as enabling structured discourse among researchers. The individuals and groups always strive to understand which ideas are being claimed as new, and assess their importance and consistency with the domain.

Shum et al. (2000) has observed that researchers have limited ways of articulating many of their interested phenomena. These phenomena, exemplified with queries, are things like: the impact of ideas, different perspectives on the issues, inconsistencies with the established and new ideas, and convergences of other research. The focus of ScholOnto (Shum et al., 2000) is to facilitate answering these questions by developing a methodology that supports research communities in the interpretation, and discussion of new and established ideas. In a sense, the main goal of ScholOnto is to supplement the information presented in the documents.
In later work on ScholOnto, Uren, et al. (2006), articulated the main aims of the Scholarly Ontologies Project. They started by pointing out that their approach is capable of representing not only the consensus among the ideas, but also principled disagreement. They argue that this capability enables different interpretations of claims by mapping the discourse relations of the knowledge domain. In ScholOnto, the concepts are captured as short pieces of descriptions on the contribution of the idea. The relations are classified into groups with similar rhetorical association: supports/challenges, problem related, taxonomic, causal, similarity and general. Furthermore, each of the relations is assigned a property to indicate whether it has positive or negative implications of the concepts.

Figure 3: ScholOnto link type ontology (Uren et al., 2006)

Figure 3 shows the link type ontology of ScholOnto project. Unlike ScholOnto approach, which offers a larger set of link types, the GOKM approach presented in this thesis focuses only on
problem-solution relations. The main aim of GOKM is to explore the means-ends relationships, which is a subset of what ScholOnto encompass.

In this thesis, we are facing the similar challenges that the ScholOnto project faced. The goal of this research is similar to that of ScholOnto in organizing the knowledge produced in domains by encoding he documents based on their main concepts including the problems, solutions, and their impacts. However, we focus on the use of means-ends and decomposition relationships to map out the concepts across the knowledge domain. The supplementary contribution links (i.e. Help, hurt, and unknown) are also used to give more meaning to the association of the concepts.

In order to evaluate our approach to knowledge mapping, we use the evaluating approach of the ScholOnto project presented in Orion et al. (2006). This evaluation approach involves creating a domain goal-oriented know-how map and written review of a domain. Uren et al. (2006) provided us with a structure and guideline on how to conduct a controlled experiment to evaluation GOKM.
3 Research Methodology

In this section, we turn our focus to the research methodology. First, we identify and describe the iterative methodology required for development of this research. After that, we refer to guidelines on design science in information system research explained by Hevner, et al. (2004). These guidelines provide a clear road map that design science research can follow to achieve its goals with less ambiguity.

3.1 Information system research: design science

In this work, we are adopting the design science paradigm, and we aim to follow the model and guidelines proposed by Hevner et al. (2004). The nature of this work is exploratory with the aim of innovating new ways of organizing and representing the technological domains. Hevner et al. (2004) pointed out that design science is inseparable side of information system research cycle, which create and evaluate IT solutions and artefacts. In the case of this research, we have devised an experiment to evaluate the GOKM approach and compare it to the written review of a knowledge domain. The comparative analysis is drawn from the experimental results to evaluate the strengths and shortcomings of GOKM knowledge maps and more traditional literature review.

Nunamaker et al. (as cited in Hevner et al., 2004) observed that the process of constructing and innovating IT artefacts creates opportunity for design science research to explore and understand the problem domain. In our research, the construction of the experiment and process documentation brought many insights and helped to explore and navigate the problem domain, and address the requirements that arise from those insights.

In the following, the seven guidelines proposed by Hevner et al. (2004) are briefly described and explained how they were followed throughout this research.
3.1.1 Guideline 1: Design as an artefact

In this work, we developed three artefacts. The first artefact is the GOKM process documentations and descriptions. The GOKM process documentations are created in order to guide the users of GOKM to create, and use GOKM models based on the knowledge domain. The documentations were created based on the processes that were followed by the researcher to create the web mining domain GOKM models. In the description part of the process documentation, each of the knowledge components is defined and explained to create a clear picture of the GOKM approach.

The second artefact created in this thesis is the goal-oriented know-how map of the web mining domain. The domain map is considered to be an artefact since there has never been an attempt to map this domain, and it is the first time that principles of GOKM are being applied in this domain.

We have also designed a proof-of-concept prototype for GOKM. As Hevner et al. (2004) has also explained, the prototype has helped to demonstrate the feasibility of both the design process and designed product. Using this prototype as IT artefact has helped to get insights about how such a system can be designed and implemented. The design of such a prototype helps to narrow down the gap between theoretical and practical development of this field.

3.1.2 Guideline 2: Problem relevance

We have identified and elaborated on the problem domain in Chapter 1. This is followed by gathering information in the form of a literature review in Chapter 2. While trying to narrow the gap between theoretical and practical development of information retrieval and classification methodologies, we also evaluate the goal structuring framework (Nasser, 2012) and propose improvements. These evaluations will be explained in later chapters.
3.1.3 Guideline 3: Design Evaluation

In summary, in this work we apply the GOKM concepts to a technological (web mining) domain to evaluate the applicability of the know-how mapping in different domains. Furthermore, the GOKM approach is evaluated and compared to a written review of the knowledge domain. The strengths and shortcomings of each approach can be analysed with comparison to the other approach. In order to avoid repetition, the evaluation process will be explained in more details with results in later chapters.

Furthermore, a preliminary evaluation of the prototype was conducted by the researcher playing the role of a user of the system. The iterative nature of prototype development, and the feedback from the researcher’s supervisor was the main source of evaluation and improvement of the prototype. However, the main goal of the user study in this research was about evaluating the general GOKM approach to knowledge structuring, and hence the user evaluation of the prototype is not within the scope of this research. The prototype was developed as a proof-of-concept to demonstrate the feasibility of the system that supports goal oriented know-how mapping.

3.1.4 Guideline 4: Research Contribution

In this research, the experimental results are the most important research contribution. With the help of experiment, new insights emerge about the GOKM approach as well as how it stacks up compared to a more traditional and well-known textual representation.

The GOKM process documentations are also an important piece of this research, where the foundations and methodologies of GOKM are explained. The documentation provides the necessary information for the users to create and browse the knowledge domains with the help of GOKM.
In this research, we have also designed the prototype as supplementary design artefact. The prototype contributes to this research by exploring and developing the technical side of theoretical developments in the field. The prototype has also let us get a deeper insight on the implications of goal oriented know-how mapping.

3.1.5 Guideline 5: Research Rigour

The outlining methodology followed in this research is iterative design. This methodology divides the project into smaller deliverable tasks. The rapid and short iterations make sure that the issues of every task are identified and handled sooner in design and parallel to the development of the process documentations and experiments as well as the prototype development. The other part of research rigour is the evaluation, which has a dedicated a section, explaining it in details.

3.1.6 Guideline 6: Design as a search process

Throughout this research, we have decomposed the huge tasks of design, development, and evaluation into smaller, more manageable tasks. The research was framed in generate/test cycles, where an idea about user interaction with the knowledge domain was first generated, then developed into potential scenario with real domain examples, and finally put to test to evaluate its strength.

3.1.7 Guideline 7: Communication of research

The process documentation and the descriptions are created to communicate the functionalities and specifics of GOKM to the end users of the knowledge domains. The process documentations and descriptions describe the tasks and activities required from creators of GOKM maps its end users.
As for the GOKM prototype, we have included screenshots of the prototype to provide the readers with the look and feel of the system. The flow of the prototype is indicated with the descriptive screenshot.

### 3.2 Research methodology steps

Following are the steps we took to in carrying out this research.

#### 3.2.1 Mapping web mining domain

In this research, we aim at exploring and assessing goal-oriented know-how mapping by applying it to a new knowledge domain. We first take the task of mapping a knowledge domain. This step starts with finding a technological knowledge domain which expresses the means-ends knowledge. The web mining domain is the technological domain that has the means-ends relationship and thus is suitable for know-how mapping. Creating the web mining know-how map involves a great deal of research to gather the publications to codify.

Knowledge mapping can be done in various ways. Following the top-down approach, publications are mapped using a source that has already made some kind of a mapping (e.g., books). Using that approach has the advantage of having a collection of related knowledge (knowledge domain) that can be used to map the domain. Adopting the bottom-up approach means that the mapping begins with a single source and as additional source are being mapped the domain map evolves. In this research, we took the top-down approach, where we used Liu (2007) as primary source of knowledge and mapped web mining domain. The task of mapping the web mining domain is an attempt to understand the challenges of know-how mapping, and document the steps involved in details.
3.2.2 GOKM process documentation

The process documentation is the detailed record of the steps we took to map the know-how of the web mining domain. These steps are documented with details that can be used in the process of creating and using GOKM.

3.2.3 GOKM prototype

The proof-of-concept prototype is developed to envision a system that can support GOKM. The prototype was developed following the initial requirements identified by the research team, and undergone many iterations to refine the requirements and its look and feel. Due to exploratory nature of the prototype, the initial requirements identified are based on the desired features that the research team expected in the functioning system. These requirements were further refined as the prototype development progresses.

3.2.4 Evaluating the GOKM approach

Lastly, we conducted a controlled experiment with twelve participants to evaluate various aspects of GOKM features such as: finding problems, finding alternatives, evaluating alternative solutions, evaluating solutions’ effects, and finding knowledge gaps. The details of how the controlled experiment is carried out are discussed in length in chapter 7.
4 Goal-Oriented Know-how Mapping (GOKM)

In this chapter, we will focus on discussing the essence of this research where all the necessary components of the work are gathered together to make the final artefact. This chapter will include the conceptual discussions on the goal oriented know-how mapping component and its approach to structure of knowledge based on the existing means-ends relationships within a knowledge domain.

First, an introductory description of web mining knowledge domain is provided to familiarize the reader with this sample technological domain. From the description of the knowledge domain, we move to the conceptual model of GOKM, where the conceptual relationships of the GOKM components are identified. Next, the GOKM components are individually explained with the help of examples from the web mining domain. The description of GOKM components includes definition, and the relationship with respect to other components.

4.1 Web Mining knowledge domain description and Know-how Map

Web mining is the branch of data mining that aims to discover and retrieve valuable information and knowledge from the web hyperlink structure, web page contents, and web usage data (Liu, 2007). Web structure mining focuses on discovering useful knowledge from hyperlinks (links) which form the web structure. The web structure has been proven to be a vital component of search engine technologies (e.g. Google’s PageRank).

Web content mining is the application of data mining techniques to discover knowledge from web page contents. Classification and clustering of web pages according to their topic, and extraction of specific information from the web pages are examples of web content mining activities. Analysis of data usage patterns, as a by-product of web mining, can help organizations
better understand the needs of their clients and ultimately serve the information needs of their clients more effectively.

Many of the web mining processes are similar to data mining processes; however, there is a difference in the methods of data collection. Data collection processes in web mining are referred to as web crawling, a process that automatically discovers new web structure and the contents.

The web mining knowledge domain is very much based on applying the lessons learned from traditional data mining domain to capture useful information from the World Wide Web contents (Liu, 2007). However, one can soon realize that web mining cannot be limited to traditional data mining, and as the complexity and volume of the WWW content increases, more new and complex solutions are needed to keep up this exponential growth. In examining the web mining domain, we identified three abstract research goals: (i) web information to be prepared for web mining, (ii) information to be searched and retrieved, (iii) web be browsed for web indexing. The high level goals identified here were based on reading multiple papers in the web mining domain literature. After identifying the initial goals from those papers, the goals are refined through iterations (described later in the process documentation) in order to cover other related literature in the domain. Based on these abstract goals, we aimed to capture as much as possible about the solutions that have been proposed to achieve the goals.

Figure 3 shows the entire Know-How map that was codified from web mining literature for the purpose of this study. This model is used as an example to explain the features of GOKM throughout this research.
4.2 GOKM components

Figure 4 is the map of GOKM components, which summarizes the relationships of the components with respect to each other. Figure 4 presents a simplified and summarized version of Figure 1 for a single publication. A knowledge domain includes many articles. These articles are bound together in a knowledge domain by the common subject being studied. The article has a goal, which is an objective set by the author to provide a solution to the problem. The solutions in GOKM are proposed in response to the goals; however, as it is sometimes the case, the solutions themselves introduce new goals in the knowledge domain by the task-goal specification (see 5.1.4). The goals that are introduced by the solution are part of a scholarly endeavour where new ideas are generated in response to finding a solution to the goal.
Furthermore, the solutions are evaluated with the help of soft-goals, which are essentially non-functional requirements of the solution. Soft-goals are the quality properties of the solution, which helps to evaluate the solution. Solutions are also based on a set of assumptions. These assumptions define the mindset that led to the proposal of the solution.

4.3 Goal Oriented Know-how Mapping components

In this section, each of the GOKM knowledge components is defined, and the approach for extracting them is discussed. The knowledge component is a piece of knowledge that can be identified as meaningful pieces of information, which makes up the knowledge domain collectively. Yang et al. (2013) have also observed that if publications can be divided into several knowledge components, then the reader can efficiently make more specific search on the knowledge domain. By dividing the documents into several knowledge components, which correspond to a more specific type of information, the time required for knowledge users to search specific domains can be reduced.

The knowledge components of GOKM include: goals, solutions, assumptions, and soft-goals. We discuss each of these components with their relationships to other knowledge components. The relationships described for each knowledge component is with respect to the component that it is being connected to. This way we can describe components relationship with other components without creating a complex mesh of relationships.

Goals (hierarchy and structure)

There are a number of goal definitions in the requirements engineering literature that can be used to describe the goals in the context of goal oriented know-how mapping. In an early description, Anton et al. (1994) described the goals as high-level objectives of the business, organization or
system, which capture the reasons why the system is needed and doing what it is doing. In the requirements engineering field, goals are used as decision guided means that provide grounds for justification and analysis of various components of the enterprise or the system. Van Lamsweerde (2000) also provided a definition of goal, where the goal is defined as an objective that the system should achieve through cooperation of the system components. This definition emphasizes the intentionality of the agents (e.g. actors, systems) and their collaboration to achieve the desired objective.

Goal-oriented know-how mapping is also using the concept of the goal to identify the objectives of the publication. A publication is set to achieve a goal(s), an intention to achieve an objective, which is set by the author. The goal concept in know-how mapping is extended to accommodate three levels of goal for each publication. The three levels of goal are distinguished in GOKM in order to provide a more meaningful knowledge representation and knowledge component categorization.

The key factor that allows GOKM to distinguish different levels of goal is the level of abstraction and intentionality. The (research) goal, (article) goal, and the goal are all representative of the publication’s goal at different level of abstraction. The (research) goal, with the highest level of abstraction is at the top of the goal abstraction hierarchy, followed by (article) goal and goal at the lowest level of abstraction. We have used the terminology of “research” and “article” to provide a context for the users of GOKM. The use of these terminologies should not be confused with the separation of goals, rather it is meant to help the GOKM users to codify the publications with consideration of different goal abstraction level.
The purpose of GOKM in distinguishing three levels of a goal is not to introduce a new knowledge component, but rather it is intended to provide more a detailed and logical representation of knowledge domains. The knowledge domains are made of various publications, each of which has their own distinctive set of goals. The abstraction level of goals in GOKM makes it possible to find a place for categorization of publication in a knowledge domain based on their goals. The hierarchy of the goals is essential in visualization and modelling of the domain. The know-how maps have the hierarchy, which was discussed in Canas et al. (2003) as a characteristic of knowledge maps. Single publication goals are merged together on different levels of goals to form the knowledge domain map. The different levels of goals and their use are described in more detail in the following section.

4.3.1 (Research) Goal

The first step taken after selection of a knowledge domain is to identify the research goal of interest to map the related publications. For the purpose of know-how mapping, a knowledge domain is usually identified with one research goal per domain map; however, a knowledge domain can have multiple research goals. The knowledge map, just like any other type of map, is limited in the amount of information it can present.

*Definition:* Research goal is the abstracted goal of the publications in a knowledge domain, which shows the general motivation of the knowledge domain. Research goal is supposed to be abstracted and may or may not be explicitly stated in the article. User may need to comprehend the research goal from the article goal. Research goals are identified and reported in the publication. We identify a research goal in order to map all the related publications to that research goal. Research goals can be helpful in a mature knowledge domain, where research can be branched into separate, sometimes mutually exclusive paths. In this research, *Information to*
be extracted from web data is taken to be as research goal, indicating that we are focusing on web mining research within the web mining domain.

Relationships:

- A research goal has no outgoing relation with any of the knowledge components. This is because only the next immediate knowledge component, article goal, has a relationship with research goal. More on this relationship is provided below.

4.3.2 (Article) Goal

Definition: An article goal is the abstracted goal of the publication. Article goals are usually stated clearly in the article, mostly in the abstract or introduction section of the article. The article goal is differentiated from goals in the level of abstraction. Article goals are more specified to the article compared to the research goal. At the same time, article goal has higher abstraction level to capture the related articles in the hierarchy. For example, web to be browsed for web indexing is an article goal that encompass articles which share the goal of exploring web crawling within web mining domain. This differentiation can create a meaningful structure for the knowledge domain model. In theory, though very rarely, a publication can have one or multiple article goals, which themselves can be further detailed into multiple goals.

The level of abstraction is a crucial aspect of establishing the article goals. At first glance, it appears that each article requires, as the name implies, at least one article goal. However, as we have observed similar to Nasser (2012), adjusting the level of abstraction on article goals can provide a coherent categorization of the publications, which are addressing the same abstracted goal. Adjusting the level of the abstraction of the goal can be done by making it more abstract to encompass more articles, or by branching off and creating a separate article goal for more
specific articles. Article goals tend to have low variability under the research goal, which is due to the flexibility of abstracting the goal that the article is trying to achieve.

In the web mining knowledge domain, we have identified three article goals: 1) web information to be prepared for web mining, 2) information to be searched and retrieved, 3) web to be browsed for web indexing. These are the abstracted goals that the publications chosen for this knowledge domain set out to achieve. In web information to be pre-processed, the article’s goal was defined in a way that can capture various articles relating to the preparation stage of information retrieval. As we dig further down into the domain, more established goals emerge, each of which deals with a more specialized aspect of pre-processing raw web page data.

**Relationships:**

- The relationship between the article goal and research goal can be means-end or decomposition relationship. This depends on the formation of research goal and the nature of the knowledge domain. The satisfaction of an article goal may be sufficient for achieving the research goal, which requires “means-ends” relationships. However, as demonstrated in the web mining knowledge domain, the research goal’s (information to be extracted from web data) satisfaction depends on the satisfaction of all three article goals. See Figure 5.

- An article goal can only belong to one research goal in the domain map. This is due to the fact that goals that are set by the article do not stray from the research goal they set to achieve.
4.3.3 Goal

**Definition:** Goals are defined as sub-sections, or detailed article goals. Goals have the lowest level of abstraction in the goal component hierarchy. The goals themselves can be further decomposed and refined into more detailed goals and should be defined in a way that clearly indicates the objective of the publication. In GOKM, the solutions are directly proposed to achieve the goal, which demonstrate the means-ends relationship. Goal formation is an important part of GOKM, which is discussed in more detail in the process documentations (refer to 5.1.3).

In web mining domain, we have defined three goals that directly go under *web information to be prepared for web mining*. These goals are 1) *Text and web pages to be pre-processed*, 2) *Web formatting to be removed from the data*, and 3) *Information blocks to be extracted from web pages*. These goals have been refined to capture the related publications under one branch, which grows with the development of the knowledge domain know-how map. Most often, the goals have to go under task-goal refinement described in section 5.3.2.2, in order to provide more detailed context for the means-ends relationships of knowledge domains.

**Relationships:**
- Goals can have means-ends relationship with the article goal in order to demonstrate the alternatives of achieving the article goal. The mean-ends relationship between goals is not part of the i* modelling approach; however, in context of GOKM, a goal might be satisfied by a subset of its alternative following goals.

- Goals can be introduced by and attached to a solution by decomposition or means-ends relationship. This is the task-goal specification that is described in 5.3.2.2. An example of means-ends relationship between the goals and solution is where Hyperlinked web to be analysed, and Information authorities to be discovered are connected to the Link structure algorithm solution. (See the figure below).

![Diagram](image_url)

Figure 6 Demonstration of means-ends relationship between goals and solutions.
4.3.4 Solution

**Definition:** Solutions are the tasks that are proposed for achieving a goal. There may be multiple solutions to a goal, which means there is more than one way of reaching the goals. In any knowledge domain, there are solutions that are being discussed in multiple publications; however, there should be only one instance of that solution appearing the GOKM model in order to maintain the know-how map information integrity. An article might have multiple solutions.

The solution can also raise a common goal that would be shared among the alternative solutions. The shared goal specification is used in this case to clarify the goal that the solution is sharing with other solutions. In GOKM, a combination of solutions is considered as a solution on its own, where the combination of solutions is demonstrated with the help of introducing the shared goal specification.

In web mining knowledge domain, we have identified many solutions addressing the goals set forth by the publications. Under the Web to be browsed for web indexing, we identified two main goals (Web pages to be discovered universally and Web pages to be discovered preferentially), and a goal (social networks to be discovered) that has not yet been explored. The last goal at the leaf node, that has no potential solution is what we refer to as knowledge gap.

*Universal crawler* has been proposed as a solution to Web pages to be discovered universally. On the other branch of Web to be browsed for web indexing, *Topic-specific crawling* and *Topical crawler* are proposed to Web pages to be discovered preferentially. *Focused crawler* and *Context-focused crawler* are different solutions for topic-specific crawling, which share the goal of Web crawling to be topic specific. See the figure below.
Relationships:

- The solution can have means-ends relationship with goals. This relation indicates the alternatives of a particular goal. A solution might belong to multiple goals, which means the same solution can be used for realizing different goals.
A solution may have an effect on soft-goal, and it is connected to it by contribution links (discussed more in the following).

A solution may have many assumptions (discussed more in the following).

4.3.5 Soft-goal

**Definition:** Soft-goals are used as evaluation criteria of solutions. Soft-goals, or Non-Functional Requirements (NFR) in the language of software requirements engineering (RE), are helpful for evaluating the qualities of tasks and their effects. NFRs are usually represented in RE models by soft-goals, which unlike the regular goals, are seldom considered accomplished or satisfied completely (Lapuchnian, 2005). Soft-goals lack the clear-cut satisfactory condition, and therefore are accompanied with the notion of satisficing. In the context of GOKM, soft-goals are qualitative values that are attached to the solution, which describe benefits, or shortcomings of the solution.

In the web mining domain, **Stemming, Duplicate detection, and Stop word removal** are solutions identified for *Text and web pages to be pre-processed*. **Stemming** has a *Hurt* impact on precision, as well as *Help* contribution to *Recall* and *smaller index*. In this demonstration, we can see that *smaller index* is shared between multiple solutions in order to avoid duplication of soft-goals. “*Stop word removal*” (*Help*) “*Smaller index*” and “*SEO*”, but (*Hurt*) “*phrase search*”. The solution contribution to the soft-goal is extracted from the publication, where the author explains how the solution have an effect on the soft-goal. For example, the original text that was used to extract some of soft-goals for stemming is: “Clearly, stemming increases the recall and reduces the size of the indexing structure. However, it can hurt precision because many irrelevant documents may be considered relevant. Stemming enables different variations of the word to be considered in retrieval, which improves the recall.” (Liu, 2007, p. 228)


**Relationships:**

- Soft-goals are affected by the solutions, which can be shown with the contribution links.

The contribution links used for GOKM are the same definitions for i* framework, which are:

- Some+/-: indirect form of contribution where solution has a slight effect on the soft-goal.
- Help/hurt: the solution is helpful or hurtful to the soft-goal.
- Make/break: strongest form of contribution where the solution has distinguishable and definitive effect on the soft-goal.
- Unknown: the contribution is unclear based on the available data.
4.3.6 Assumption

Definition: An important part of understanding the solutions along with their associated goals requires an understanding of the assumptions on which the solutions are based. An assumption might be shared among different solutions. Solutions can also be based on many assumptions. Extracting useful assumptions in a knowledge domain requires some degree of judgment from the knowledge codifier, where he/she has to make a judgment call as to whether some statement can be considered as a useful assumption that needs to appear in the knowledge domain. The judgment call entails thinking about the clarity of the domain map, value that the assumption adds to understanding of the solution, and degree of expertise the map users have. This judgment greatly depends on the degree of familiarity of the knowledge user with the knowledge domain. Assumptions can also be used for evaluating solutions. Validation and evaluation of the solution based on the assumption is based on the knowledge user’s familiarity with knowledge domain.

In a web mining domain, we identified predefined template schema exist as the assumption that was used for proposal of Automatic information block extractor.
**Relationships:**

- Assumptions are connected to the solutions with association links as shown in figure 9.
- Multiple assumptions can be connected to one solution, and assumptions can be shared from multiple solutions.
5 Goal-Oriented Know-How Mapping process documentation

In order to create a set of process documentation on how to create and use GOKM models, we have recorded detailed steps of the codification and modelling process taken to form the web mining knowledge domain. The key issues in each step are systematically addressed in the process documentation. The process documentation is accompanied with examples to help readers to envisage the GOKM concepts and processes.

The goals of the process documentation are: first, help the readers and users understand the GOKM concepts and methods. Second, help GOKM users by creating a detailed manual, which can be used as a reference for utilizing GOKM. With these goals in mind, the process documentation captures detailed steps involved in setting up the knowledge domains with GOKM approach.

The process documentations are organized into four main sections: basic processes, codification processes, modelling processes, and GOKM for research. The first set of processes expands on the basic concepts and steps of GOKM. These concepts range from the explanation of means-ends and decomposition relationships, formation of goals and special task-goal specification. The process documentations also cover the preliminary process of knowledge gathering, in which more details of how and what types of publications contribute to the knowledge domain.

In the codification processes, the focus is the process of extracting the knowledge components from the publication. Each of the knowledge components has been described and characterized in the previous chapter, and therefore the process documentations focus on the process of extracting and formatting these components.
The modelling process concentrates on visualizing the knowledge domain by modelling the knowledge components codified in previous steps. The modelling processes are divided into two sections: first is the process of modelling of a single publication. The second part of this documentation is dedicated to the process of aggregating a single publication model into the larger knowledge domain model.

Lastly, the GOKM for research processes focuses on how the GOKM models can be interpreted and used by the knowledge user. These steps point to the questions and explain how and from where the questions can be answered using GOKM.

**5.1 Basic processes**

The basic processes are intended for the introduction of basic modelling techniques and concepts used in GOKM. These processes also include the preliminary process of knowledge gathering, which is the starting point of GOKM.

**5.1.1 Knowledge gathering**

Before any steps can be taken towards GOKM, we need to identify the appropriate knowledge domain. The knowledge domain represents a collection of publications within a related domain of interest. A collection of publications in a domain can be identified using two general approaches. First is top to bottom where an existing source is used to gather all the publications. An overview book, annotated bibliography, or the researcher’s implicit knowledge of the domain can be a good source of gathering domain related publication. Another approach to publication gathering is where the researcher builds the collection from the bottom up; meaning the collection is built as the researcher navigates through the domain.
The following steps can illuminate the process of information gathering.

1) Knowledge domain identification: The knowledge domain that can be used for purposes of GOKM needs to be “problem-solution” based. The problem-solution based knowledge domain allows us to study and evaluate the means-ends relationships that arise from solutions to the problems. The one category of knowledge domains that best fits the problem-solution criteria are technological domains, where explicit solutions are proposed for explicit problems. In this research, Web Mining is chosen as a knowledge domain, where new solutions are developed and proposed for the domain problems.

2) Identify problem-solution publications: It is important that the publications of the knowledge domain are also problem-solution based. The problems and solutions presented in a publication enable the study of means-ends relationships. A candidate publication should have stated explicitly, or sometimes implicitly, the problems that are being addressed with explicit solutions.

5.1.2 Means-ends vs. decomposition relationships

The means-ends or ‘OR’ relationship (Figure 11) in goal-oriented know-how mapping is representing alternatives. This means that satisfying one of the alternatives is sufficient for satisfying the parent goal. Means-ends relationship allows for refinement of the alternatives that exist for the goals in knowledge domain, and therefore provide a richer model of the knowledge domain. In other words, the solutions are the “means” of achieving the desired “ends”.

The decomposition or ‘AND’ relationship is the way of identifying the relations where the satisfaction of all goals or solutions are required for satisfaction of the parent goal. This is a more
restrictive form of relation, where all of the solutions need to be achieved, in order to achieve the objective of the parent goal.

![Means-ends Decomposition](image)

Figure 11: Means-ends and Decomposition links

### 5.1.3 Knowledge goal formation

The goals of a publication are embedded in it, and therefore need to be extracted and refined for the GOKM. This requires two steps: to first identify the goal, and second, refine the extracted goal to the goal-oriented systematic format.

A goal is an objective that the publication set to achieve. The goals may be formulated at different levels of abstraction, which ranges from high level, dealing with strategic concerns, to low level, dealing with more detailed and technical concerns (Lamsweerde, 2001). The flexibility of goal formulation with level of abstraction, and motivation is used for distinguishing different levels of goals, which is discussed in length in 4.3.

The main goal structure in requirements engineering (RE) has been established as “X to be Y”. This format is originally taken from goal-oriented requirement engineering literature. This format was adopted as it explicates the notion of means-end which is a central part of the adopted goal-oriented approach. However, as Nasser (2012) has observed, the “X to be Y” needs
to be extended to accommodate the needs of knowledge management. Nasser has proposed that goals can be formatted in “X to be Y preposition Z” which can provide more contexts to the goal structure. The two approaches cannot be considered mutually exclusive, but rather can be used to provide a more meaningful and more specific goals. Table 1 provides examples of these two goal formatting techniques.

<table>
<thead>
<tr>
<th>The original goal (as stated in the publication)</th>
<th>“X to be Y”</th>
<th>“X to be Y preposition Z”</th>
</tr>
</thead>
<tbody>
<tr>
<td>“This paper addresses this question of how to build a practical large-scale system which can exploit the additional information present in hypertext.” (Brin &amp; Page, 1998)</td>
<td>Hypertext link structure to be analysed</td>
<td>Hypertext link structure to be analysed from web pages.</td>
</tr>
</tbody>
</table>

Table 1 Examples of goal formation.

- Note: the formation of goal may be easily confused with the statement of task. The goal should be stated in a way that demonstrates the objective and the desired state of the knowledge component. Goals should not be mistaken for statement of tasks, where the task to be done is considered as a task. In the case of web mining knowledge domain, the difference between well-formed goal and statement of the task was noticed where the initial goal was set to be web to be crawled, but it became apparent that this goal doesn’t express the objectives as the goal should do, rather it is the task to be done. Therefore, this was changed to web to be browsed for web indexing, which is more telling of the objective, and less telling of the tasks.

### 5.1.4 Task-goal specification and its use

Nasser (2012) has observed that when dealing with knowledge components, there is a need for addressing the specialization of solutions, which share the same goal. In this scenario, a solution that is proposed to achieve a goal is specialized, meaning that there are alternatives on how the
solution can be achieved. The solution can raise multiple goals by means-end or decomposition relationship. Figure 12 shows an example of the use of this specification in the web mining domain. In here, automatic information block extractor is a solution that is provided for information blocks to be extracted from the web page. However, the solution itself can be taken from different approaches, which are ContentExtractor and FeatureExtractor. These two solutions share a same goal, and that is primary information to be extracted. Using the task - goal specification, we are able to show where the solutions can share a common goal, where the goal itself arises from a solution.

Figure 12 Example of shared goal specification
5.2 Codification processes

Here we discuss the process of extracting the knowledge components needed for mapping of a knowledge domain. There are four essential knowledge components needed for GOKM: goal, solution, soft-goal, and assumption. In each of these components, we outline the process of what to look for, and what should be the outcome of each step.

5.2.1 Goals

Goal extraction is the bulkiest process of codification in the GOKM. This process is divided into three steps, one for each of the goal levels. We are going to start from the codification of goals, and move on to the article and research goal. Building up the goals from the lowest level of abstraction can give the codification process the flexibility to play with the abstract higher goals, which is the most important part of goal codification.

- **Note:** It is worth noting that the frequency of the goal codification is directly related to the level of the goal, with the goal being most frequent, and research goal less frequent of all goals. This correlation is due to the fact that less new higher-level goals are generated as the knowledge domain grows, and only low-level goals are added to the domain.

**Goal:** As mentioned previous chapter, the goal has the lowest level of abstraction, and is used to demonstrate the detailed objective that the publication addresses. This characteristic of the goal makes them relatively easy to identify in the publication. In Table 1, the original text from the article is shown along with the extracted goal from that passage. The publications usually set their goal, most explicitly, in their abstract or introduction sections. Once the goal is identified, the goal formation techniques should be applied to achieve a well-formed goal.
(Article) Goal: The next step of goal codification is to abstract the goal further to get to article goal. The article goal is not unique to one publication; rather, the article goal aims to cover all the publications that have a more or less a similar abstract goal. In most publications, the article goal may not be explicit, and therefore it should be derived from the codified goals with lower hierarchy. It should be noted that the process of deriving the article goal from goals is greatly improved as the knowledge domain map grows to have established article goals. The established article goals can facilitate the article goal codification by providing a placeholder, into which the publications can fall.

(Research) Goal: Research goals are the highest level of goals, and thus are rarely introduced to the knowledge domain. The research goal is intended to capture the objectives of the domain article goals. That being said, knowledge domains know-how map are expected to have one underling research goal, which is representative of the research in the knowledge domain. It is a good practice to determine the research goal at the earliest stages of GOKM codification (knowledge gathering stage being the best starting point). Research goal is the point of interest in knowledge domain, which captures the related publications. Even though the research goal is codified in the early stages of the knowledge domain codification process, it should not be considered as static; research goals, like other goals, should be treated as dynamic and changed to accommodate the necessary objectives of the research in the knowledge domain.

Know-how domain maps are expected to have one (research) goal in order to reduce the complexity and cluttering of information. This is an arbitrary division of knowledge domain to create a map. Knowledge domains however, are not bound to one research goal.
5.2.2 Solutions

The process of solution codification starts with looking for the solutions that the authors offer for achieving the goal. Solutions are proposed in response to addressing a problem defined as a goal. Solutions are usually easy to spot in the publication since they are what the publication is written about. Solutions are usually stated in the beginning sections of the publication, either in abstract, or introduction section where the authors state the domain problem and their solution for it. For example, in Debnath, Mitra, & Giles (2005) stated “in this paper, two new algorithms, ContentExtractor, and FeatureExtractor are proposed”. From this, one can identify ContentExtractor and FeatureExtractor as solutions that the article proposes to the main goal.

The solutions might appear in different forms; sometimes the solution is given a name by the authors, and sometimes solutions are presented without any recognizable name. In the first case, the name of the solution is simply taken for the purposes of GOKM. However, a solution without recognizable name has to be formatted in the form of a meaningful task (e.g. incremental web page template detection).

*Note:* With the introduction of task-goal specification, a new solution component must be added to the model. This solution has to be broad enough to be meaningful for the subsequent solutions that are sharing the same goal and its general solution. In Figure 13, there is a case where two layers of task-goal specification is introduced for accommodating the publication in the knowledge domain (i.e. template detection, batch web page processing, and page level template detection). Here, there are more than one way to achieve “template detection”, and therefore we introduce the goal of “web templates to be detected” to show the possible solutions that share this goal. Further, “Batch web page processing” solution approach has alternative techniques, and that is
why we introduced a goal “web templates to be processed in batch”, which is a shared goal to the solutions of “Site Style Trees (SST)” and “Shingle”.

The solution in the task-goal specification may not be explicitly stated in the publication, and it is up to the codifying researcher to capture the essence of the solution in the task-goal specification. This process should become clearer with the task-goal specification documentation in section 5.3.2.2.

5.2.3 Soft-goals
The soft-goals are qualities that describe the effects of the solution. In most of the publications, soft-goals are usually followed by the proposal of the solutions. Introduction, contribution, future works, or limitation sections of a publication are a good place to look for soft-goals when codifying an article.

In any knowledge domain, there are many cases where the soft-goals for a solution can be extracted from different articles. This is a case where the publications talk about the available solutions in the domain and evaluate them. The soft-goals for the solution can be extracted from analysis of the solution in other publications.

5.2.4 Assumptions
The assumptions are usually discussed before or after proposal of the solution. The codification of the assumptions does not require the assumption to be in any specific format, which makes it easier than the other knowledge components. However, the assumptions may greatly depend on the discretion of the researcher doing the codification. As mentioned in the description of the assumption in the previous chapter (4.3.6), the degree of familiarity with the knowledge domain plays an important role in deciding what is and should be considered as an assumption.
5.3 Modelling processes

The modelling process is divided into two sections. The first section of the process is dedicated to describing how to map the knowledge component relationships. The second portion of process focuses on the merging process of the aggregating single publication models into the knowledge domain model. The single publication modelling is presented as a separate procedure in the documentation for more clarity; however, in practice, the publication can be modelled and merged into the domain model directly. The output of this process is a single publication know-how map fragment that will be used later on to merge it into the domain map.

5.3.1 A single publication model relationships

This is instructions on how the single publication can be modelled after the codification process. The main components of this process is shown in Table 2, where the relationships between the know-how components are summarized.

The different levels of goal hierarchy are separately shown in this table to provide more detail about the relationships goal’s hierarchical levels. As mentioned in Chapter 4, these goals are not meant to be considered as different types of goals. Furthermore, the means-ends relationship among the goals is an extension of the GOKM approach to i* modelling technique. The means-ends relationships among the goals in GOKM are needed to denote the alternative paths that can be taken to achieve the goals with higher hierarchy.

<table>
<thead>
<tr>
<th>Origin</th>
<th>Relationship</th>
<th>Relationship Type</th>
<th>Destination</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Research) goal</td>
<td>Belongs to</td>
<td></td>
<td>Knowledge domain</td>
</tr>
<tr>
<td>(Article) goal</td>
<td>Belongs to</td>
<td>Means-ends</td>
<td>(Research) goal</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Decomposition</td>
<td></td>
</tr>
<tr>
<td>Goal</td>
<td>Belongs to</td>
<td>Means-ends</td>
<td>(Article) goal</td>
</tr>
<tr>
<td>---------------</td>
<td>------------</td>
<td>------------</td>
<td>----------------</td>
</tr>
<tr>
<td>Solution</td>
<td>Belongs to</td>
<td>Means-ends</td>
<td>Goal</td>
</tr>
<tr>
<td></td>
<td>Raises</td>
<td>Means-ends / Decomposition</td>
<td>Goal</td>
</tr>
<tr>
<td>Soft-goal</td>
<td>Affected By</td>
<td>Help Hurt Make Break Some + Some – Unknown</td>
<td>Solution</td>
</tr>
<tr>
<td>Assumption</td>
<td>Belongs to</td>
<td>Based on</td>
<td>Solution</td>
</tr>
</tbody>
</table>

Table 2: Knowledge component relationships

5.3.2 Goal-oriented knowledge domain

Here we explain the process of aggregation of multiple publications to form a goal-oriented know-how domain map. These steps include goal-merging, introduction of task-goal specification, and soft-goal merging. These processes together can transform the single publication model to fit into the domain model.

5.3.2.1 Goal merging

In Figure 13 (graph on left), the article goal for Yi et al. (2003) has been defined as *templates and contents to be detected*. However, when this article is merged into the domain, the more comprehensive *web information to be pre-processed* article goal replaced the publication goal to provide better context for the knowledge domain. The process of modifying how the goal (all levels of goal) merge into each other, either by change of abstraction level, or by introduction of task-goal specification, is what we call goal merging.
The purpose of this step is to reconcile a single article graph with the know-how domain goal graph. A knowledge user has to approach this step with discretion and make a judgment on whether the goals have enough in common that can be merged into one goal. The merged goal has to express all the qualities that the merging goals expressed. Article and research goals can be merged by altering the level of abstraction the original goals were based on.

Figure 13: Demonstration of goal merging process and task-goal specification. The left portion of the graph is a partial map of Yi et al (2003).
5.3.2.2 Task-goal specification

After an initial goal merging, a knowledge map creator might need to introduce the task-goal specification in order to fit a single publication into the knowledge domain. The process is very much the same as adding task-goal specification to single publication, with the difference that here the user is implementing this specification to accommodate the shared goal for the different solutions in the knowledge domain. The task-goal specification is used to represent the shared goal of alternative solutions. It is very often that different solutions to a similar goal are discussed in different publications, but these solutions are hidden at the time of codifying a single publication. When codifying a single publication, the knowledge map creator has no idea that some other publication is proposing a different solution to the same goal. That is why solutions to same goal might be hidden to codifier. It is therefore necessary to add an extra layer of shared goal for the different solutions in the knowledge domain.

In the web mining domain, Site Style Trees (SST) is proposed for Web formatting to be removed from data (as seen on the left side of the Figure 13). At the time of merging this publication into the domain, it soon becomes apparent that Site Style Trees (SST) has to come after some intermediate goal-task specification to allow for other solutions and goals discovered from other publications. This is because there are alternatives to achieve web formatting removal. To allow for more alternatives to be incorporate into the know-how map we introduce the intermediate goal. In order to make necessary adjustments for merging publications into domain graph, Template detection and Web templates to be detected has been added to add a new layer for alternative solutions. The alternative solutions Batch web page processing and Page level template detection are sharing the same goal of Web templates to be detected. Batch web page processing is further refined with the addition of new task-goal specification (Batch web page
processing \(\rightarrow\) Web template to be processed in batch), which enables us to demonstrate alternatives of batch web page processing.

Need for intermediate task-goal specification becomes apparent in the context of domain, when codifier notices that there are other solution to the same goal in the domain. The addition of task-goal specification is possible as the knowledge domain grows and more publications with their goals and solutions come into light. In Figure 13, the modified branch for getting to Site Style Trees (SST) has been highlighted. The additional task-goal specification, can be used in two ways: first they allow for creating new branches of alternative, as well as serving as complementary information holders that provide better understanding of Site Style Trees (SST) as a solution.

5.3.2.3 Soft-goal merging

Some soft-goals among different articles may be common and have to be shared when articles are being merged to form the domain model. Soft-goal merging is a step where the duplicated soft-goals are merged together and shared between different solutions. The duplication of soft-goals usually happen between the alternative solutions, where the solutions affect same types of soft-goals.

Soft-goals are also subjected to the hierarchy. That means the soft-goals and their contributions apply to the branches of the same hierarchy. Soft-goals hierarchy doesn’t look any different than normal soft-goal. When all the solutions of a goal share a same soft-goal with same contribution, soft-goal can be move up to the base solution (task-goal specification). This feature provides the ability to demonstrate the soft-goals, which are shared among the solutions stemming from original base solution.
It is recommended that in case where the alternatives of a solution (which is denoted by task-goal specification) share a same soft-goal with same contribution, the soft-goal can be moved to the original solution. This optional step helps the model to maintain the hierarchical structure of the soft-goals in the knowledge domain with less information duplication.

In Figure 14, an example of soft-goal merging can be drawn to Crawling efficiency, which is shared, between the focused crawler and context-focused crawler, with different contribution links. Crawling updatability is also shared between these two solutions with a same contribution, and therefore it can be moved up Topic-specific crawler. From the model, it can also be assumed that both alternatives have also been Helping ‘Low search time’ and Hurting ‘training set independent’.
Figure 14: Demonstration of soft-goal merging process and knowledge gap discovery
5.4 GOKM for research

We outline how the GOKM models can be used for research and knowledge retrieval. The evaluation process documentation helps the user to optimize their use of GOKM models.

5.4.1 Knowledge gap discovery

One advantage that GOKM can provide is the ability to identify the knowledge gaps of a knowledge domain. Knowledge gap is described as the goals or problems that have not yet been studied and no solution is available for them. The knowledge gap might occur for various reasons in any knowledge domain; Lack of time or effort, infeasibility, strong status quo are all examples of how the knowledge gap might occur in a domain. The knowledge gap may also originate from the future research sections within the publications, and there has not been a solution proposed to them yet. Knowledge gaps are visually distinguishable in GOKM where the goal is the leaf of the graph with no connecting solutions.

It should be noted that in case of GOKM maps, knowledge gaps might be the result of the incomplete domain codification where the model needs to be completed. Knowledge users should use GOKM knowledge gap discovery with discretion to distinguish between the incomplete knowledge domain graph and an incomplete knowledge domain. In this case, knowledge gaps are filled as the knowledge domain graph matures.

In the web mining domain (Figure 14), social networks to be crawled has been left out as a knowledge gap, and it can be easily identified as a goal as a leaf node. This gap in knowledge was found in Liu (2007), where he mentioned more research is needed on this topic.
5.4.2 Solution effects evaluation

Another useful aspect of the GOKM is its ability to visualize the solution’s evaluation. Soft-goals, as we have previously defined them, are the effects of the solution and deal with quality aspects of the solution. Evaluation of the solution is done through study of contribution links between the solution and soft-goal. Soft-goals can be greatly used in evaluating the different alternatives of the goal.

In the web mining domain (Figure 14), Universal crawler has been identified to Help ‘Low crawling overhead’, ‘Incremental index update’, and ‘Support universal search’. However, Universal crawler is biased toward popular web pages rather than authoritative sources, and hence it ‘Hurt’ Authoritative web page bias.

Taking another example, we can evaluate the alternative solutions of topic-specific crawling, focused crawler’ and ‘context-focused crawler’. Crawling updatability is Helped by both alternatives, whereas Crawling efficiency has Some – (negative) impact by Focused crawler. Finding a high yield crawling pathway is Hurt by Focused crawler.

5.4.3 Solution assumption evaluation

Knowledge user is also provided with the ability to evaluate the solutions based on the assumptions that the solution is based on. By evaluating and analysing the assumptions that the solution is based on, knowledge users can achieve more in-depth insight about the available solutions.

In the example of web mining domain (Figure 15), topic-specific crawling is taking the assumption that ‘text classifiers and seed pages are available for crawling’. It is important to note that the branch of alternatives under topic-specific crawling is sharing this assumption due
to the hierarchical structure of the GOKM modelling approach. Further,context-focused crawler is taking the assumption that reverse links for a fraction of the seed set documents exists at a known search engine.

**Figure 15: Knowledge domain assumption evaluation**

### 5.5 GOKM process documentation use case

Use cases describe user-system interactions. They focus on the functional interactions between a role (AKA actor) and a system in order to achieve a goal. The actor can be a human or an external system. The use case diagram (Figure 16) is designed in order to summarize the processes described earlier in this chapter, and identifies the users, their interaction when creating, and using GOKM.
As illustrated in the use case diagram, the main users of the GOKM proposed are researchers and information users. Researchers or domain experts play a significant role in building the GOKM model for the problem-solution based knowledge domains. They are responsible for codifying and identifying the knowledge components of the publications including goal(s), recommended solution(s), assumptions that the solution(s) is built on, as well as the soft goals that are affected by the solution. They are also responsible for building the GOKM model for every article separately in order to show the relationship between the knowledge components within the publication. Furthermore, merging process begins where the publication model is merged into the with the domain model, which represents the available literature within the domain. The merged model will allow researchers to overview and visualize the available knowledge with the domain, thereby giving them insight for their future research.

Researchers as well as other information users will benefit from the GOKM model while searching for information. They would be able to search by goals or solutions with respect to a problem-solution based knowledge domain resulting in retrieving more accurate results as opposed to keyword-based search. The information user who may be a newcomer to a domain can also use the GOKM model as a guide to educate themselves concerning the domain. The GOKM facilitates the learning process as it largely focuses on the main knowledge components within the domain. However, this does not imply that the model can replace the publications.
Figure 16: GOKM process documentation use cases
6 GOKM prototype

In this chapter, we present the prototype that was built to facilitate the GOKM. The prototype was developed in response to the lack of development in information systems that support GOKM. Observing the gap between the theoretical aspects of GOKM and its technological implications, we have decided to come up with a proof-of-concept prototype.

Palm et al. (2013) noted that the prototyping is a way of defining a milestone where the realization of a subset of system aspects is the main goal. Prototypes can take different forms and configurations depending on the demands of the projects. Palm et al. (2013) also pointed out that prototyping is most useful if the proof-of-concept is missing and the system requirements are unclear or unstable. At this point of research, the requirements for the system that supports GOKM are still unclear and the system needs to be conceptually proven more than anything else.

Following this rationale, we develop a prototype with the primary goal of exploring and overcoming the issues related to the implementation of GOKM approach. With the help of this prototype, we are able to envision a fully implemented system that supports and promotes GOKM principles and procedures.

This prototype was developed using WAMP (Windows, Apache, MySQL, PHP) stack. This web application development bundle is widely used in the industry, and provides the flexibility and functionality needed for this prototype. Further, we tested the prototype using the data from web mining domain to ensure that the prototype meets the expected functional requirements.
6.1 Publication data entry

![Image of.GOKM prototype - Publication data entry](image)

The first step of GOKM prototype is adding the publication information to the system. The users are asked to enter the publication basic information that will be used throughout the knowledge codification process. The publications are added to a knowledge domain. This procedure creates a library of articles that represent the knowledge domain. When the user submits the publication by clicking on Add Publication, they receive a confirmation page acknowledges that the publication has been added to the database.

- At the early stages of prototyping, we examined integration of GOKM prototype with the online digital libraries like Mendeley (Mendeley.com). Integration of GOKM with Mendeley through their API gives GOKM the advantage of utilizing Mendeley’s publication collection management. The integration also brings more collaboration.
between the GOKM users with the help of the collaborative nature of Mendeley networks.

6.2 Codification form

The main process of GOKM prototype is focused on codification of a publication. In this phase, the user enters the knowledge components codified from the publication. At the top, users can select the knowledge domain, and article, which they entered in the previous stage. Figure 18 shows an example of how a codified knowledge component of one publication can be entered into the system. Goals are first entered with respect to their hierarchy and relations. The two solutions that the publication introduced are entered with respect to the goal that they solve. Soft-
goals are also entered based on the contribution that they receive from the solution. Lastly, assumptions, if any, are attached to the solution. After submission, users can see a confirmation page, which demonstrates the knowledge components that have been added to the system.

6.2.1 One-to-many knowledge component relationships

This is when the knowledge component has relationships with more than one component (e.g. solution with multiple soft-goals). In order to address this, relations are captured with the help of knowledge component’s hierarchy (abstract) levels. The form is designed in a way that each component specifies its dependency to the knowledge component that it belongs to.

- For each of the knowledge element, user can add multiple entries as needed (as shown in Figure 18). The interface supports the add/remove functionality for each record.
- The form data are dynamically handled in the back end and prepared to be inserted into right database tables.

6.2.2 Data duplication

Handling the duplication is a major issue in data integrity. Knowledge components entered by the user should be checked and made sure that they are unique. For example, high-level goals can be easily duplicated if the user formats the goal differently, or user is entering a same solution, which was already codified in the system (Figure 19). In order to minimize data
duplication, records from knowledge domain are retrieved and presented to the user at the time of data entry to encourage use of existing knowledge components and avoid creating unnecessary duplications. This way, users can see what the database already holds, which firstly saves time on re-entering, and more importantly, minimizes data duplication. The use of different terminology for knowledge components is an issue that should be addressed by widening the range of suggestions given to the user.

**6.2.3 Dynamic form pre-populated**

As part of data duplication validation and user interface enhancement, we have designed a form, which is pre-populated with the knowledge domain information. On the form load, a database query is executed which fetches all the knowledge components of the domain, and populates the form’s input fields. These components are displayed to the user once they start typing, and they can choose from the options.

**6.2.4 Validations**

In GOKM prototype, one type of validation is on the empty field or partial entries, which are not allowed for data integrity. These are mandatory fields that cannot be left empty when the form is submitted. For example, in publication entry page, publication’s author(s) and title are mandatory fields, therefore user cannot submit without filling those fields. Similarly, in codification form, we mandate that the knowledge component's entry to be complete, and don’t accept partial form submission. The other type of validation takes the preventive approach by pre-populating the form with knowledge domain entries stored in database previously. We have described this process in the preceding section.
6.3 GOKM Search

We have also envisioned a search feature for the GOKM prototype. With the help of GOKM search, users are able to search for GOKM knowledge components, and the system generates the portion of the model that corresponds to the search criteria. The search results narrow down the domain map to show the user the most related knowledge components. The search can be narrowed down by selecting the type of knowledge component as part of the search criteria. It should be mentioned that as part of proof-of-concept prototype, the search functionality was not developed fully so that it generates the perfect model for the result. Figure 20 is a sample search result to validate the search functionality of the GOKM prototype. As part of search results, the article information in which the solution was found is also displayed.
6.4 GOKM help

In the help section of the prototype, we provide the complete set of documented GOKM processes and definitions. The help section is the part of the prototype where the users familiarize themselves with the GOKM concepts and processes. The process documentation are used from this thesis with modifications that makes it more readable on the web; however, it is worth noting that more user friendly and effective help strategy should be envisioned to adequately address the complexity of GOKM.
6.5 Lessons from GOKM prototype
Throughout the iterations of prototyping, we have learned many lessons. We have alluded to many of these lessons in this chapter so far. Here are some of the points to summarize this chapter.

In the process of creating this prototype, we have created and used a relational database to store the data. We opted to utilize the relational database since it is widely used and available freely. However, other types of databases, particularly object-relational databases might offer superior capabilities for this prototype. In an object-relational database, classes and inheritance are supported in database schemas and in the query language. We believe that the object-oriented nature of these databases can be advantageous in capturing the GOKM knowledge components and their complexity more effectively. Employing such databases can improve how the data is stored, as well as improving how the data is retrieved. The object-oriented database might be a better choice for retrieving data for generating the GOKM model.

In designing GOKM prototype, we face a trade-off between accessibility, and formality. For example, use of ontologies and web semantic techniques like OWL and RDF schemas is viable for such project. However, we believe that at this point of project, these approaches would equip know-how map with too much information that might make it harder for users to adopt GOKM.

Following the need for the different database management system, we also have learned that due to the complexity of the data, and exponential growth in size of database, there is a need for optimization task. We believe that the complexity of the data structure can diminish the performance of the system in the long run, and therefore, it is reasonable to invest time in optimization task.
7 Evaluating the GOKM approach

In this chapter, we evaluate the GOKM approach in three complementary ways. First, we apply the GOKM approach to the web mining domain as a case study in another domain to demonstrate the applicability of the approach. Second, we perform a feature analysis of the GOKM approach based on the characteristics of knowledge maps we identified in the literature reviews. Lastly, we experimentally evaluate the GOKM approach with respect to a textual literature review. We provide two different sources of information about web mining domain to the participants: GOKM map, and a written review. The performance in terms of correctness and time to answer the questions about web mining domain is then analysed.

7.1 Applying GOKM to the Web mining domain

In order to assess the applicability of GOKM, we need to extend the variety of the domains to apply the GOKM approach. Nasser (2012) applied know-how mapping on Geo-engineering domain, which is a technological domain, Gross et al. (2013) applied know-how mapping to the ranking algorithm domain. Nasser (2012) and Gross et al. (2013) took the approach map know-how of the domain, and used goal-oriented concepts and modelling techniques to create their maps. Nevertheless, more domains need to be explored by the GOKM approach in order to further evaluate the potential and applicability of the approach.

The domain we chose to apply GOKM in this work is web mining. As alluded to earlier, the GOKM approach is aimed at technological domains. The web mining domain is a technological domain that fits the characteristics of GOKM. Following the GOKM process documentations, we mapped the domain based on major publications. We adopt the top-down approach for mapping
the web mining domain. We start with Liu (2007), which is a comprehensive book on web-mining domain, and we used it as a starting point to gather other publications.

In mapping the web mining domain, we explored the domain based on the problem-solution structure, and identified the means-ends and decomposition relationships of problems and their solutions. The other knowledge components, such as soft-goals and assumptions were also analysed in the process of creating a know-how map of the domain.

The other product of GOKM application in web mining domain is the GOKM process documentations. These documentations are a reflection of the actual process of creating the domain know-how map. The documentation of the processes helped us to evaluate the applicability of the GOKM approach on the web mining domain.

7.1.1 Lessons learned:

Applying the GOKM concepts to map out the web mining domain we gained the following understandings. First, it requires one to have a good understanding of the available literature within the domain in order to be able to codify the domain and identify the relationship between different knowledge components discussed in various publications within the domain. Although this is not a mandatory requirement, we believe that mapping task would be unclear if map maker is unfamiliar with the context of knowledge domains. Second, in order to be able to create the know-how map, the map maker requires to have decent knowledge of the goal-oriented concepts to be able to extract and codify the GOKM components. Third, know-how map requires maintenance. That is, with know-how map, there is a need for a domain expert / map maker to review the knowledge map in order to ensure the correctness and accuracy of the map thereby increasing the quality of the know-how map. Fourth, updatability is an important requirement for
know-how map. Similar to other knowledge mapping techniques, in order to have a useful know-how map, the map needs to be updated regularly.

Even though the creation and maintenance of know-how map can be deemed as a time-consuming activity, we believe it adds great value to the domain. Know-how maps can provide a detailed overview of domains. That is, it allows one to visualize the available literature as well as the knowledge gaps. Additionally, based on our experience, codification and mapping helped us gain a new perspective about publications in terms of GOKM components (i.e. Goals, solutions, soft goals, etc.). This perspective facilitated us in learning and reviewing domains.

7.2 Feature-based analysis of GOKM

In this section of GOKM evaluation, we look at the GOKM approach in light of the desired characteristics of a knowledge map. We have identified these characteristics and needs earlier in the literature reviews. The knowledge map characteristics are coming from the literature review on characteristics and the needs of knowledge map in Chapter 2. Furthermore, we seek to express various ways that GOKM relates and addresses the knowledge map characteristics and needs.

Underlying theory:

The underlying theory, which refers to assimilation theory (Canas et al., 2003), is at the base of GOKM approach. In GOKM, problems-solutions are mapped with relation to the domain, and therefore, provide a map of which knowledge components relate to the neighbouring pieces of domain knowledge. The GOKM approach provides the means to map the problems and solutions, as well as the qualification criteria for the solution. This feature of GOKM, we believe,
is essential for effective learning and understanding of the domain. Mapping the knowledge domain and relating the pieces of knowledge together is also aligned with what Borner et al. (2003) identified as the primary goal of knowledge visualization, which is to paint the big picture of the domain. Moreover, GOKM maps are a means of communication between the map creators and users (Wexler, 2001). The efficiency of the domain learning and the ability of the map users to navigate the know-how map as a means of communication are the main objectives of the upcoming GOKM evaluation.

**Semi-hierarchical structure:**

The semi-hierarchical structure of the GOKM allows users to follow the direction of generalization to more detailed parts of the domain (Canas et al., 2003). The semi-hierarchical structure provides more readability for the map users, and enables the mapmakers to find the right places for their knowledge components. The evaluation of semi-hierarchical structure, which includes making and using of GOKM, requires more focused evaluation. This is due to the fact that hierarchy that the mapmaker has in mind, might be ambiguous to the map users; thus causing a discrepancy between creating and using of the know-how map.

**Shared syntax and semantics:**

The GOKM process documentation provided a syntax and semantics that can be shared between the map creators and users. Labelled links and definition of nodes are other characteristics of GOKM, which has been improved upon by the process documentation. In addition, the self-correcting action identified by Wexler (2001) was not included in the scope of GOKM evaluation in this research, and it is not yet clear how GOKM can address the issue.
**Purposeful communication medium:**

The knowledge maps are designed to facilitate the communication of between the mapmakers and its users. GOKM as a communication medium is designed to improve the breadth of knowledge transfer by engulfing the multiple aspects of the domain. GOKM is also designed to minimize the miscommunication by use of defined knowledge components and links, which can reduce errors in knowledge transfer.

As part of usability criteria of the knowledge map, Wexler (2001) pointed out that the map should be made for the purpose of addressing a problem. The use of the knowledge map depends on how well it addresses the needs of its creators and users. So far, we have made the case that technological domains are well suited for GOKM approach, and we have applied GOKM to different domains. This aspect of GOKM requires more application of GOKM in different domains to enable us with more generalized evaluation and conclusion.

**Information search and discovery:**

Knowledge maps should be used as a means of information search and discovery. The GOKM approach is a way of facilitating information search and discovery in technological domains. GOKM provides a map with an overview of the existing problems and solutions in a domain. The maps produced by GOKM approach can be utilized to search and discovery of various knowledge components (i.e. problems and solutions). Knowledge gap discovery is at the core of GOKM approach, where the map users can identify the gaps in problems and solutions of the domain by looking for a leaf node goal.

**Information pattern and data reduction:**

An effective knowledge map should aim to reduce the data by finding patterns or models in data, thus avoiding information cluttering (Grinstein & Fayyad, 2002). GOKM takes the approach to
reducing the information by summarizing the domain into knowledge components (i.e. goal, solution), relationships (i.e. means-ends and decomposition), and attributes (i.e. soft-goals and assumption). Data reduction is also important component of reducing miscommunication among knowledge map users as it reduces the cluttering information that complicates the knowledge transfer. However, this type of information summarization obscures the definitions of goals and solution for the map users; therefore, the know-how map should be used in junction with full text publication to maximize the efficiency of knowledge transfer and learnability.

Table 3 summarizes the relation of GOKM with the desired characteristics of knowledge map.

<table>
<thead>
<tr>
<th>Characteristics</th>
<th>GOKM</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Underlying theory</strong> [Borner et al. (2003), Canas et al. (2003)]</td>
<td>Problems and solutions, qualification criteria for solutions mapped with relation to the domain.</td>
</tr>
<tr>
<td><strong>Semi-hierarchical structure</strong> [Canas et al. (2003)]</td>
<td>Directional information flow from general to more details about knowledge domain. Provides high-level overview of the domain.</td>
</tr>
<tr>
<td><strong>Shared syntax and semantics</strong> [Canas et al. (2003)]</td>
<td>GOKM process documentation provides definitions and process of creating and utilizing the know-how map.</td>
</tr>
<tr>
<td><strong>Purposeful communication medium</strong> [Wexler, M. (2001)]</td>
<td>Facilitate knowledge transfer by reducing miscommunication.</td>
</tr>
<tr>
<td><strong>Information search and discovery</strong> [Wexler, M. (2001)]</td>
<td>Facilitate search for problems, solutions, and knowledge gaps in technological domains.</td>
</tr>
<tr>
<td><strong>Information pattern and data reduction</strong> [Grinstein &amp; Fayyad (2002)]</td>
<td>Avoids information cluttering by summarizing information and use of links for relations and attributes.</td>
</tr>
</tbody>
</table>

Table 3: GOKM characteristic summary
7.3 Controlled experiment

To be able to assess the effectiveness of using GOKM, we conducted a controlled experiment. In the following, we describe the objectives of the experiment, its design process and execution, as well as the results, their analysis and discussion.

7.3.1 Objectives

The main objective of the controlled experiment is to evaluate the effectiveness of understanding a knowledge domain by comparing it to a widely used technique – a written review. The written review is designed to mimic a short literature review of the domain. We compare the two approaches in terms of the level of understanding of the domain, and the relative time it takes the knowledge user to find answers to given questions. A by-product of the experiment is to check the extent to which one understands a know-how map generated by the others. We also analyse how the familiarity of the participants with goal-oriented concepts affects comprehension of the domain.

7.3.2 Hypotheses

The objectives of the controlled experiment are focused on the effectiveness of understanding a knowledge domain. First, we hypothesize that the GOKM approach is an effective way of knowledge transfer in terms of understanding the domain, and time efficiency. Further, another hypothesis is that the GOKM approach is able to capture and transfer knowledge, considering aspects of: finding problems, finding alternatives, evaluating alternative solutions, evaluating solutions’ effects, and finding knowledge gaps.

We also hypothesize that the GOKM approach and written review can convey the same amount of information to the participants; however, the differences between the two methods of
information delivery should result in observable effects in terms of domain comprehension and time efficiency.

We have also formed a hypothesis around participants’ familiarity with goal-oriented concepts and modelling approach. The hypothesis is that the participant’s performance over time and comprehension of the domain is positively affected, if they have a background in goal-orientation concepts and modelling.

7.3.3 Experiment design

In order to test the comprehension of a domain, we first design a written review and a know-how map. Next, we devise a questionnaire to examine the comprehension of the domain from various aspects, such as: finding problems, finding alternatives, evaluating alternative solutions, evaluating the solution’s effects, and finding knowledge gap.

In the first part of the experiment, background information of the participants is collected along with rating scales questions about their level of experience in different fields. The main body of the experiment is comprised of eleven open-ended questions about the web mining domain. The experiment ends with six reflection questions, which ask participants about their experience with retrieving information while doing the experiment.

Sreejesh et al. (2013) identify and discussed six steps in experiment design. We followed these steps in the design of our experiment as a guideline. These steps in experiment design are described in Appendix 5.

7.3.3.1 Principle of Information equivalence

During the design of the experiment, a special attention was given to the principle of information equivalence and its consequences on the questions. We have designed an experiment that
provided two different types of material to the participant to answer the questions. As already mentioned, two groups were provided with the GOKM maps of the web mining domain, and the other group was given the written review of the domain. This creates the implication that in order to create a fair and reliable experiment, all groups should receive the same information as well as questions that can be answered both ways. The information provided to all groups had to be carefully aligned together with iterative reviews, and matched each other in terms of amount of information, as well as the structure of the information. This is what we refer to as principle of information equivalence.

7.3.3.2 Correctness scoring system

In order to assess the correctness of the results submitted by the participants, we created a gold standard solution with the correct answer to each question (refer to Appendix 2). Solutions were created along with the creation of the questions to ensure that the answers are aligned with what the question intended to ask. Thus, the gold standard solution was used as a reference at the time of marking the questions to make sure all participants are evaluated based on a single point of reference, thus reducing subjectivity.

The scoring system used in the evaluation of the experiment weighted all questions equally. This approach is similar to the approach of Uren et al. (2006), in which all questions are given equal mark, which in our experiment was set to a range between 0 to 5 for each question, giving a maximum score of 55 in total. The scoring system allows us to assign partial marks to the questions based on the completeness and correctness of each question.
7.3.3.3 Independent and dependent variables

There are two categories of variables that need to be defined before going further. These two categories of variables are independent and dependent variables. These two categories of variables define the nature of relationships among occurrences (McArdle et al., 2006). For causal relationships, the manipulation of independent variable alters the dependent variable.

The primary independent variable identified for this research is the approach the participants take to answer the questions. These approaches are the GOKM and written review provided to the participants as the main source of information to answer the questions. The dependent variables of this experiment were set to the time it takes the participants to complete the questions and the quality of their answer.

We have also identified the participants’ familiarity with the goal-oriented concepts and modelling to be an independent variable that would affect their performance and time in the experiment. The familiarity of the participant with the goal-oriented concepts can change how the participant explore and understand the model and relationships. This factor applies only to the participants who took the GOKM material for their experimentation.

7.3.3.3.1 Factors affecting the relationships among variables

There are some factors that can potentially affect the relationships among variables (McArdle et al., 2006). An understanding of the dynamics of these factors, and addressing them in design, execution and analysis of experiment are crucial parts of any research.

Experimental testing effects are the effects on the participants, which often change their behaviour due to the changes in the environment (McArdle et al., 2006). The potential changes in the behaviour of the participants may raise questions about the generalization of results based on
the sample population. In order to limit the experimental testing effects, we have planned a comfortable experimental environment. We also believe that the participant experience in research and information searching works in advantage of reducing the experimental testing effects since this is not a new task for them.

Measurement errors during data collection can be categorized into technological error and recording error (McArdle et al., 2006). The technological errors are the errors that are inherited from the instability of the measuring devices or means. In this research, the technological errors are minimal since we are only using pen and paper as a means of collecting data, and it all of the participants is well familiar with these means. The recording error include the inaccuracies associated with observation and recording of the phenomena. This research is subject to potential recording errors since we intended to record the time for each of the participants. In order to address this challenge, we provided the large digital clock display in the experimental setting, visible to all of the participants.

The researcher’s expectation effect is the phenomena where the anticipation of a particular outcome of research may cause alternation in the actual results of the experiment (McArdle et al., 2006). The researcher’s neutrality must be maintained throughout the experiment design, execution and analysis in order to mitigate the expectation effects. During the design of the experiment, we reviewed the experiments and materials provided to the participants time after time to limit any type of disadvantage to any specific group of participants. Also during the experiment, different groups of participants are unaware of the fact that their approach is used to in comparison to the other approach, which we believe was an essential component of our neutrality to the participants and their answers.
7.3.4 Experiment execution

In this section, we first describe the participants of the study. Following that, we move on to the environment in which the experiment is carried out, the training, the experiment’s tasks, and threats to validity.

7.3.4.1 Experiment participants

We divided the participants into three groups: in the G0 group, we include participants who received a GOKM model, but were not familiar with goal-orientation concepts before participating in this experiment. In the G1 group, we include participants who received a GOKM model and were previously familiar with the goal-orientation concepts. Lastly, in group W, we include participants who were given a written review of the domain. The researcher chose the participants of G1 from the colleagues and former classmates who have the knowledge of goal-orientation concepts. The remainder of the participants were selected randomly (out of those who volunteered). The participants were motivated with a monetary compensation for completing the questionnaire. In Table 4, participants’ background information is presented. Table 5 contains the information about the experience of the participants with the goal-oriented concepts (for group G0 and G1) and experience with written reviews.

As shown in Table 5, 75% of G1 participants identified themselves to be very familiar with the web mining domain. On the other hand, none of the participants of G0 have any familiarity in this field, and 75% of W participants are somewhat familiar with web mining domain.
<table>
<thead>
<tr>
<th>Participant</th>
<th>Gender</th>
<th>Age</th>
<th>English Proficiency</th>
<th>Education Level-Field of Study</th>
<th>Scholarly Search</th>
<th>Problem-Solution Search</th>
</tr>
</thead>
<tbody>
<tr>
<td>G0 1</td>
<td>Male</td>
<td>30</td>
<td>Advanced</td>
<td>Master’s - business</td>
<td>Usually</td>
<td>Every day</td>
</tr>
<tr>
<td>G0 2</td>
<td>Male</td>
<td>23</td>
<td>Advanced</td>
<td>Master’s - mechanical engineering</td>
<td>Usually</td>
<td>Every day</td>
</tr>
<tr>
<td>G0 3</td>
<td>Female</td>
<td>19</td>
<td>Advanced</td>
<td>Bachelors - life science</td>
<td>Usually</td>
<td>Usually</td>
</tr>
<tr>
<td>G0 4</td>
<td>Male</td>
<td>29</td>
<td>Advanced</td>
<td>PhD - mechanical engineering</td>
<td>Every day</td>
<td>Usually</td>
</tr>
<tr>
<td>G1 1</td>
<td>Male</td>
<td>24</td>
<td>Advanced</td>
<td>Master’s – information technology</td>
<td>Usually</td>
<td>Every day</td>
</tr>
<tr>
<td>G1 2</td>
<td>Female</td>
<td>30</td>
<td>Advanced</td>
<td>Master’s - information technology</td>
<td>Usually</td>
<td>Usually</td>
</tr>
<tr>
<td>G1 3</td>
<td>Male</td>
<td>30</td>
<td>Advanced</td>
<td>Master’s - business administration</td>
<td>Usually</td>
<td>Usually</td>
</tr>
<tr>
<td>G1 4</td>
<td>Female</td>
<td>26</td>
<td>Advanced</td>
<td>Master’s - information technology</td>
<td>Usually</td>
<td>Occasionally</td>
</tr>
<tr>
<td>W 1</td>
<td>Male</td>
<td>32</td>
<td>Advanced</td>
<td>Master’s - mechanical engineering</td>
<td>Occasionally</td>
<td>Sometimes</td>
</tr>
<tr>
<td>W 2</td>
<td>Female</td>
<td>26</td>
<td>Advanced</td>
<td>Master’s - information technology</td>
<td>Occasionally</td>
<td>Usually</td>
</tr>
<tr>
<td>W 3</td>
<td>Male</td>
<td>31</td>
<td>Advanced</td>
<td>Master’s - electrical engineering</td>
<td>Usually</td>
<td>Usually</td>
</tr>
<tr>
<td>W 4</td>
<td>Male</td>
<td>24</td>
<td>Advanced</td>
<td>Bachelors - electrical engineering</td>
<td>Occasionally</td>
<td>Usually</td>
</tr>
</tbody>
</table>

Table 4: Participant demographic information

<table>
<thead>
<tr>
<th>Participant</th>
<th>Goal Oriented Modelling Concepts</th>
<th>Literature Reviews</th>
<th>Knowledge Mapping</th>
<th>Web Mining Domain</th>
<th>Exploring knowledge domains</th>
</tr>
</thead>
<tbody>
<tr>
<td>G0 1</td>
<td>Unfamiliar</td>
<td>N/A</td>
<td>Somewhat familiar</td>
<td>Unfamiliar</td>
<td>Very familiar</td>
</tr>
<tr>
<td>G0 2</td>
<td>Unfamiliar</td>
<td>N/A</td>
<td>Unfamiliar</td>
<td>Unfamiliar</td>
<td>Very familiar</td>
</tr>
<tr>
<td>G0 3</td>
<td>Unfamiliar</td>
<td>N/A</td>
<td>Unfamiliar</td>
<td>Unfamiliar</td>
<td>Very familiar</td>
</tr>
<tr>
<td>G0 4</td>
<td>Unfamiliar</td>
<td>N/A</td>
<td>Unfamiliar</td>
<td>Unfamiliar</td>
<td>Somewhat familiar</td>
</tr>
<tr>
<td>G1 1</td>
<td>Very familiar</td>
<td>N/A</td>
<td>Somewhat familiar</td>
<td>Very familiar</td>
<td>Expert</td>
</tr>
<tr>
<td>G1 2</td>
<td>Very familiar</td>
<td>N/A</td>
<td>Unfamiliar</td>
<td>Unfamiliar</td>
<td>Somewhat familiar</td>
</tr>
<tr>
<td>G1 3</td>
<td>Very familiar</td>
<td>N/A</td>
<td>Very familiar</td>
<td>Very familiar</td>
<td>Expert</td>
</tr>
<tr>
<td>G1 4</td>
<td>Very familiar</td>
<td>N/A</td>
<td>Very familiar</td>
<td>Very familiar</td>
<td>Somewhat familiar</td>
</tr>
<tr>
<td>W 1</td>
<td>N/A</td>
<td>Unfamiliar</td>
<td>Unfamiliar</td>
<td>Unfamiliar</td>
<td>Unfamiliar</td>
</tr>
<tr>
<td>W 2</td>
<td>N/A</td>
<td>Very familiar</td>
<td>Very familiar</td>
<td>Somewhat familiar</td>
<td>Very familiar</td>
</tr>
<tr>
<td>W 3</td>
<td>N/A</td>
<td>Very familiar</td>
<td>Somewhat familiar</td>
<td>Somewhat familiar</td>
<td>Very familiar</td>
</tr>
<tr>
<td>W 4</td>
<td>N/A</td>
<td>Somewhat familiar</td>
<td>Unfamiliar</td>
<td>Somewhat familiar</td>
<td>Somewhat familiar</td>
</tr>
</tbody>
</table>

Table 5: Participant’s research experience information
7.3.4.2 Experiment environment

The experimental environment was designed to be the same for all three groups in order to mitigate the environmental variables. The participants filled out their questionnaire independently, while working with their assigned groups. Conducting the experiment in groups was important since this way we are confident that the participant received the same amount of introductory information, and their environmental effects are controlled and equal.

In order to enable the participants to record the time spent for each question, we provided them with a large digital clock on a laptop screen that was visible to all of the participants.

7.3.4.3 Training

As Wexler (2001) pointed out, knowledge map makers and users must share the same set of symbols and definitions in order to utilize the knowledge maps. The participants of G0 and G1 were given a demo on a simple GOKM graph to familiarize them with basic concepts of GOKM. This demo is required for both groups since even though G1 participants had knowledge goal-orientation concepts, the GOKM components and approach was new to them. The demo focused on demonstrating the problem-solution aspect of GOKM and explained the means-ends and decomposition relationships among the components. The demo domain is described in Appendix 4.

7.3.4.4 Experiment task

The primary task of participants in this experiment is to explore and navigate the domain material in order to answer questions about the domain. Participants have to refer to the domain material (either GOKM or written review) to answer the questions. The experiment instructions and the questions appear in the Appendix 1.
While answering the questions, participants also have the task of recording the start and end time for each of the questions. This task involves looking up the time provided to them as part of the experimental setting, and writing it down before they start looking for the answers, and when they are done with each question.

7.3.4.6 Threats to validity

Asking the wrong question is the main source of threats to validity (Kirk & Miller, 1986). The threat of asking the wrong question was one of the key issues that challenged us in the process of questionnaire design. The clarity of the questions was a primary concern for this experiment, and we have taken extensive measures to ensure that the right questions are asked with clarity and simplicity. Appendix 5 provides some of the steps we took in order to ensure the quality of the questions.

Learning effect is likewise an issue during the experiment design. Learning effect is the process that affects the participants’ ability to answer the questions. Learning effect can influence how the participants understand the questions as well as how they look for the answers. We believe that the learning effect is not a threat to this experiment. This is because as also pointed out by Mosheiov (2001), in realistic settings, learning effect comes as a result of repeating the same or similar activities. In the realistic setting, we expect that users of GOKM benefit from repeating the similar task in exploring the knowledge domain, and therefore, improve with time.

The participant's background in the web mining domain can potentially be a threat to the validity of the overall results. This can skew the results in favour of those with the previous knowledge of the domain. However, we believe that this threat is minimized in this experiment because of the following.
First, the focus of the experiment is on the understanding of knowledge components and links in GOKM, and not the knowledge domain itself. We expect to see differences in performance across groups in different aspects of GOKM evaluation. The threat to validity that is caused due to the participant’s background in web mining should manifest itself with recurring results across all evaluation aspects. In our experiment, since the number of participants is small, we can expect the average score to be greatly effected if a participant completely miss a question. In detailed results description, we explain these cases based on what the actual answers given by the participant. Second, web mining is a large domain with many specializations; therefore, background knowledge of the participants is not enough to skew the results. Third, the questions are specific to the material provided to the participants; thus, we assume that participants could not rely on their memory to answer the questions.

This threat could have been neutralized further if the experiment included a question about the familiarity of the participant with the material provided to them. However, based on the justifications above, we believe participant’s background knowledge of the domain has minimal threat to the validity of the results.

Furthermore, as the number of the sample population increase, we have greater changes of receiving valid data. Although the number of participants was more than experiments of a similar kind (i.e., Uren et al. (2006) and Nasser (2012)), it is still not representative of the general population, and therefore, we have to refrain from making generalized conclusions.

**7.3.4.7 Scoring bias analysis**

Even after the use of the gold standard solution to mark the questions, there remains the bias of subjectivity of the researcher in marking each question. In order to address this subjectivity, we
used the second marker, who marked the questionnaire independently based on the gold standard solutions. Further, we analysed the correlation of the two marking results based on the Pearson correlation coefficient.

This approach provided us the confidence in the marking approach, which reduced the subjectivity of one marker. Based on the second set of scores, we concluded that the correlation coefficient for all questions is positive, which indicate that the two marker scores are in strong agreement with each other. Basing on this finding, we chose the original marking scores of the researcher as the primary data for this research.
7.4 Overall Results and analysis

In this research, we examine how the users of GOKM maps can perform when asked questions about the knowledge domain. In the experiment, we used factual questions about the domain that allowed us to investigate the overall quality of participant’s understanding of the GOKM approach and how they answered each question. Similar to the evaluation approach of Uren et al. (2006), we produced a written review of the knowledge domain, which provided a point of comparison for the GOKM approach. As Uren et al. (2006) pointed out, it is impossible to create the substance of the two artefacts (GOKM graph and written review) wholly identical, which has led them to be cautious about the general conclusions they draw from the experiment. This level of caution is also shared in the conclusions of this experiment. In the following, we demonstrate the overall results of the experiment in Table 6.

To further analyse the results, we look at the various aspects we mentioned before, namely: finding problems, finding alternatives, evaluating alternative solutions, evaluating solutions’ effects, and finding knowledge gaps. The questions are categorized into these aspects to evaluate GOKM approach. Table 7 presents the results for both comprehension and time divided into groups and aspects, whereas Figure 22 and Figure 23 visualize these results. In the following, we discuss the various aspects. For each one we provide the related questions and discuss the results.
### Table 6: Overall results per question per group

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Q 1</td>
<td>100%</td>
<td>0%</td>
<td>0:01:45</td>
<td>0:00:33</td>
<td>75%</td>
<td>43%</td>
<td>0:01:13</td>
<td>0:00:28</td>
<td>93%</td>
<td>13%</td>
<td>0:01:36</td>
<td>0:00:51</td>
</tr>
<tr>
<td>Q 2</td>
<td>100%</td>
<td>0%</td>
<td>0:01:19</td>
<td>0:00:07</td>
<td>100%</td>
<td>0%</td>
<td>0:00:54</td>
<td>0:00:08</td>
<td>100%</td>
<td>0%</td>
<td>0:02:07</td>
<td>0:00:57</td>
</tr>
<tr>
<td>Q 3</td>
<td>50%</td>
<td>50%</td>
<td>0:01:37</td>
<td>0:00:25</td>
<td>100%</td>
<td>0%</td>
<td>0:00:48</td>
<td>0:00:20</td>
<td>100%</td>
<td>0%</td>
<td>0:01:22</td>
<td>0:00:36</td>
</tr>
<tr>
<td>Q 4</td>
<td>75%</td>
<td>43%</td>
<td>0:01:19</td>
<td>0:00:15</td>
<td>100%</td>
<td>0%</td>
<td>0:00:48</td>
<td>0:00:06</td>
<td>100%</td>
<td>0%</td>
<td>0:01:56</td>
<td>0:00:44</td>
</tr>
<tr>
<td>Q 5</td>
<td>100%</td>
<td>0%</td>
<td>0:00:52</td>
<td>0:00:28</td>
<td>50%</td>
<td>50%</td>
<td>0:00:56</td>
<td>0:00:19</td>
<td>50%</td>
<td>50%</td>
<td>0:02:03</td>
<td>0:00:56</td>
</tr>
<tr>
<td>Q 6</td>
<td>100%</td>
<td>0%</td>
<td>0:00:51</td>
<td>0:00:17</td>
<td>100%</td>
<td>0%</td>
<td>0:00:56</td>
<td>0:00:09</td>
<td>100%</td>
<td>0%</td>
<td>0:01:03</td>
<td>0:00:44</td>
</tr>
<tr>
<td>Q 7</td>
<td>100%</td>
<td>0%</td>
<td>0:01:12</td>
<td>0:00:16</td>
<td>100%</td>
<td>0%</td>
<td>0:01:28</td>
<td>0:00:23</td>
<td>30%</td>
<td>17%</td>
<td>0:01:47</td>
<td>0:01:06</td>
</tr>
<tr>
<td>Q 8</td>
<td>95%</td>
<td>9%</td>
<td>0:02:41</td>
<td>0:01:45</td>
<td>74%</td>
<td>22%</td>
<td>0:02:11</td>
<td>0:00:41</td>
<td>43%</td>
<td>44%</td>
<td>0:02:10</td>
<td>0:00:41</td>
</tr>
<tr>
<td>Q 9</td>
<td>55%</td>
<td>30%</td>
<td>0:02:56</td>
<td>0:01:36</td>
<td>65%</td>
<td>36%</td>
<td>0:02:46</td>
<td>0:01:32</td>
<td>35%</td>
<td>38%</td>
<td>0:01:37</td>
<td>0:00:41</td>
</tr>
<tr>
<td>Q 10</td>
<td>25%</td>
<td>43%</td>
<td>0:04:49</td>
<td>0:03:32</td>
<td>75%</td>
<td>43%</td>
<td>0:01:44</td>
<td>0:00:51</td>
<td>25%</td>
<td>43%</td>
<td>0:02:04</td>
<td>0:01:05</td>
</tr>
<tr>
<td>Average</td>
<td>73%</td>
<td>5%</td>
<td>0:24:42</td>
<td>0:06:05</td>
<td>79%</td>
<td>7%</td>
<td>0:18:58</td>
<td>0:03:25</td>
<td>67%</td>
<td>13%</td>
<td>0:30:52</td>
<td>0:07:26</td>
</tr>
</tbody>
</table>

### Table 7: Overall feature results

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Finding Alternatives</td>
<td>100%</td>
<td>0%</td>
<td>0:01:32</td>
<td>0:00:27</td>
<td>88%</td>
<td>33%</td>
<td>0:01:06</td>
<td>0:00:25</td>
<td>96%</td>
<td>10%</td>
<td>0:01:51</td>
<td>0:00:56</td>
</tr>
<tr>
<td>Finding Problems</td>
<td>63%</td>
<td>48%</td>
<td>0:01:28</td>
<td>0:00:23</td>
<td>100%</td>
<td>0%</td>
<td>0:00:48</td>
<td>0:00:15</td>
<td>100%</td>
<td>0%</td>
<td>0:01:39</td>
<td>0:00:44</td>
</tr>
<tr>
<td>Evaluating Alternative Solutions</td>
<td>63%</td>
<td>43%</td>
<td>0:02:23</td>
<td>0:01:36</td>
<td>53%</td>
<td>43%</td>
<td>0:02:20</td>
<td>0:01:19</td>
<td>47%</td>
<td>44%</td>
<td>0:02:46</td>
<td>0:01:49</td>
</tr>
<tr>
<td>Evaluating Solution Effects</td>
<td>100%</td>
<td>0%</td>
<td>0:01:02</td>
<td>0:00:19</td>
<td>100%</td>
<td>0%</td>
<td>0:01:12</td>
<td>0:00:24</td>
<td>65%</td>
<td>37%</td>
<td>0:01:25</td>
<td>0:01:00</td>
</tr>
<tr>
<td>Finding Knowledge Gap</td>
<td>25%</td>
<td>43%</td>
<td>0:04:49</td>
<td>0:03:32</td>
<td>75%</td>
<td>43%</td>
<td>0:01:44</td>
<td>0:00:51</td>
<td>25%</td>
<td>43%</td>
<td>0:02:04</td>
<td>0:07:26</td>
</tr>
</tbody>
</table>
7.4.1 Domain comprehension and scores

The aim of the study was to examine how the GOKM approach performs in exploring and understanding a knowledge domain. As demonstrated in Figure 24, the results of this experiment point to the conclusion that GOKM performance is reliable when it comes to exploration and
understanding of the domain. Group W ranked slightly lower than G0 in the average correctness of the answers to the experiment with an average of 67%. Group G0 is second, with improvements with an average of 73%. Group G1, with an average mark of 79% is the top performer in our evaluation. Table 8 shows the overall score for each participant.

![Average performance score across the groups](image)

**Table 8: Total score**

<table>
<thead>
<tr>
<th></th>
<th>G0</th>
<th></th>
<th>G1</th>
<th></th>
<th>W</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total mark:</td>
<td>78%</td>
<td>75%</td>
<td>64%</td>
<td>75%</td>
<td>82%</td>
</tr>
<tr>
<td>Average:</td>
<td>73%</td>
<td></td>
<td>79%</td>
<td></td>
<td>67%</td>
</tr>
<tr>
<td>Std dev:</td>
<td>5%</td>
<td></td>
<td>7%</td>
<td></td>
<td>13%</td>
</tr>
</tbody>
</table>

The difference between the group G0 and G1 can be mainly due to their difference in experience with goal-orientation. The other possibility of this result could have been the participants' background knowledge in web mining domain, however we ruled out this as a threat to the validity of the results (see Section 7.3.4.6). Furthermore, it can be concluded that both groups of GOKM (group G0 and G1) approach performed relatively better than the written review
participant. Based on this observation, we can conclude that within this sample population, GOKM approach has the advantage, and can improve the performance of its users. In comparison, in the ScholOnto project experiment (Uren et al., 2006), participants who used the written review outperformed the group that used the Claim Network approach to knowledge mapping.

7.4.2 Time

One of the factors that we used in this research to determine the relative difficulty of the participant’s answer was the time it took them to find and answer the questions. The time measured in this process can be skewed by factors like how fast the participants can write down answers to questions, or how much writing was done for each question. For the first factor, we have no way of measuring the speed of handwriting for the participants. However, we have observed that almost the same amount of information was written down across all groups of participants for each question. This is because the answers to the questions were intended to be short and precise and to the point. With all these factors taken into account, we observed that the average amount of time that took the participants to complete the experiment varied dramatically.
As demonstrated in Table 9, it took an average of 30:52 minutes for participants of group W to complete the experiment. Participants of group G0 ranked second, with 20% improvement in average of 24:42 minutes. Lastly, it only took 18:58 minutes for participants of group G1 to complete the experiment; that is almost 40% improvement in time compared to participants of group W who were given the written review. The average time difference between group G0 and G1 can be related to their background and familiarity with goal-oriented models. Table 9 shows the time spent by the participants on answering the questions.

Both groups G0 and G1 did well compared to group W, which leads us to conclude that GOKM approach can save time in exploring and understanding the domain. This conclusion is supported.
by the difference in the average time of the two groups that took the GOKM approach to answer the questions. In their evaluation of ScholOnto approach, Uren et al. (2006) observed the opposite effect, where the participants using the ScholOnto claim-based knowledge mapping approach as their main source of information performed dramatically worse than the written review group. The familiarity of the group W with written review seems to be irrelevant since participants of group G0 were presented with the GOKM model for the first time.

7.5 Detailed experiment results

To further move into detailed results of the experiment, we focus on the results in terms of different aspects. Questions from the experiment are grouped together based on the type of information they ask for. These groups of questions have been utilised in the process of creating the user evaluation experiment in order to allow us to handle different aspects of know-how mapping.

In the following, for each set of questions, we identify the questions, and present the results gathered from the conducting the experiment. For each question, we scored the answers from participants based on the quality and completeness of the answers. We also provide a combined mark for each group of questions, which allow us to inspect the overall performance of knowledge feature per each group.
7.5.1 Finding Problems

The purpose of question 3 and 4 is to examine whether participants are able to correctly identify the problems. The participants are required to find problems based on the solutions given to them in the question.

- (Q 3) What is the problem that HITS and SALSA methods aim to solve?
- (Q 4) What is the problem that FeatureExtractor method aims to solve?

The majority of the participants within the GOKM group were able to identify the problems correctly. However, for Q 4, two participants in group G0 focused on the soft goals that are affected by the proposed solutions rather than the problems that the solutions aim to solve. The two participants made the same mistake while answering the Q 4. The foremost reason for this mistake can be that the participants found the questions vague, which we believe this is highly unlikely to be the case, as the majority of the participants were able to answer the questions correctly implying that they interpreted the question as well as the concept that the goal graph conveyed. The second reason might be that the difference between the two concepts: problems and soft-goals were not very clear to the participants. This is more likely to be the case, as this misunderstanding occurred with the participants who used the GOKM approach but are not familiar with the goal-orientation concepts.

On the other hand, participants of group W were all able to understand, and answer the two questions correctly. This implies that users are likely to identify the problems that solutions aim to solve via a well-written text that give them full exposure to the context.
As demonstrated in Figure 22, G1 participants scored full score in finding the problem with the lowest time. Participants of W were also able to achieve the full mark; however, their time almost doubles the time that G1 participants took to answer the questions. This is an indication that GOKM users can find problems based on the solutions.

### 7.5.2 Finding alternatives

The focus of these questions is on solution finding in the knowledge domain. Participants are required to find the alternative solutions proposed to the problem in the question.

- (Q 1) What are the alternatives for discovering hyperlinked web structure?
- (Q 2) What are the alternatives for text and web page pre-processing for data mining?

Both groups G0 and G1, with the exception of one participant were able to address the solution alternatives flawlessly for both questions, which entails the concept that the introduced approach expresses was well understood by the participants. One participant in G1 was not able to answer the question because he/she did not pay attention to the keyword of “discovering” in the first question. This seems like a trivial mistake, as the participant successfully answered the other question on finding the alternatives.

On the other hand, within group W, one of the participants was not able to identify all three of the alternatives to discovering hyperlinked web structure problem. He/she was able to identify only two of solution alternatives, but not all. The reason could be that participants either missed the solution within the text, or forgot to write it down, or did not understand that the third solution, was an alternative to the other two solutions.
Based on this result, as shown in Figure 22, we can conclude that GOKM approach performed almost as well as written review, which is in line with the overall claim that GOKM can support alternative finding. In terms of time to complete the task, once again the subjects that who got the GOKM map did it in less time than the subject who got the literature review.

### 7.5.3 Evaluating alternative solutions

In this set of questions, we aim to study the ability of the participants to evaluate the alternative solutions. These questions require the participant to identify the alternative solutions, and make a comparison between them.

- (Q 5) The solutions for the web template detection problem are batch web page processing and page level template detection. Are these solutions both required to overcome this problem.
- (Q 8) Provide a summary of the solution trade-offs for crawling the web for specific topics.
- (Q 9) Which of the alternatives to batch web template processing has the advantage? Please explain.
- (Q 10) What are the distinguishing effects of focused crawler and context-focused crawler?

The aim of question 5 is to examine if participants understand the existing means-ends relationship, and its meaning whether such relationship is provided explicitly or not. According to the results, half of group W, and half of group G1 were not able to answer the question correctly. All other participants were able to identify the existence of this relationship, and what
it communicates. It is likely that group W missed the existence of the means-ends relationship, as it is implicit within the provided written review. However, within group G1, the reason that two of participants were incapable of answering the question correctly can be that the participants did not pay attention to the means-ends relationship and considered it as a decomposition link, or that they did not understand the concept that means-ends relationship communicates correctly. Group G0, who were new to the goal-orientation concepts did best in this question. This could be as a result that group G0 paid more attention to the description of the relations between GOKM components, and group G1 simply missed out to refresh their memory.

In question eight, we are examining the ability of the participants to identify the trade-offs of solution alternatives. By trade-off, we expected to get a list of all effects (soft-goals) that differentiate the alternative. The answer to this question was expected to be relatively easy compared to the other questions in this section, but as it turns out, this question was the lowest performance point for participants of group G0 and G1, while group W performed relatively well. None of the participants in group G0 could answer this question correctly, while only one participant in group G1 answered the question correctly. After reading and answers of group G0 and G1 more carefully, we found out that there is an occurrence of common mistake in their answers. In the GOKM graph provided to the group G0 and G1, the solution of topic-specific crawler is expanded with task-goal specification of web crawling to be topic-specific with two alternatives. Most of the participants in group G0 and G1 provided the trade-offs of the task-goal specification alternative, where the intention of the question was to identify the trade-off between topic-specific crawler and topical crawler. Although participants of group G0 and G1 struggled
with the question, participants of group W seem to have understood the intention of the question with no problem, and provided adequate answer for it.

We believe that such a gap in performance might be due to the way that the information was provided to the participants. First, group G0 and G1 may have been confused about the hierarchy of the goals and solutions, which led them to look for the trade-off in the task-goal specification. Another factor resulting this might be due to the similarity of keywords used in the task-goal specification, which can result in poor judgment of question’s intended meaning. On the other hand, group W participants maximize their privilege of having written review by following the flow of the text, and performed better in the answers.

With question 9, we aim to examine how participants evaluate the alternative solutions of a problem based on their quality effects, or soft-goals. Although all three groups answered the question within the same average time, but the quality of the answers varies in different ways than most of the other questions in the experiment. Group G0, with an average of 95% ranks the highest average for this question, with group G1 following it on average of 74% and group W scoring the lowest average of 43%. Group G0, although unfamiliar with goal-orientation concepts, seems to grasp the value of alternative analysis more than any other groups. Two participants of group G1 miss to provide any reasoning and justification for their answer.

In group W, one participant who achieved the full mark for this question, wrote down everything related to batch template processing in order to leave out any room for error. The rest of the group W seems to fall short in writing down, and missed some points, or missed the question by evaluating alternatives of a different problem.
In question 10, similar to previous questions, the goal is to evaluate participants’ understanding of solution alternative analysis. Although on average, it took participants of groups W one minute less than group G0 and G1 to answer the question, their quality of answer falls short at 35%. Group G0 participants ranked second, while the participants of group G1 answered this question most accurately with an average of 65%. In this question, we expected the participants to provide a complete set of the distinguishing factors of the alternatives. Only 50% of the participants in group G1 answered the question completely, whereas this only 25% of group G0 and G1 achieved the complete answer. One of the reasons for the low average mark on this question might be related to the way the question is phrased. Although all participants were instructed and encouraged for the full answer to each of the questions, this question is phrased in a way that led the participant to provide the short form of answer. The short form of the answers, which 50% of group W and 25% of group G0 and G1 provided was focused on only one attribute that they thought is the “distinguishing effect” between the solution alternatives. Similar to previous questions, some participants of group G0 and G1 struggled with contribution links and their effects. We believe that this relates to visibility (i.e. “Some +/-” links lacked enough visual distinction) of the contribution links in the printouts of the model provided to the participants.

As demonstrated in Figure 22 G0 participants achieved the highest average for the alternative evaluation feature. Furthermore, G1 participants are also affected by their low performance in question 8; however, they too have an advantage over the W participants. The average time for all groups is in the same range (Figure 23). This result can be interpreted in support of GOKM as an effective way of evaluating the alternative solutions in a knowledge domain.
7.5.4 Evaluating solutions’ effects

Question 6 and 7 aims to test participants' understanding of soft goals, and contribution links. The participants are required to find the solution in the material first, and further describe the effects of that solution.

- (Q 6) How does Site Style Trees (SST) affect the noise data in the web page?
- (Q 7) Describe the effects of ‘stopword’ removal.

All of the participants of group G0, G1 and W were able to answer Q 6 correctly. This implies that within the group G0 and G1, participants were able to read the contribution link, and understand how it serves. On the other hand, within group W, it can be concluded that as long as the participants are able to find the solution, and as long as the text is clear enough, participants are able to understand solutions’ effects.

With regards to the question 7, all participants within the group G0 and G1 were able to find the named solution, and via studying the soft goals and the contribution links were able to provide accurate answers which validates that they understood the solution, soft-goal, and contribution concepts well, and are capable of navigating the model. Then again, none of the participants of group W were able to resolve this question completely. The reason for this can be either the participants found it hard to find the solution within the text, or not all of the soft goals were clear to them.
The GOKM approach clearly has the advantage in evaluating the solution's effects both in terms of comprehension and in time (Figure 22 and Figure 23). The participants of G0 and G1 were able to achieve the full score in both questions, which indicate that GOKM approach is an effective way of evaluating the solutions.

7.5.5 Finding knowledge gap

In this question, we aimed to validate participants’ understanding of a knowledge gap, and examine their ability to identify a knowledge gap within the provided material. This question is the most open-ended question in the experiment, and perhaps it is most fundamental one as well. We did not expect that the expertise of participants in web mining domain to influence this question, since the domain is so broad and their short exposure to the domain is not sufficient to reveal insights on the domain knowledge gaps.

Originally, we have defined knowledge gaps in GOKM as problems or goals that do not have any solutions proposed to them and therefor there is a need for improvement. The knowledge gap is distinguishable in the domain map as a goal leaf node that has no solutions attached to it. In order to make this a fair comparison of all three groups, during the experiment, participants were instructed to think in problem-solution context, and see if in that context they can identify a gap in knowledge. We believe that this way of interacting was fair since participants had no way of identifying missing goals (due to their lack of expertise in the domain), and that should let them think in terms of missing solutions. In the written review provided to group W, we provided them with a sentence stating “The rise of social networks in recent years has opened up a new field for web crawlers to improve and innovate.” We believe that this should have been fair way of pointing out the knowledge gap in written review, as statements like this are common in future works of many publications.
(Q 11) In the provided material, can you identify any gaps in the knowledge in this domain? Explain how you are able to identify these gaps (if any).

75% of participants in group G1, and 25% of those of group G0 were able to identify the knowledge gap from the provided model. Within the group W, only 25% of participants were able to identify the knowledge gap.

In this question, we were only considering problems that no solution has been proposed to them as knowledge gap; however, we came across interesting and correct answers. There were a couple of participants who identified different type of knowledge gaps. One participant in group G1, in addition to identifying the correct answer of social networks to be discovered, mentioned that solutions without relating soft-goals are also considered as a knowledge gap. Although all of solution nodes in the GOKM model had soft-goals, three solutions which were used as intermediary solutions did not have any soft-goals. This is an important insight into how the readers of GOKM think about knowledge gaps. The participants identify the lack of soft-goals as a gap in knowledge of understanding the solutions and their effects. One participant of group G0 believes that goals (problems) with only one solution should be considered as a knowledge gap. This is also valuable insight since it indicates that the participant was able to look for places inside the GOKM graph, and hypothesize about where the potential knowledge gaps may reside.

To conclude, we can see that the G1 participants had the advantage in terms of correctness of the answer as well as in time. The other groups share the same number of correct answers, although the time for W participants is half of G0 participants. Although G1 participants show great advantage other the other participants, this result is not sufficient to make generalized
conclusions about the knowledge gap discovery of GOKM. This is because G0 participants show no advantage compare to W, and it is only G0 that has the advantage.

7.5.6 Reflection questions
To better understand the participants' experience regarding the use of two approaches (GOKM and written review) we asked them to answer a set of questions (as part of their task) as shown in Table 10
<table>
<thead>
<tr>
<th>R 1</th>
<th>The taken approach helps a newcomer to a domain to learn a domain in a short amount of time</th>
</tr>
</thead>
<tbody>
<tr>
<td>R 2</td>
<td>The taken approach facilitates a newcomer to learn a domain comprehensively</td>
</tr>
<tr>
<td>R 3</td>
<td>The approach enables researchers to identify knowledge gaps in a short amount of time</td>
</tr>
<tr>
<td>R 4</td>
<td>The approach enables researchers to easily identify knowledge gaps</td>
</tr>
<tr>
<td>R 5</td>
<td>Please provide a scenario where such an approach would be useful and beneficial for understanding and exploration of a knowledge domain.</td>
</tr>
<tr>
<td>R 6</td>
<td>Please suggest areas of improvements or shortcomings of the approach.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Table 10: Reflection questions</th>
</tr>
</thead>
</table>

<table>
<thead>
<tr>
<th></th>
<th>P 1</th>
<th>P 2</th>
<th>P 3</th>
<th>P 4</th>
<th>P 1</th>
<th>P 2</th>
<th>P 3</th>
<th>P 4</th>
<th>P 1</th>
<th>P 2</th>
<th>P 3</th>
<th>P 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>R 1</td>
<td>Strongly Agree</td>
<td>Unsure</td>
<td>Disagree</td>
<td>Agree</td>
<td>Agree</td>
<td>Agree</td>
<td>Agree</td>
<td>Agree</td>
<td>Disagree</td>
<td>Disagree</td>
<td>Unsure</td>
<td>Agree</td>
</tr>
<tr>
<td>R 2</td>
<td>Strongly Agree</td>
<td>Agree</td>
<td>Unsure</td>
<td>Strongly Disagree</td>
<td>Unsure</td>
<td>Disagree</td>
<td>Unsure</td>
<td>Unsure</td>
<td>Strongly Disagree</td>
<td>Unsure</td>
<td>Agree</td>
<td>Strongly Agree</td>
</tr>
<tr>
<td>R 3</td>
<td>Strongly Agree</td>
<td>Strongly Agree</td>
<td>Disagree</td>
<td>Strongly Disagree</td>
<td>Strongly Agree</td>
<td>Agree</td>
<td>Strongly Agree</td>
<td>Agree</td>
<td>Unsure</td>
<td>Disagree</td>
<td>Disagree</td>
<td>Agree</td>
</tr>
<tr>
<td>R 4</td>
<td>Strongly Agree</td>
<td>Strongly Agree</td>
<td>Disagree</td>
<td>Disagree</td>
<td>Strongly Agree</td>
<td>Agree</td>
<td>Agree</td>
<td>Agree</td>
<td>Unsure</td>
<td>Disagree</td>
<td>Disagree</td>
<td>Strongly Agree</td>
</tr>
</tbody>
</table>

| Table 11: Participants answer to reflection questions |

<table>
<thead>
<tr>
<th>P 1</th>
<th>P 2</th>
<th>P 3</th>
<th>P 4</th>
<th>P 1</th>
<th>P 2</th>
<th>P 3</th>
<th>P 4</th>
<th>P 1</th>
<th>P 2</th>
<th>P 3</th>
<th>P 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>R 1</td>
<td>Strongly Agree</td>
<td>Unsure</td>
<td>Disagree</td>
<td>Agree</td>
<td>Agree</td>
<td>Agree</td>
<td>Agree</td>
<td>Disagree</td>
<td>Disagree</td>
<td>Unsure</td>
<td>Agree</td>
</tr>
<tr>
<td>R 2</td>
<td>Strongly Agree</td>
<td>Agree</td>
<td>Unsure</td>
<td>Strongly Disagree</td>
<td>Unsure</td>
<td>Disagree</td>
<td>Unsure</td>
<td>Unsure</td>
<td>Strongly Disagree</td>
<td>Unsure</td>
<td>Agree</td>
</tr>
<tr>
<td>R 3</td>
<td>Strongly Agree</td>
<td>Strongly Agree</td>
<td>Disagree</td>
<td>Strongly Disagree</td>
<td>Strongly Agree</td>
<td>Agree</td>
<td>Strongly Agree</td>
<td>Agree</td>
<td>Unsure</td>
<td>Disagree</td>
<td>Disagree</td>
</tr>
<tr>
<td>R 4</td>
<td>Strongly Agree</td>
<td>Strongly Agree</td>
<td>Disagree</td>
<td>Disagree</td>
<td>Strongly Agree</td>
<td>Agree</td>
<td>Agree</td>
<td>Agree</td>
<td>Unsure</td>
<td>Disagree</td>
<td>Disagree</td>
</tr>
</tbody>
</table>
In the first two questions, we ask the participants about their learning experience as a newcomer to the domain. Our emphasis is on the time and comprehensiveness of the approach they took when learning about a new domain. As shown in Table 11, six out of eight participants in G0 and G1 felt comfortable in learning a new domain in a timely manner; however, group W showed a limited support for timely learnability of the domain. As for the comprehension aspect (R2), most of the participants across all groups expressed concern about the comprehensiveness of the approach they took to learn the domain. This can be understandable since GOKM lacks the context and has a summarized format, and therefore users may not feel they have comprehensively learned the domain. The same argument may hold for the participants of the written review group W, since the written review provided to them was very short, and lacked many details.

Furthermore, R3 and R4 focused on the experience of the participants in finding the knowledge gaps in the knowledge base. Although only 75% of the G1 participants found the knowledge gap correctly, they all seem to have an unwavering conviction that GOKM can be helpful in defining the knowledge gap in terms of both time and effort. G0 participants are divided into two groups, 50% of them positively assess this issue, and the other half fell dissatisfactory about it. In the case of the literature review group W participants, the only one who provided a positive feedback is the one who answered the questions on knowledge gap correctly. The rest of the participants felt incapable of finding the knowledge gap in writing reviews. This might be due to their misunderstanding of what knowledge gap meant in the context of the experiment. The other reason might be that although they knew what the knowledge gap meant, they were lost in the written review text that affected their ability of knowledge gap discovery.
In R5, we ask an open-ended question for the suggestions about where the approach they took could be useful and beneficial. Table 12 illustrates all the results collected from the participants. One of the participants in the G0 group noted that similar concepts were employed in managerial and strategy course to evaluate the alternative solutions and trade-off. Another interesting possibility was brought up by participants of G0 group was the use of the GOKM modelling approach to find opportunities for starting a new business. One of the subjects within the e G1 group suggested that a similar approach could be used in transferring knowledge among organization’s members and teams. Other participants in G1 noted that a similar model could be utilized in creating a high-level knowledge about the field. One participant within the W group pointed out that such written review would be useful for preparing for exams, and compare the written review of the lecture notes.

Lastly, we ask the participants to suggest an area for the improvement. Virtually all of the G0 participants pointed out the amount of information in GOKM can be overwhelming as the model grows. Two of the participants in that group suggested that sub-categories or zones could support domain graph partition. One participant, who suggested zoning as a solution to reduce the complexity of the model, also suggested providing a list of all solutions and problems along with the zone number to assist the information searching. There have been also suggestions about increasing the readability of the model by improving on fonts and colour scheme used in the model.

In the G1 group, the general theme was about improving on the readability of the model. One participant subject pointed out that the contribution links were not readable and like the other participant suggested improvements in size and colour schemes of the model. There is also a proposal for categorizing the sections of the model, which was suggested by participants across
G0 and G1. The suggestions put forward by group W are vague or irrelevant to the main objectives of this research.
<table>
<thead>
<tr>
<th>Group G0</th>
<th>Participant 1</th>
<th><strong>R 5</strong></th>
<th>Participant 2</th>
<th><strong>R 6</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td><strong>Group G0</strong></td>
<td><strong>Participant 1</strong></td>
<td>Similar concept was applied in management and strategy courses to allow students to evaluate the alternative solutions by identifying the advantages and disadvantages, and identifying the gaps.</td>
<td>This approach can be a mess if we include many solutions/alternatives. If possible, there should be sub categories under each headline to help the reader better understand the concept without being lost in the process. Colour code and font can be improved.</td>
</tr>
<tr>
<td></td>
<td><strong>Participant 2</strong></td>
<td><strong>Identification of the business to start in a town focusing on finding the need for a new business.</strong></td>
<td>The amount of information could be overwhelming, some need the prejudice before making a decision, i.e. one might not need to know the alternative solution, simply provide the best option (similar to weighted metrics used to decide between projects).</td>
<td></td>
</tr>
<tr>
<td></td>
<td><strong>Participant 3</strong></td>
<td><strong>The model provided could be used for real life situations and goals such as planning for medical school, which courses are supposed to be taken, and what would be a trade-off to your GPA, etc.</strong></td>
<td>The information provided in this model was not necessarily easy to follow for a novice in the field, the time provided was not sufficient, and there could be a better description of the research. Participants should have been better informed on the purpose, as I felt lost during the study. The model was too complex.</td>
<td></td>
</tr>
<tr>
<td></td>
<td><strong>Participant 4</strong></td>
<td><strong>This would be very helpful for the design and execution in the engineering field</strong></td>
<td>The efficiency in search speed would improve if all the solutions and problems were separately listed and sorted and the researcher was provided with such a list. The model could be separated into zones. The zone number could be included in the list too.</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Group G1</th>
<th>Participant 3</th>
<th><strong>Marketing approaches to promote a business.</strong></th>
<th><strong>Participant 3</strong></th>
<th><strong>Marketing approaches to promote a business.</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td><strong>Participant 3</strong></td>
<td><strong>Marketing approaches to promote a business.</strong></td>
<td>I believe organizing and structuring the model to be easier to read and trace would help a lot.</td>
<td></td>
</tr>
<tr>
<td>Participant 4</td>
<td>It can be useful when you need to provide an overview of web mining and identify tasks for each of the goal systems.</td>
<td>The model could be bigger in order to make it easier for readers. I had some problems in finding specific tasks or goals.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Participant 1</td>
<td>Not sure.</td>
<td>Not sure.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Participant 2</td>
<td>This approach can be useful when studying for an exam. It can be considered like lecture notes, which only cover the important concepts, and so facilitate and speed up the learning and review process as opposed to reading multiple textbooks and/or documents.</td>
<td>The larger the text, the harder it is to understand and comprehend the material. Additionally, it is hard to remember the covered concepts, when the text is too long. Furthermore, it is hard to identify knowledge gaps.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Participant 3</td>
<td>Research in information technology domain, in general.</td>
<td>Not sure.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Group W Participant 4</td>
<td>Any technical field would be fine being represented with text.</td>
<td>It would be more comprehensive if the crawlers were explained before hyperlink structure.</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Table 12: Reflection questions**
8 Conclusion and Future Work

Uren et al. (2006), Nasser (2012) and many others have observed and reported that with the introduction of information technologies, researchers are benefiting from improved access to the documents through various technological means. Nevertheless, the growth of access to information brings its own challenges. One challenge is that keeping up with the changes in the information landscape, and how the information can be structured and presented to the users. Uren et al. (2006) took on these challenges by developing ScholOnto project, and utilize the notion of ‘sense making’ process in which the individuals and groups construct meaning from set of information. Nasser (2012) took a different approach to the challenge of expanding on means-ends relationships within the knowledge domain, and map the domain based with help of goal-oriented modelling language. In this research, we expanded on the approach taken by Nasser (2012) to improve upon the Goal-Oriented Know-How Mapping.

8.1 Research contribution

The contributions of this research are the following:

- We have provided a clear definition of the Goal-Oriented Know-How Mapping (GOKM) and its knowledge components. We describe what each of the knowledge components means, how these components can be discovered, and identified their relationships to each other.

- We have provided GOKM process documentation, which aids the researcher to codify publications. These documented processes describe in details what are the required steps of GOKM, and how they should be carried out to produce GOKM maps. The
documentation includes the process of codifying a single publication, formation of goals and solutions, and aggregation of the single publication into a larger domain.

- We have evaluated the GOKM approach by three complementary ways: a case study, feature analysis, and a controlled experiment.
- We have also taken initiatives in the development of proof-of-concept prototype that support GOKM process. The prototype developed for GOKM helped us to envision the system that encompasses all process involved in GOKM, from the codification of publication, to searching and browsing of the domain graphs.

Summarizing the contribution, the work has demonstrated the benefits of using the GOKM approach.

8.2 Limitations and future work

Although the GOKM approach to know-how mapping greatly help in organizing and highlighting know-how in technological domains, it is not without restrictions. Some of the limitations that need to be addressed remain the same as discussed in a dissertation by Nasser (2012).

In the work of Nasser (2012), domain experts are only eligible to codify scholarly articles within technology domains. Although, in this thesis we developed a knowledge repository and process documentation where the main concepts of the articles can be stored manually, the contribution of the experts is still required for knowledge codification. The process documentation developed in this thesis need to be validated by means of a user study in order to examine how individuals without expertise can utilize the documented processes and codify new publications for GOKM.
Introducing a mechanism that would allow a computer system to automatically codify the knowledge and generate the GOKM model, which corresponds to a user query, remains a work to be done. However, as Nasser (2012) also pointed out, such system may require the publications to follow a standard structuring format. Furthermore, the work of Hui et al. (2004) on automatic extraction of knowledge components can be extended to be used for GOKM purposes. In addition, we envision incorporating crowdsourcing principles into GOKM prototype to enable a larger number of contributors in the domain. The crowdsourcing principles focus on crowd knowledge, which enables us to attract more contributors to create and use GOKM approach.

Validation of the framework via applying the model to more technological domains, and involving more participants in the user study are also regarded as the future work of this thesis.

The final modification or improvement is concerning the experimentation and evaluation. In the experiment presented here, we mainly focused on evaluating the participants’ overall understanding of the GOKM features. We also validated their understanding of concepts such as goals, problems, solutions, mean-ends and decomposition relationships, as well as contribution links. The two remaining concepts that need to be validated include information sources (i.e. Publication’s title or author’s name), and assumptions. Furthermore, the GOKM prototype needs to undergo a user usability evaluation to validate its features.
References


Yang, S., Hou, J., Chen, J. (2013). A knowledge component extraction technology using figures and tables. Journal of Experimental & Theoretical Artificial Intelligence (June 2013), 25 (2), pg. 147-175

Web Mining knowledge domain references


Appendices:

Appendix 1 – Experiment questionnaire
Part 1 – Background Information

1. Gender: _____________________________
2. Age: _______________________________
3. English proficiency:
   - Beginner  Intermediate  advance
4. Education level:
   - Bachelor’s  Master’s  PhD
     Other (please specify): _______________________________
     Field of study: ______________________________________
5. How often do you search for information in scholarly articles:
   - Never  Occasionally  Sometimes  Usually  Every day
6. How often do you search for information about solutions to problems:
   - Never  Occasionally  Sometimes  Usually  Every day
7. Please rate your familiarity with goal-oriented modelling concepts and languages (e.g., i*, GRL, KAOS):
   - Unfamiliar  Somewhat familiar  Very familiar  Expert
8. Please rate your familiarity with knowledge management and concept mapping:
   - Unfamiliar  Somewhat familiar  Very familiar  Expert
9. Please rate your familiarity with the data mining domain.
   - Unfamiliar  Somewhat familiar  Very familiar  Expert
10. Please rate your experience in exploring technical domains.
    - Unfamiliar  Somewhat familiar  Very familiar  Expert
Web mining is a branch of data mining that aims to discover and retrieve valuable information and knowledge from the web hyperlink structure, web page contents, and web usage data [8]. Web structure mining focuses on discovering useful knowledge from hyperlinks (links) which form the web structure and has been proven to be a vital component of search engine technologies (e.g., Google’s PageRank).

Web content mining is the application of data mining techniques for discovering knowledge form the web page contents. Classification and clustering of web pages according to their topic and extraction of specific information from the web pages are examples of web content mining activities. Rather than looking for information in the web structure or contents, web usage mining is the effort to gather information of user information access patterns.

Many of the web mining processes are similar to data mining processes; however, there is a difference in data collection methods. Data collection processes in web mining are referred to as web crawling, a process that automatically discovers new web structure and the contents.
In order to improve information search and retrieval, researchers resort to the understanding and analysis of hyperlinks and the structure of the web [8]. A major part of this process is the discovery of hyperlink structure. Various techniques were proposed for that task. For example, PageRank [1] is a high search precision algorithm with the advantage of being query independent as well as ability to control spams. However, PageRank cannot identify the pages that are authoritative on the query topic. Another example is Anchor Text [1] in which the page is associated with the pages the link points to. This has two advantages: first, anchors often provide accurate description of web pages than the pages themselves, and second, anchors make the search possible on non-text documents. Topic-Sensitive PageRank [10] as an alternative has the advantage of being scalable as well as being accurate.

Another approach for analysing the link structure is to use information authority discovery. HITS [6], takes the approach of authority discovery and is search query sensitive. The main strength of HITS is its ability to rank pages according to the query topic, which may be able to provide more relevant authority and hub pages. HITS does not have the anti-spam capability like that of PageRank. Another drawback of HITS is its high overhead. Alternatively, SALSA [7] improves the search precision and overhead as well as spam control [8].

Before analysing the link structure, there is a need to discover and gather the Web content. Web crawlers are programs that automatically download Web contents. There are different types of crawlers: universal and preferential crawlers. Universal crawlers maintain their indices by taking advantage of incremental index updates and reduced crawling overhead [8]. Preferential crawlers are divided into two classes of topical and topic-specific crawlers. Topical crawlers are able to take advantage of being independent of the training set and the fresh crawling page hits. However, the search is slow compared to a traditional search engine [8]. Topic-specific crawlers gain significant savings in crawling overhead and help in keeping the crawl more up-to-date. However, these crawlers are dependent on the training sets that guide them. Within that group, focused crawlers are not considered an adaptive and efficient crawler and depend on training classifiers [8]. Context-Focused Crawlers, also depended on training classifiers, help the crawling efficiency [8]. The rise of social networks in recent years has opened up a new field for web crawlers to improve and innovate [8].
Before the documents in a collection can be used for information retrieval, some pre-processing tasks are usually performed. For texts and web pages the tasks are stopword removal, stemming, and duplicate detection [8]. Stemming [8] increases the recall and reduces the size of the indexing structure. However, it can hurt precision. Duplicate detection [8] can reduce the index size. Stopword removal may cause lower performance on phrase search, however, SEO (Search Engine Optimization) is improved due to better indexing [8].

To better process the information web page formatting need to be removed. One way of doing that is through template detection. The two approaches to template detection are page level template detection and batch web page processing [9]. Page level template detection can be done either by Isotonic smoothing [2] that detects templates effectively and accurately, or by incremental web page template detection [5] that reduces the required storage and results in high speed data refreshing.

Another approach to template detection is batch web page processing, which can be achieved by applying Site Style Trees (SST) [3] that reduces the noise data in web pages, neglects any visual cue information, helps in structure and content analysis, and improves the search precision. Alternatively, it can be done by applying Shingle [9] which also improves search precision, however, performance declines over large scale data.

Extracting the actual information from a Web page is usually done by finding the primary informative blocks. The reason for that is that Web pages contain several kinds of items that cannot be classified as the primary content and therefore informative blocks needs to be extracted [4]. The two kinds of informative block extractions are FeatureExtractor and ContentExtractor. FeatureExtractor shows poor precision and recall, nevertheless it improves the search time as well as reducing the index size [4]. ContentExtractor produces excellent precision and recall as well as reducing the index size, decreases the search time, and eliminates redundancies [4].
GOKM material (participants were provided a larger and a coloured version of this model):
References:


Based on the provided material on the web mining domain, please answer the following questions. You will be able to refer back to this material when answering questions. You are also allowed to mark up or highlight the model if you wish.

We kindly asked you to record the time for each of the questions. Please record the time in minutes and seconds.

Questionnaire starting time: ____________

1. *Start time: (_______________) What are the alternative methods for discovering hyperlinked web structure?*

2. *Start time: (_______________) What are the alternative methods for text and web page pre-processing for data mining?*

3. *Start time: (_______________) What is the problem that HITS and SALSA methods aim to solve?*

4. *Start time: (_______________) What is the problem that the Feature Extractor method aims to solve?*
5. **Start time:** (_______________) Solutions to the web template detection problem include batch web page processing and page level template detection. Are both of these techniques required to address this problem?

6. **Start time:** (_______________) How do Site Style Trees (SST) affect noise data in the web page?

7. **Start time:** (_______________) Describe the effects of “stopword” removal.

8. **Start time:** (_______________) Provide a summary of the solution trade-offs when crawling the web for specific topics.
9. **Start time: (_______________)** Which of the alternative techniques for batch web template processing has the advantage? Please explain.

**End time: (_______________)**

10. **Start time: (_______________)** What are the distinguishing effects of focused crawlers and context-focused crawlers?

**End time: (_______________)**

11. **Start time: (_______________)** In the provided material, can you identify any gaps in the knowledge in this domain? Explain how you are able to identify these gaps (if any).

**End time: (_______________)**

Completion time: _________
Part 3 – Reflection

Having just completed the above tasks, we kindly ask you to reflect on your experience. For each of the statements below, please mark your level of agreement and provide additional comments. In the following, “the approach” refers to the goal-oriented know-how map of the domain of web mining.

1. The approach helps a newcomer to a domain learn about the domain in a short amount of time.
   
   Strongly disagree  disagree  unsure  agree  strongly agree
   
   Please comment on your answer:

2. The approach helps a newcomer to obtain an overview of the domain comprehensively.
   
   Strongly disagree  disagree  unsure  agree  strongly agree
   
   Please comment on your answer:

3. The approach enables researchers to identify knowledge gaps in a short amount of time.
   
   Strongly disagree  disagree  unsure  agree  strongly agree
   
   Please comment on your answer:
4. The approach enables researchers to easily identify knowledge gaps.

<table>
<thead>
<tr>
<th>Strongly disagree</th>
<th>disagree</th>
<th>unsure</th>
<th>agree</th>
<th>strongly agree</th>
</tr>
</thead>
</table>

Please comment on your answer:


5. Please provide a scenario where such an approach would be useful and beneficial for understanding and exploration of a knowledge domain.


6. Please suggest areas of improvements or shortcomings of the approach.


Appendix 2 – Questionnaire solution’s manual
1. What are the alternative methods for discovering hyperlinked web structure?
   - Anchor text
   - PageRank

2. What are the alternative methods for text and web page pre-processing for data mining?
   - Stemming
   - Duplicate detection

3. What is the problem that HITS and SALSA methods aim to solve?
   - Information authorities to be discovered

4. What is the problem that the Feature Extractor method aims to solve?
   - Primary information to be extracted from web pages

5. Solutions to the web template detection problem include batch web page processing and page level template detection. Are both of these techniques required to address this problem?
   - No, only one of the alternatives would be sufficient

6. How do Site Style Trees (SST) affect noise data in the web page?
   - Helps noise reduction
7. Describe the effects of “stopword” removal.

- Help smaller index and SEO
- Hurt phrase search

8. Provide a summary of the solution trade-offs when crawling the web for specific topics.

- Help low crawling overhead,
- Help crawling updatability,

9. Which of the alternative techniques for batch web template processing has the advantage? Please explain.

SST has the advantage since it has:
- Helps page structure & content analysis,
- Helps noise reduction,

10. What are the distinguishing effects of focused crawlers and context-focused crawlers?

Focused crawlers have:
- Some negative to crawling adaptability,
- Some negative efficiency,
- Hurt training classifier dependency.

11. In the provided material, can you identify any gaps in the knowledge in this domain? Explain how you are able to identify these gaps (if any).

“Social networks” need to be explored for solutions.
### Appendix 3 – Experiment’s results

<table>
<thead>
<tr>
<th>Q 1</th>
<th>Not Familiar with Goal Oriented Concepts (group G0)</th>
<th>Familiar with Goal Oriented Concepts (group G1)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Participant 1</td>
<td>Participant 2</td>
</tr>
<tr>
<td></td>
<td>Time Spent</td>
<td>0:01:30</td>
</tr>
<tr>
<td></td>
<td>Mark</td>
<td>5.00</td>
</tr>
<tr>
<td></td>
<td>Time Spent</td>
<td>0:01:20</td>
</tr>
<tr>
<td></td>
<td>Mark</td>
<td>5.00</td>
</tr>
<tr>
<td></td>
<td>Time Spent</td>
<td>0:01:45</td>
</tr>
<tr>
<td></td>
<td>Mark</td>
<td>5.00</td>
</tr>
<tr>
<td></td>
<td>Time Spent</td>
<td>0:01:40</td>
</tr>
<tr>
<td></td>
<td>Mark</td>
<td>5.00</td>
</tr>
<tr>
<td></td>
<td>Time Spent</td>
<td>0:00:30</td>
</tr>
<tr>
<td></td>
<td>Mark</td>
<td>5.00</td>
</tr>
<tr>
<td></td>
<td>Time Spent</td>
<td>0:00:45</td>
</tr>
<tr>
<td></td>
<td>Mark</td>
<td>5.00</td>
</tr>
<tr>
<td></td>
<td>Q 8</td>
<td></td>
</tr>
<tr>
<td>-------</td>
<td>-------</td>
<td>-------</td>
</tr>
<tr>
<td>Time Spent</td>
<td>0:01:33</td>
<td>0:01:12</td>
</tr>
<tr>
<td>Mark</td>
<td>5.00</td>
<td>5.00</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Time Spent</td>
<td>0:03:50</td>
<td>0:02:14</td>
</tr>
<tr>
<td>Mark</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Time Spent</td>
<td>0:05:35</td>
<td>0:00:53</td>
</tr>
<tr>
<td>Mark</td>
<td>5.00</td>
<td>5.00</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Time Spent</td>
<td>0:02:20</td>
<td>0:01:13</td>
</tr>
<tr>
<td>Mark</td>
<td>3.00</td>
<td>1.00</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Time Spent</td>
<td>0:10:55</td>
<td>0:02:59</td>
</tr>
<tr>
<td>Mark</td>
<td>0.00</td>
<td>5.00</td>
</tr>
<tr>
<td>Total:</td>
<td>43.00</td>
<td>41.00</td>
</tr>
<tr>
<td>Question</td>
<td>Participant 1</td>
<td>Participant 2</td>
</tr>
<tr>
<td>----------</td>
<td>---------------</td>
<td>---------------</td>
</tr>
<tr>
<td><strong>Q 1</strong></td>
<td>Start Time: 8:12:55</td>
<td>End Time: 8:14:05</td>
</tr>
<tr>
<td></td>
<td>Start Time: 8:14:14</td>
<td>End Time: 8:16:52</td>
</tr>
<tr>
<td></td>
<td>Start Time: 8:17:00</td>
<td>End Time: 8:20:22</td>
</tr>
<tr>
<td></td>
<td>Start Time: 8:19:12</td>
<td>End Time: 8:21:02</td>
</tr>
<tr>
<td></td>
<td>Start Time: 8:20:35</td>
<td>End Time: 8:21:07</td>
</tr>
<tr>
<td></td>
<td>Start Time: 8:26:30</td>
<td>End Time: 8:26:18</td>
</tr>
<tr>
<td></td>
<td>Start Time: 8:28:30</td>
<td>End Time: 8:28:30</td>
</tr>
<tr>
<td><strong>Q 2</strong></td>
<td>Start Time: 8:43:32</td>
<td>End Time: 8:43:26</td>
</tr>
<tr>
<td></td>
<td>Start Time: 8:45:41</td>
<td>End Time: 8:46:39</td>
</tr>
<tr>
<td></td>
<td>Start Time: 8:47:36</td>
<td>End Time: 8:48:39</td>
</tr>
<tr>
<td></td>
<td>Start Time: 8:50:39</td>
<td>End Time: 8:51:53</td>
</tr>
<tr>
<td></td>
<td>Start Time: 8:52:53</td>
<td>End Time: 8:53:53</td>
</tr>
<tr>
<td><strong>Q 3</strong></td>
<td>Start Time: 8:55:03</td>
<td>End Time: 8:55:07</td>
</tr>
<tr>
<td></td>
<td>Start Time: 8:57:20</td>
<td>End Time: 8:57:24</td>
</tr>
<tr>
<td></td>
<td>Start Time: 8:59:42</td>
<td>End Time: 9:00:42</td>
</tr>
<tr>
<td></td>
<td>Start Time: 9:02:04</td>
<td>End Time: 9:02:08</td>
</tr>
<tr>
<td><strong>Q 4</strong></td>
<td>Start Time: 9:04:23</td>
<td>End Time: 9:04:27</td>
</tr>
<tr>
<td></td>
<td>Start Time: 9:06:47</td>
<td>End Time: 9:06:51</td>
</tr>
<tr>
<td></td>
<td>Start Time: 9:08:34</td>
<td>End Time: 9:08:38</td>
</tr>
<tr>
<td><strong>Q 5</strong></td>
<td>Start Time: 9:10:55</td>
<td>End Time: 9:10:59</td>
</tr>
<tr>
<td></td>
<td>Start Time: 9:12:15</td>
<td>End Time: 9:12:19</td>
</tr>
<tr>
<td><strong>Q 6</strong></td>
<td>Start Time: 9:16:55</td>
<td>End Time: 9:17:00</td>
</tr>
<tr>
<td></td>
<td>Start Time: 9:18:34</td>
<td>End Time: 9:18:38</td>
</tr>
<tr>
<td></td>
<td>Start Time: 9:20:15</td>
<td>End Time: 9:20:19</td>
</tr>
<tr>
<td><strong>Q 7</strong></td>
<td>Start Time: 9:22:33</td>
<td>End Time: 9:22:37</td>
</tr>
<tr>
<td></td>
<td>Start Time: 9:24:15</td>
<td>End Time: 9:24:19</td>
</tr>
<tr>
<td></td>
<td>Start Time: 9:26:34</td>
<td>End Time: 9:26:38</td>
</tr>
<tr>
<td><strong>Q 8</strong></td>
<td>Start Time: 9:28:15</td>
<td>End Time: 9:28:19</td>
</tr>
<tr>
<td></td>
<td>Start Time: 9:30:33</td>
<td>End Time: 9:30:37</td>
</tr>
<tr>
<td></td>
<td>Start Time: 9:32:51</td>
<td>End Time: 9:32:55</td>
</tr>
<tr>
<td><strong>Q 9</strong></td>
<td>Start Time: 9:35:15</td>
<td>End Time: 9:35:19</td>
</tr>
<tr>
<td></td>
<td>Start Time: 9:37:33</td>
<td>End Time: 9:37:37</td>
</tr>
<tr>
<td><strong>Q 10</strong></td>
<td>Start Time: 9:39:51</td>
<td>End Time: 9:39:55</td>
</tr>
</tbody>
</table>

**Time Spent:**
- **Q 1:** 0:01:10
- **Q 2:** 0:00:58
- **Q 3:** 0:03:04
- **Q 4:** 0:01:10
- **Q 5:** 0:00:58
- **Q 6:** 0:00:39
- **Q 7:** 0:02:12
- **Q 8:** 0:03:12
- **Q 9:** 0:02:07
- **Q 10:** 0:00:26
<table>
<thead>
<tr>
<th></th>
<th>0.00</th>
<th>5.00</th>
<th>1.00</th>
<th>1.00</th>
</tr>
</thead>
<tbody>
<tr>
<td>Q 11</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mark</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>5.00</td>
</tr>
<tr>
<td>Start Time</td>
<td>8:30:41</td>
<td>5:22:32</td>
<td>8:51:45</td>
<td>8:39:02</td>
</tr>
<tr>
<td>End Time</td>
<td>8:34:30</td>
<td>5:24:26</td>
<td>8:53:28</td>
<td>8:39:54</td>
</tr>
<tr>
<td>Time Spent</td>
<td>0:03:49</td>
<td>0:01:54</td>
<td>0:01:43</td>
<td>0:00:52</td>
</tr>
<tr>
<td>Total:</td>
<td>27.00</td>
<td>47.00</td>
<td>34.00</td>
<td>39.00</td>
</tr>
</tbody>
</table>
Appendix 4 – Demo domain

Accommodation to be acquired

Place to be taken under agreement

Property to be purchased

Residential place to be taken for rent

Find property

Take Hotel Apartment

Rent

Make an offer on property

Possession to be exchanged for rent (money)

Get mortgage

Rent Condo

Perform credit check

Rent Room

Mortgage request to be approved by bank

Rent basement

Offer to be placed on property

Rent Condo

Make an offer on house

Owner chooses tenant

Room-mate to be found

Rent Condo

Make an offer on condo

Tenant finds his/her room-mate

Get mortgage

Complete property inspection
Appendix 5 – Questionnaire design steps

1.1 Preliminary decisions

This part of experiment design requires the researcher to recognise and understand the experiment’s objectives before taking further steps. This is to ensure that questions are designed to address the research objectives. As also suggested by Sreejesh et al. (2013), a researcher should go through research studies that are similar to the current study, which helps the planning of the survey. In pursuit of this goal, we have reviewed the work of Nasser (2012), Shum et al. (2000) and Uren et al. (2006), which are similar to this research. Inspired by the work of Uren et al. (2006), we also took the path of creating an experiment that enables us to determine whether the GOKM could be used to communicate the domain’s information in comparison to a written review of the same domain.

1.1.1 Target participants and sample size

At this early stage of experiment design, target participants should be considered. Sreejesh et al. (2013) point out that sampling frame would be drawn after the target respondents are defined. For this research, we needed three sets of target participants. The first group should consist of participants that are familiar with goal-oriented requirement engineering and modelling. The second group should include participants with no experience with goal-orientation concepts. The third group should include participants who have some background on literature review.

As this is an exploratory research, we chose to have a sample size of twelve participants in total, which are divided into groups of four, as described above. Our design is also supported by previous studies, e.g., Nasser (2012) and Uren et al. (2006), in which the sample size was set to six participants. In this work, we have increased the sample size to get richer and more meaningful data from the experiment.
1.2 Question’s content

In this part of experiment design, Sreejesh et al. (2013) points out that irrespective of the type of research, a researcher has to find answers to five major questions while deciding the question content. These questions are provided and discussed in relation to this research in the following.

- What is the utility of the data collected? Each question should have a clear contribution to the experiment, and the research in general. Question screening, which analyses the usefulness of the data is a step that should be taken to ensure the utility of the questions. During the design of this experiment, questions were rigorously screened to select those that produce value to the research.

- How effective is a question in producing the needed data? This part of experiment design requires that each question passing from initial utility screening, goes through another round of screening to ensure that it produces effective data for analysis. Some double-barrelled questions that touched more than one issue separated into two questions or removed altogether.

In this part of experimental design, we had to pay special attention that both groups of participant can answer the same questions based on the material provided to them. In all versions of the questionnaire, we asked the same questions from the participants, and provide them with different type of material (Goal-oriented know-how maps and written review). The process of modifying and screening the questions goes hand in hand with the principle of information equivalence (see Section 7.3.3.1) to ensure that both groups can effectively answer the questions about the knowledge domain.

- Can the respondent answer the question accurately? It is important to make sure that the respondent understands the question in a way that the researcher intended, in order to
eliminate potential incorrect answers. Sreejesh et al. (2013) points out to the three sources of respondent’s inability to answer the questions: genuine ignorance, inability to recollect, and inability to verbalize the answer.

The experiment is designed with the knowledge that the participants are not familiar with the web mining domain, and depend entirely on the information provided to them with the experiment. This helped to avoid respondents’ ignorance to answer the questions. Secondly, the experiment included instruction to respondents to refer back to the domain material and mark as they wish. This helps to avoid any potential mistakes that are caused by respondent's inability to recall the answer. Lastly, in order to limit the inability to verbalization of the answers, the questions were designed plainly and to the point.

- Is the participant willing to answer the question accurately? The experiment was designed in a form that had no impact on the willingness of the participant to answer the questions. In order to mitigate this, no personal or sensitive questions are asked of the participants.

- What is the probability of the responses being influenced by external events? The effects of external events are controlled by setting up the experiment environment beforehand. We also provide enough time for participants to complete the questionnaire.

### 1.3 Response format

In the design of the experiment, we used different types questions for specific response types. In the main part of the experiment, which is the domain questions, we used open-ended questions, which are useful to enable the participant to provide answers based on their understanding of the question and the material. The open-ended questions encourage participants to respond in their own words without being limited to pre-defined answers. In this case, participants have greater freedom of expression and bias is limited due to wider response range. However, this type of
question has a downside since coding open-ended questions is difficult and time consuming for the researcher. The questions in this section of the experiment were marked according to their correctness and completeness, and these marks are used to get an average for each question among the participants.

1.4 Question wording

When designing the experiment, researcher should focus on effective communication of what is being asked for and what is provided to the participants. Sreejesh et al. (2013) identified some consideration factors while framing a questionnaire. Following these factors, we also describe the factors that we had taken into matter when designing the questions.

➢ Shared vocabulary

The experiment is about exchange of ideas between the researcher and the participant, and therefore it is imperative for the language to be kept simple and clear between both parties (Sreejesh et al., 2013). In this research, most of the participants were not familiar with the vocabulary of web mining domain; however, the questions were designed in simplest and clearest form to express what the researcher is looking after. The appropriate choice of words ensures the simplicity of the questions as well as reducing the ambiguous or vague questions.

➢ Unsupported assumptions

An experiment should not be designed in a way that contains assumptions not explained in the questions or hidden to the participants. An experiment should not leave anything for the participants to interpret. Unsupported or implied assumptions tend to produce a misleading response in the experiment (Sreejesh et al., 2013). In this experiment, if the participant would
have felt that there are hidden assumptions to particular questions, they get a chance to explain themselves.

- **Biased wording**
Experiments should avoid committing a bias. One type of bias is biased wording, which influence the responses of the participants in a predetermined ways. Biased and loaded words should be avoided to eliminate biased responses. Reformatting the questions in different way is one way of avoiding bias wording. Stone (1993) points out that the a question may appear unbiased until the responses are being interpreted and analysed. The objective here is to ensure that the question is no more likely to trigger one kind of response than the other. In order to reduce wording bias, we designed the neutral questions, which avoid favouring or hinting one type of answer over other potential answers.

- **Double-barreled questions**
Double-barreled questions, which combines two different questions into one should be avoided. These types of questions tend to confuse the participant in answering, and produce invalid data for analysis since two or more ideas are included. One means of avoiding such questions is to divide the questions into two separate questions. In the initial draft of the experiment, we encountered the double-barreled question problem, and replaced the question to avoid the invalidity of the data.

**1.5 Experiment sequence**
The sequence in which the participants are presented with the questions is an important aspect of experiment design. In the design of the experiment, we start with the demographics and background questions, which are a necessary component of the experiment. This is because responses to this experiment could not be effectively analysed unless they are classified based on
the background of the participant. The next section of the experiment is dedicated to the questions related to exploring of the web mining domain which are more specific questions that requires the participant to go other the material and answer the questions. The order of the questions in this section was arranged so that doesn’t hint to the participants about what the research is going to evaluate from their answers. The participant’s ability and the quality of the answers is not expected to be affected by the order of the questions.

The later part of the experiment consists of reflection questions, which are a mixture of open-ended questions and scale-based questions. The intention behind the reflection questions is to gather data about how participant reflect on their experience of exploring the knowledge domain while answering the questions.

1.6 Experiment pre-test, revision and final draft

In order to eliminate the potential remaining flaws and problems that might remain hidden throughout the design of the experiment, Sreejesh et al. (2013) urge the researchers to pre-test the experiment with a small sample of participants. Pre-testing includes testing different aspects of the experiment, including the content and sequence of the questions, and material provided to the participants. Pre-testing, or piloting the experiment enables the researcher to revise the experiment by identifying the flaws and ambiguity of the questions or the procedure.

We have also ran a pilot with one participant from written review group and GOKM group before conducting the actual experiment. Pilot experiment procedure is identical to the actual experiment, except the number of participants. The piloting of the experiment proved to be an import stage of the experiment especially for the textual representation group. After the pilot, we notice some issues with the written review of the domain that we revised for the actual
experiment. The piloting also helped to identify the questions that needed more clarification at the time of introduction.