NTNU
Innovation and Creativity

# Finding and Mapping Expertise Automatically Using Corporate Data

**Audun Vennesland**

Master of Science in Informatics
Submission date: July 2007
Supervisor: Trond Aalberg, IDI

Norwegian University of Science and Technology
Department of Computer and Information Science

# Assignment

An Expertise finder is a common term for computer applications aiming at locating, organizing and presenting an organizations' knowledgeable employees. Traditionally, this kind of applications have often used semi-automatic methods often following a database approach where the the employees themselves describe their expertise areas- and levels. This approach carries several flaws. What is needed in this field are Expertise finders which locate, organize and present an organizations' expertise in an automatic and objective manner. This thesis shall evaluate the state-of-art in expertise finding and suggest practical principles that might overcome the limitations caused by semi-automatic Expertise finders.

## Abstract

In an organization, both management as well as new and experienced employees often have a need to get in touch with experts in a variety of situations. The new staff members need to learn how to perform their job, the management need - amongst other things - to man projects and vacancies, and other employees are often dependent on others' expertise to accomplish their tasks.

Traditionally this problem has often been approached with computer applications using semi-automatic methods involving self-assessments of expertise stored in databases. These methods prove to be time-consuming, they do not consider the dynamics of expertise and the self-assessed expertise is often difficult to validate.

This report presents an overview of issues involved in expertise finding and the development of a simple, yet effective prototype which tries to overcome the mentioned problems by using a fully automatic approach. A study of the Urban Development area at the Municipality of Trondheim is carried out to analyze this organizations' possessed expertise, sought after expertise and to collect necessary information for building the expertise finder prototype. The study found that a lot of expertise evidence is found in the formal correspondence archived in the case handling systems' document repository, and that the structure and content of these documents could fit a fully-automatic Expertise finder well.

Four alternative test cases have been evaluated during the testing and evaluation of the prototype. One of these test cases - where expert profiles are modelled on-the-fly based on employees' names occurring in formal documents - is able to compete with- and in some cases outperform evaluation scores presented in related research.

# Acknowledgements

I would like to thank my supervisor, Trond Aalberg, for his help and guidance throughout this thesis' lifecycle. I would also like to thank my father, Øystein Vennesland, for thorough proof-reading and constructive advice. I also owe great thanks to all involved personnel at the Urban development area for keeping a positive and helpful attitude during the case study. Last, but not least, I wish to thank my wife Tone for being patient those times I gave more attention to this thesis than to her.

# Contents

# List of Figures

# List of Tables

# Chapter 1

# Introduction

## 1.1 Motivation

Knowledge Management Systems (KMS) comprise various kinds of computer applications aiming to facilitate the creation, sharing and application of peoples knowledge. A special field within knowledge management is focused on the finding and mapping of peoples expertise. Applications within this domain use names such as ”People Finders”, ”Corporate Yellow Pages” or ”Expertise finders”. Traditionally these applications have often often been using a database approach where the expertise is described by the personnel themselves, and then it is stored in a database with a searchable and/or browseable user interface.

This approach has several limitations. Firstly, expertise is highly dynamic. Most people gain new knowledge every day and specifying new expertise areas into the database-stored expertise profile is often a time-consuming task.

Secondly, it is difficult to evaluate and accurately describe ones expertise areas and levels. How do you compare your expertise to other employees in the organization? Will different people with similar expertise describe it with similar terms? Is your competency about UML modelling really worth mentioning or should someone else be rated as the uppermost expert in this area?

Thirdly, expertise is not only dynamic, it is also highly subjective in that it really depends on who’s asking for it; if you are told to describe your expertise to someone with little knowledge of the domain in question you may describe it in general terms, but if you were to describe your expertise to someone with major knowledge you would probably describe your expertise in a more detailed manner.

Fourthly, validating other peoples expertise is difficult. How do you discriminate the real expert from someone who merely has some mediocre knowledge of a subject and all you base your validation upon is their self assessment of their expertise? These are important questions related to the location of expertise, and prove that automatic and objective methods are needed.

## 1.2 Problem statement

A central objective in this thesis is to investigate different approaches to the finding and mapping of expertise. This investigation should form a basis for an alternative approach that will try to overcome the limitations caused by semi-automatically driven expert finders. Problems needing answers in this context are:

- How to define expertise?
- Which corporate sources contain expertise evidence?
- How should the experts be presented?
- How to validate the presented experts?

### How to define expertise?

Before locating experts in an organization we need to know what kind of characteristics define an expert. These characteristics may be domain specific or general, depending on the domain in question. For instance, programming skills or certifications may be an important criterion in one line of work, but totally irrelevant in another. Expertise is often mentioned in the same context as knowledge and competence. What seperates these three concepts, and how do they relate to what expertise finders are meant to provide? An analysis of literature within this domain and a case study in an organization will hopefully provide some answers to this problem.

### Which corporate sources contain expertise evidence?

A number of different sources has previously been used to locate expertise automatically: e-mail archives[6], intranet web pages [28], software source code [27] are some of them. Organizations often have several different sources where expertise evidence might be found. For instance, an organization may have several different document storages housing different kinds of document themes (e.g. project reports, formal letters, rules and guidelines) with different kinds of formats (Word, PDF, HTML etc.). All these storages might hide important evidence of expertise and could be utilized to find experts in an expert finder system.

### How should the experts be presented?

The seeker of expertise should be presented with a profile of the experts the expertise finder system has located given a query. This profile ought to be kept updated as a person's expertise accumulates as he performs new tasks and receive new knowledge. What kind of information this profile should contain depends on the following conditions:

1. **Sources**: To successfully implement and use an expertise finder based on an automatic approach, there has to be some way to couple the sources' content with a person or several persons.
2. **Expertise indication**: What kind of indications exists that makes it possible to discriminate one expert from others, and how can these indications enable the expertise seeker (the user) to validate that the recommended experts actually possess the demanded expertise?
3. **Contact information**: What kind of contact information is available to connect the expert seekers with the experts? This kind of information can be either impersonal information such as the physical or logical location of the experts or the experts' telephone number, or it can be more personal information that makes it easier to get familiar with the expert, such as an image or education information.

### How to validate the presented experts?

To validate if a person actually is an expert you either need to know the expert candidates well enough to be able to judge whether they possess the demanded expertise or not, receive references from other people in the same situation as you who have previously gotten expert advice, or receive references to the experts' previous work to be able to choose the right expert.

## 1.3 Research design

The problem statements defined in the previous section were suited for a combination of three different research techniques: a literature review, a case study and an experiment.

### Literature review

The concept definitions requires a literature review, and especially the notion of expertise - which is rather context dependent - needs a thorough review of literature in the domains of psychology, sociology and technology. Besides this, the literature review forms a basis for conducting the meetings and interviews in the case study and also a foundation for developing the expertise finder prototype.

### Case study

A case study is ideal to investigate a lot of information about an empirical restricted resource - like an organization [34]. The case study is carried out to examine an organization's personnel, potential expertise, inquired expertise, sources where expertise evidence might be located, and other findings related to the development of the expertise finder prototype.

### Prototype experiment

A prototype expertise finder is being developed to perform testing of the underlying principles in this thesis. To evaluate these principles, prototype experimenting using evaluation measures found in the field of information retrieval is used.

The research process is illustrated in figure 1.1.

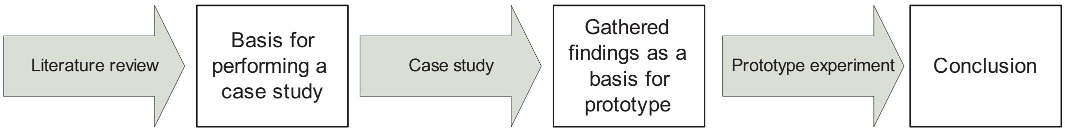

Literature review
Basis for performing a case study
Case study
Gathered findings as a basis for prototype
Prototype experiment
Conclusion

Figure 1.1: Research design

## 1.4 Report organization

The remaining part of the report is structured as follows:

Chapter two focuses on the concept of expertise and compares this phenomenon with the related terms competence and knowledge.

Chapter three is dedicated to state-of-the-art in expertise finding. Expert finders can be put into three different categories of approaches. These categories are:

the database approach, expertise finding based on social networks and expertise profiles based on electronic evidence.

Chapter four contain a run-through of the information retrieval domain and explains some techniques often used in the creation, execution and evaluation of expertise finders as well as other information retrieval systems.

In chapter five the study carried out at the City Development area in the Municipality of Trondheim is described. The chapter begins with a description of the main objectives and then follows a presentation of the case and the findings from the study.

Chapters six and seven describe the design and development of an expertise finder prototype.

Chapter eight describes the testing and evaluation of the expertise finder prototype. Evaluation metrics presented in chapter four are applied to evaluate the underlying principles of the prototype.

Chapter nine contain the thesis' discussion and the conclusions of this thesis are presented in chapter ten along with a description of future work.

# Part I

# Theory from Literature Review

# Chapter 2

# Expertise

## 2.1 Expertise, Competence or Knowledge?

Expertise, competence and knowledge are three highly interconnected concepts, all relevant to what expert finders are meant to deliver. The first part of this chapter is devoted to a clarification of concepts and is setting the stage for the rest of the report. Further, the fields of Knowledge Management (KM) and Competency Management are two complementary fields gaining momentum in organizations of today. Expertise finders are first and foremost regarded as a Knowledge Management tool, but they are also highly relevant within Competency Management. They both consider expertise as the main premise for a sustainable competitive advantage, but with a slightly different emphasize. How Expertise finders can help facilitate these two management fields will end this chapter.

### 2.1.1 Knowledge

In the fields of knowledge and knowledge management it is common to separate knowledge into an explicit dimension and a tacit dimension. Explicit knowledge is knowledge which is easy to formulate and communicate, and is easily translated into documents, rules and procedures. Explicit knowledge can be based on objects or rules [8]. Knowledge is object based when it is expressed in words or exists in physical artefacts, e.g. documents, physical models or patents, and rule based when it exists within the organizations' rules, procedures and routines.

Tacit knowledge - on the other hand - is the kind of knowledge that exist within the individual and is difficult to express and transfer to other individuals or to documents. Nonaka and Takeuchi [30] uses the terms socialization, externalization, internalization and combination to explain the processes where tacit knowledge is transferred into another persons' tacit knowledge and where tacit knowledge is translated into explicit knowledge and vice versa (figure 2.1). Socialization is the process where tacit knowledge in one person is transferred into tacit knowledge in another person. This kind of mechanism may for instance happen during face-to-face communication. An example of this is the mentor-apprentice relationship. Externalization is the process where tacit knowledge is translated into explicit knowledge. This process may realize through metaphors, analogies, hypothesis or models that make abstract knowledge concrete. Internalization is the process where explicit knowledge is turned into tacit knowledge. This can be achieved e.g. through learning-by-doing activities or creating individual mental models of explicit knowledge. For explicit knowledge to become tacit, it helps if the knowledge is verbalized or diagrammed into documents, manuals or oral stories. When people consume written, explicit knowledge in documents, they may create their own mental models of this knowledge; they internalize the knowledge and hence make it tacit. Combination is when various explicit knowledge sources are combined into new explicit knowledge [30].

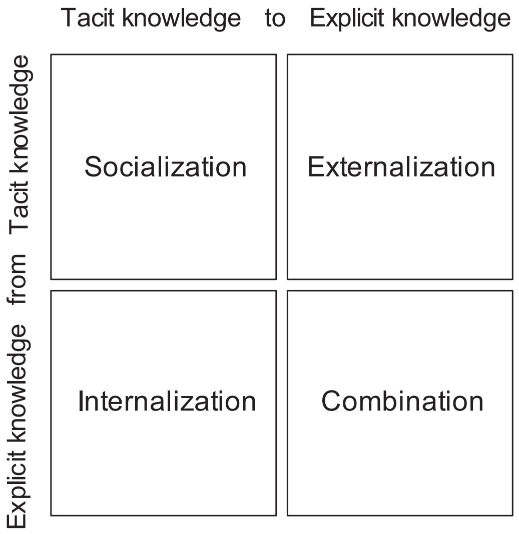

Tacit knowledge to Explicit knowledge
Explicit knowledge from Tacit knowledge
Socialization
Externalization
Internalization
Combination

Figure 2.1: Knowledge conversion

Besides the separation of knowledge into explicit and tacit dimensions, it is also common to see knowledge in the context of data and information (e.g. [10]) where data is a set of artefacts with no context. When you add context to the data, it becomes information. And when you finally add a portion of beliefs, commitment and action aspects to this information it becomes knowledge (figure 2.2). A contrast to this view comes from [35] referenced in [1] who argues that the situation is reverse; knowledge has to be produced before information and data exist. Even the most elementary piece of data has already been influenced by knowledge processes leading to its identification and collection.

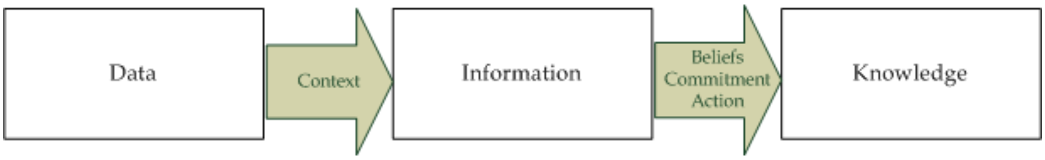


Figure 2.2: From data to knowledge

### 2.1.2 Competence

Lai [22] defines competence in this way:

> ”Competence is the gained knowledge, skills, abilities and attitudes that makes it possible to carry out the relevant functions and tasks according to the defined requirements and goals.”

Competence is what Davenport and Prusak [10] is talking about when they are talking about ground truth[1]. Ground truth means knowing what really works and what doesn’t. Competence both consists of, and is an enhancement of knowledge. You are not competent just as long as you are knowledgeable; you also need practical skills and abilities to be able to fulfil your tasks, i.e. ground truth.

Lai [22] defines different kinds of competence (figure 2.3). At a superior level we find a separation of formal competence and informal competence. Formal competence is the easily measurable competence found in e.g. résumés and curriculum vitas, whereas informal competence is built up by work and personal experiences, and is difficult to quantify. At a lower level we find a separation between professional competence, management competence, personal competence and social competence. At the lowest level, we find a separation between top competence and base competence where base competence denotes basic, more general knowledge and skills applicable to a wide range of areas. Top competence denotes competence at high professional level, and usually involves a high degree of specialization, not unlike expertise. Top competence is also called expertise competence [22].

### 2.1.3 Defining Expertise

> ”[The expert...] straightaway does the appropriate thing, at the appropriate time, in the appropriate way.” (Aristoteles)

[1] A concept the U.S Army’s Center for Army Lessons Learned (CALL) uses to describe the difference between learning by doing (practical learning) and learning by reading (theoretical learning) when they observe real military operations to gather and share knowledge discovered from these operations

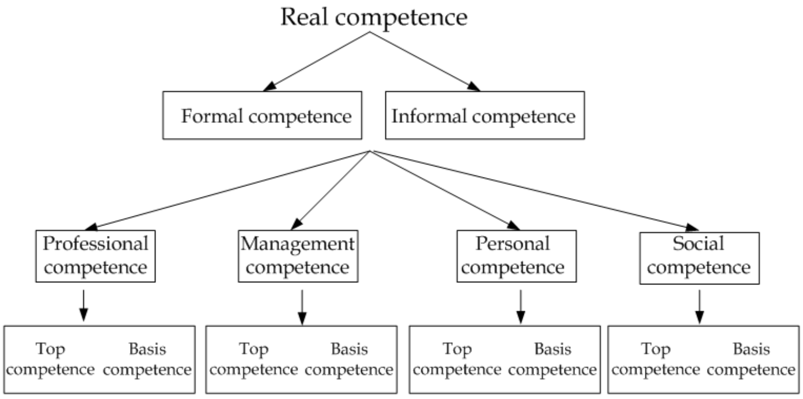


Figure 2.3: Competence hierarchy

Aristoteles description of an expert is a simple and easily understandable definition of expertise. It also point out why expertise is such an important asset for an organization; an organization need knowledgeable employees that possess the (tacit) knowledge and competence needed to make swift and proper decisions and to act upon it. However, this definition is too vague to explain how we might operationalize the notion of expertise so as to locate it. To clarify the concept of expertise even more, Ericsson [13] is a bit more precise:

> ”Expertise refers to the characteristics, skills, and knowledge that distinguish experts from novices and less experienced people”.

This definition holds several interesting aspects. One of them is that this definition implicitly states that expertise has a strong coherence with experience. He also combines skills and knowledge with the more indistinct notion of *characteristics* to define a persons’ expertise. What characterizes expertise and experts is much discussed within the psychological literature and a lot of research focuses on what significant skills separates experts from others. [7] found that experts often excel in the following seven areas:

1. **They find optimal solutions:**
   Compared to novices the experts excel by finding optimal solutions to problems, and they achieve it faster and more precise.

2. **Better ability to discover and recognize:**The experts are capable of discovering and recognizing patterns and features in different type of situations. They are also capable of discovering a deeper structure and the complexity in problems than what novices are capable of.

3. **Perform qualitative analysis:**
   Experts use a lot of time on analyzing the problem at hand, and considers both domain specific and general limitations before the correct strategy is

chosen.

4. **Have better control:**
   As experts have more knowledge, they are better capable of monitoring processes, finding errors and estimate tasks' complexity.

5. **Better at finding the right strategy:**
   Experts have an improved skill to find the proper strategy in solving problems than novices, and even though the same strategy is chosen, the experts perform the strategy better than novices.

6. **Are often more opportunistic:**
   Experts have a larger repertoire than novices. They often see alternative solutions and means for problem solving than the novice is capable of.

7. **Use less cognitive effort:**
   The expert is capable of collecting relevant domain knowledge and suitable strategies with minimal cognitive effort. Experts also have - gained through experience - an ability to employ their skills automatically.

Expertise, knowledge and competence are three different concepts, all carrying their significance, but at the same time they have a mutual relationship in that all three concepts are closely related to action. It is the tacit knowledge in an experts' mind that ensures that the expert is able to automatically see the solution to a given problem and it is the top competence that makes him able to act upon it. The combination of them - expertise - ensures that the expert is able to solve problems in a correct and swift manner according to the definitions by Aristoteles and Ericsson.

Yimam-Seid and Kobsa [36] identified two motives for seeking expertise. The first motive is to find someone who might provide useful information. This motive may be based on various information needs including the need for nondocumented information, a need for information that helps specify and explain problems, a need to leverage on others' expertise to e.g. filter out useful information, a need for interpretation and a socialization need. The second motive is to find someone who might perform a given organizational or social function. This motive requires a more structured search than the first motive. When searching for people who might provide the relevant information, one is interested in finding out ”who knows about topic x?' i.e. one is interested in finding someone possessing sufficient knowledge to answer a question or solve a problem. Whereas when searching for someone who might fill some function one is interested in how much they know about topic x, i.e. if they are competent of performing this function [36].

To keep things simple, this report continues to use the term expertise to explain what Expertise finders are meant to deliver, whether it is competence or knowledge the expertise seeker is pursuing. However, we still don't know how an Expertise finder can apply these definitions in mapping humans to expertise evidence found in various sources. One of the most obvious parameters used to signalize expertise is experience.

### Expertise and experience

According to e.g. [13] and [19] you need at least ten years of experience to be an expert in a given field. This is a rather non-balanced and generalistic point of view, but there is little doubt that to reach a level of expertise, you have to go through a long maturity process. Dreyfus and Dreyfus [23] try to illustrate this process in a five-stage framework (Figure 2.4).

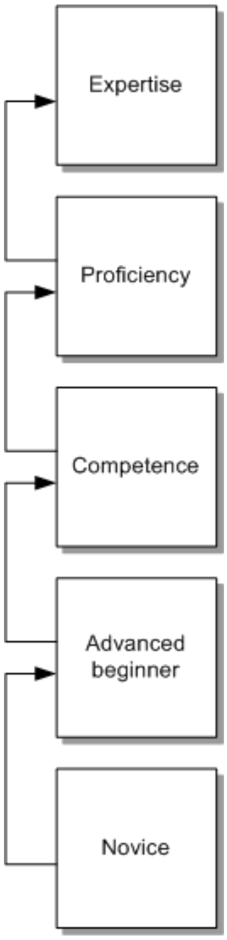


Figure 2.4: Expertise framework

The expertise process starts off with a presentation of the task at hand. The novice (Level 1) gets acquainted with some context free features that he is able to recognize (and understand) without any particular competence. Then the novice is given some ground rules to decide what strategies he will use to complete the task. These rules are remembered and crammed through practice. What is missing at this level is an understanding of the tasks' surrounding context. Without this context the given information does not make sense. When the performer gains experience by using this information in real situations, he develops an understanding for the relevant context, and may recognize other meaningful aspects related to the task. At this point the performer is transformed to an advanced beginner (Level 2).

What is still lacking to assure the necessary competence (Level 3) to understand and complete the task is the ability to focus on the important elements in the task and filter out the less important elements. When the performer gets more experience he will be able to recognize potentially relevant aspects connected to the task, but without the necessary competence to separate the really relevant from the less important aspects he will see the task as overwhelming and confusing. To reach the level of competency the performer need to learn how to formulate plans or strategies on how to focus on the relevant elements and rule out the non-relevant ones. Rules and procedures help the performer to choose the strategy and plan, but he is still not capable of choosing the strategy and plan for any deviation that may occur. At this stadium a lot of wrong decisions are made, but also some successful ones. If the performer is able to reflect over the errors made as well as the right decisions, the performer might rise to the next level.

Proficiency (Level 4) is achieved when the experience acquired in the previous stages lead to an intuitive pattern of reactions in stead of a complex rational evaluation in any occurring situation. At this stage the actions and decisions are perceived as easier and less stressing because the performer sees what needs to be done in stead of leaning to rules and procedures to complete the task. At this point the performer knows what is needed to get the job done, but he still needs to choose the right strategy to solve it.

What separates a proficient actor from an expert (5) is that the expert immediately sees what is needed to complete the task, and also how to solve it. There is no need for rules or procedures, calculations or computations. With the sufficient experience from different situations the experts mind carries out a decomposition of the task into sub-tasks that all requires a certain response. This leads to an immediate and intuitive response that characterizes expertise.

An interesting additional level to this framework is a level six; master. In a situation where the experts' expertise is to be shared with others, there is little use of an expert who is not capable of explaining his expertise so that others may appreciate and make use of it. A master is in this context a member of an elite group of experts who empowered by their dominant position among experts are qualified to share their knowledge with those at a lower level [7].

### 2.1.4 Expertise finding in Knowledge management

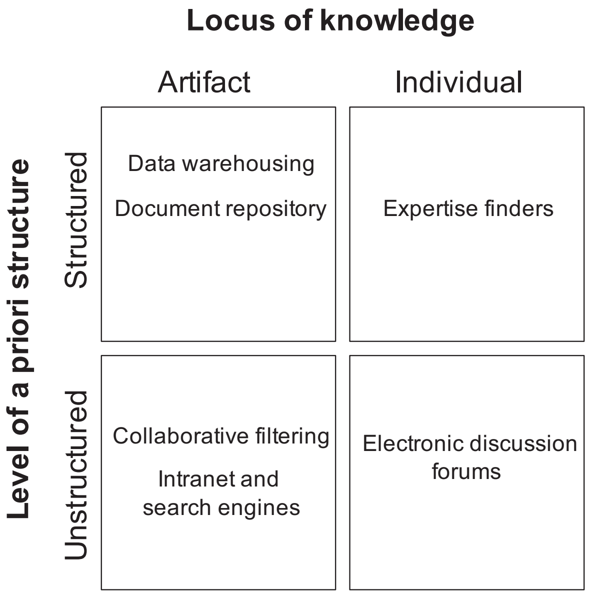
> "If there is no system in place to locate the most appropriate knowledge resources, employees make do with what is most easily available. That knowledge may be reasonably good, but in today's competitive environment reasonably good is not good enough" [10].

Knowledge management consists of strategies on how to get employees to create, share and apply knowledge in the best possible way. Hansen et al [17] separates knowledge management strategies into two different approaches; the codification approach and the personalization approach. The first one applies strategies suitable for an organization developing products or services based on explicit knowledge, whereas the strategies following the personalization approach are suitable for organizations utilizing mainly tacit knowledge to develop their products or services.

Both approaches use computer applications to support their objectives, but since they handle two different kinds of knowledge, they have to focus on different solutions. Computer applications fitting the codification approach often include storage solutions and taxonomies such as knowledge repositories which enables employees to browse through categories of relevant subjects or forums for transferring explicit knowledge such as discussion forums. Computer applications supporting the personalization approach focus on facilitating the knowledge transfer and the establishment of communication channels between employees. Hence, Expertise finders are one of the computer applications suitable for this approach.

Hahn and Sabramani [16] suggests a framework to classify computer applications supporting knowledge management. The framework has two dimensions. The horizontal dimension describes whether the knowledge is embodied within individuals (tacit knowledge) or whether it exists in artefacts (explicit knowledge). These two directions match computer applications used in the personalization approach and codification approach respectively. The vertical dimension describes whether the knowledge is structured (e.g. a database) or unstructured (e.g. dynamically created documents such as Intranet documents).

As figure 2.5 illustrate, Expertise finders or yellow pages of experts are placed in the cell where the knowledge is embodied within the individual and the level of a priori structure is considered structured. This is only partly true, because obviously Expertise finders contains individual knowledge and the names of the experts are static information are often stored in some kind of employee database.



| | | Locus of knowledge | |
|---|---|---|---|
| | | Artifact | Individual |
| Level of a priori structure | Structured | Data warehousing<br>Document repository | Expertise finders |
| | Unstructured | Collaborative filtering<br>Intranet and search engines | Electronic discussion forums |

Figure 2.5: KMS framework

However, unless the Expertise finder uses a database approach (see chapter 3.1) where the expert profiles is stored in a static manner, the information found in Expertise finders' expert profiles is often dynamically created as knowledge continuously accumulates. Thus Expertise finders, at least the ones employing dynamically created expert profiles can also belong to the cell where the locus of knowledge is individual and where the level of a priori structure is unstructured.

In a knowledge management perspective, Expertise finders can support especially the creation and sharing of knowledge in an organization. By establishing contact between expertise seekers and experts, this may facilitate a transfer of knowledge through informal communication leading to the creation of new knowledge through socialization or externalization and the general knowledge level in the organization might increase, presumably leading to higher efficiency and overall quality. However, even with the assistance from IT as an enabler in establishing the proper communication channels, knowledge management and the transfer of expertise is a difficult area without the employees' faith.

Hinds and Pfeffer [18] separates the complicating factors when it comes to sharing expertise into motivational and cognitive limitations. Some of the motivational limitations are that the competition for salary raises and promotions prevent experts from sharing their expertise as this is in fact their competitive edge. Another motivational limitation is when an organizational structures based on strict rules- and line of commands might prevent people from signalling their expertise as people working in the lower part of the organizational hierarchy might be reluctant to share their expertise with senior employees residing at a higher hierarchy level.

Even when the motivation to *share* ones knowledge is in place, the actual *transfer* of knowledge between an expert and a novice is complicated, with or without the assistance of technology. Some of the complicating causes are 1) It is difficult for the expert to recognize the knowledge level for the recipient which makes it difficult to decide the granularity level to use in the knowledge transfer, and 2) The way the expert abstracts and simplifies his knowledge.

One reason explaining why the expert is so unaware of the recipients knowledge level is explained by Bromme et al. [5]:

1. The correspondence hypothesis: One assumes that the other part has knowledge of a subject one self has knowledge about.

2. Overestimation hypothesis: One is likely to overestimate the general knowledge of a subject if one self possess it

3. Expertise hypothesis: A person who possess exceptional knowledge about a subject has a tendency to overestimate another persons' knowledge of this subject.

With experience and practice the cognitive state in humans change, and most experts organize parts of information into larger, logical chunks in memory [14]. This leads to that experts often abstracts and simplifies the problem which makes them able to process information more rapidly. The novice - at the other end - sees only the different parts of the same information, which makes the problem appear as complex and difficult (figure 2.6).

Because of the experts' ability to simplify the problem he is able to find solutions to problems quicker than novices. The downside to this is that this abstraction makes it difficult for the expert to recall the complexity and details the novice requires to understand the solution. Additionally, experts' knowledge is mainly tacit rather then explicit. Because tacit knowledge lies in the unconscious level it's difficult to formulate, which makes it difficult to transfer to others [18]

The above discussion goes to show that even though expert finders are able to establish communication channels between expert seekers and experts, the path to successful knowledge transfer is filled with obstacles, and both cognitive and motivational aspects must be resolved before the knowledge management can be successful.

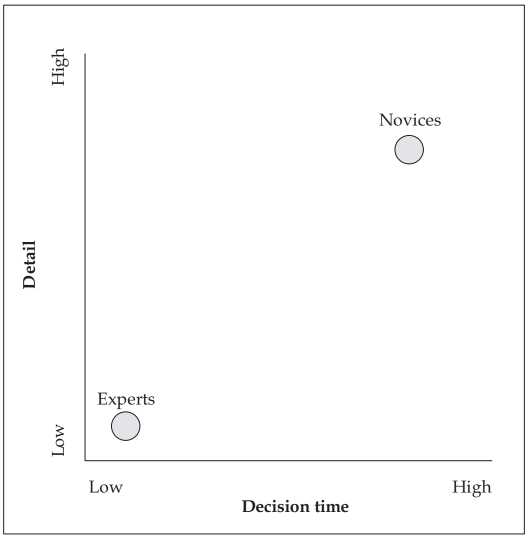


Figure 2.6: Experts and novices

### 2.1.5 Expertise finding in Competency management

Competency management is by Ley [24] defined to:

> ”[...]encompass all instruments and methods used in an organization to systematically assess current and future competencies required for the work to be performed and to assess available competencies of the workforce”

In contrast with knowledge management, which focuses on the creation, sharing and application of what employees know, competency management is more concerned with what employees actually are able to do, i.e. how their knowledge is applied in practice.

Lai [22] discriminates competency management from knowledge management in that knowledge management is setting the stage for competency management by emphasizing the social aspects, the informal competence building (i.e. informal methods for transferring tacit knowledge), and the sharing of knowledge in informal networks. As such, a relationship exists between these areas where knowledge management is responsible for creating a culture for knowledge creation, sharing and application, while competency management uses this culture as a foundation to locate and map competencies from motivated personnel.

Identical to knowledge management, competency management involves a mixture of technology and methods that facilitates the location of competent people or in some cases the lack of them. Competence Management Systems (CMS) is often used to evaluate the employees’ skills and competencies against certain measures set by the management of an organization. Competency management concerns the more hands-on properties related to employees, and thus both the formal and informal competencies may be easier to measure, locate and map than in the field of knowledge management where the ultimate goal is to create a knowledge culture for the creation, sharing and application of knowledge.

Expertise finders can facilitate the location of employees' skills and also areas where skillful employees do not exist and where the organization need to hire new employees or upgrade the competence of its existing staff. Hence, the main difference of how Knowledge Management and Competency Management can utilize Expertise finders is that in a Knowledge Management perspective they are used to establish communication channels that facilitates the transfer and creation of knowledge, whereas in a Competency Management perspective one can use them to find skillful employees who might fill demanded functions and to locate areas where skills are missing.

# Chapter 3

# State-of-the-art

Computer applications focused on the finding and mapping of expertise may be more or less automatically driven. This chapter will present different approaches used to find an organizations expertise. This genre of computer applications goes by names such as people finders, expert locators, corporate yellow pages and Expertise finders. This report has been using, and will continue to use the latter to describe this kind of systems.

Expertise finders can be classified into three different categories [29]:

1. Database approach
2. Social networks approach
3. Expertise profile based on electronic evidence

## 3.1 Database approach

The database approach to finding expertise requires the storage of self assessed expertise descriptions. The employees themselves define their expertise areas and levels. These descriptions are stored in a database structure, and come with a user interface that enables other employees to search and/or browse for expertise.

Skills Manager [12] is one Expertise finder following this approach. Skills Manager is used by the knowledge management company Computas to locate employees with the demanded expertise. The expertise seekers may choose from 250 particular expertise areas in a taxonomy and further find what level of expertise the nominated experts possess, from ”expert” as the highest level to ”non relevant” as the lowest level of expertise. The employees are requested to evaluate their profile as new expertise areas are entered into taxonomy or when they have updated their expertise in one of the existing areas.

This kind of self assessment that is found in Skills Manager and other Expertise finders following the database approach is problematic because of several reasons. Some of them are:

- The difficulty of evaluating ones own expertise
- Difficult to validate others expertise
- Deliberate underestimation or overestimation of ones expertise
- Time consuming effort updating ones expertise profile

### Difficult to evaluate ones own expertise

It is difficult to compare ones expertise with others. Several Expertise finders in this genre operate with self assessments of competence levels [12], which is unreliable due to many reasons, e.g. what criterions can be applied to assess ones expertise; how do you estimate your level of expertise relatively to other employees' expertise? And, in many cases expertise contains a tacit element of knowledge that the expert himself is not aware of [31];[32].

### Difficult to validate others expertise

When projects are being manned it can be difficult for the project manager to decide whether the self assessed expertise descriptions are actually valid or not. It may also be too time costly to perform some objective validation of these profiles and difficult to normalize the results. In small or medium-sized companies where the employees know each other well, some of these problems may not be an issue, but in larger companies where the employees don't know each other, a validation is necessary to establish the credibility and expertise of the project members.

### Deliberate underestimation or overestimation of ones expertise

In cases where one knows that a characterization as an expert lead to a lot of extra work assignments, it might be tempting to avoid this work load by underestimating ones expertise level. Another danger in putting your expertise level high in a Expertise finder system is that you may be in danger of getting assigned to the same kind of projects over and over based on your extensive experience in this field, and missing the opportunity to get assigned on new, exciting project areas [11].

In other cases, to achieve prestige among the other employees it could be tempting to "upgrade" ones expertise. This may also be the case if you want to get involved with a project that requires a certain expertise area or level that you do not possess.

**Time consuming to update ones expertise profile**

If you need to update your expertise profile each time you gain new knowledge, this will be a time consuming process, and especially if the expert himself feel there is no reward for doing so, it's relatively optimistic of an organization to expect that all employees will update their profile at every new accumulation of knowledge. Maybe this approach is not in the best interest to the organization either. If all employees use one hour to formulate this weeks gained knowledge or competence this is quite a loss of resources that could be applied elsewhere.

### 3.1.1 An objective database approach

Instead of the employee assessing his own expertise, some organizations employ knowledge stewards or expertise supervisors to gather expertise information. These knowledge stewards perform interviews of the employees, analyze expertise areas and levels, and insert this data into the Expertise finder database. In this way there is an objective element in the expertise localization stage that can verify the truthfulness of the expertise descriptions and transfer some of the workload and time effort from the employees to these stewards designated for this job. The knowledge stewards need to fulfil some important criteria because extracting expertise is a difficult task. Karhu et al [20] have developed a framework that describes the necessary steps to perform the knowledge steward process (figure 3.1).

This figure illustrates that the first objective is to create confidence between the expert and the knowledge steward. Then interviews of the employee are held and the knowledge steward performs a socialization[1] and develops his own mental models of the dialogue in a way that makes him able to describe this expertise explicitly. When the expertise seekers get access to the explicitly described expertise, another interpretation process is carried out where the seekers create their mental models of this explicit expertise and internalize it.

## 3.2 Social network approach

Social networks mean in this context an electronic network of nodes, often depicted as graph structures, where the nodes are the information handlers (sender

[1] See chapter 2.1.1

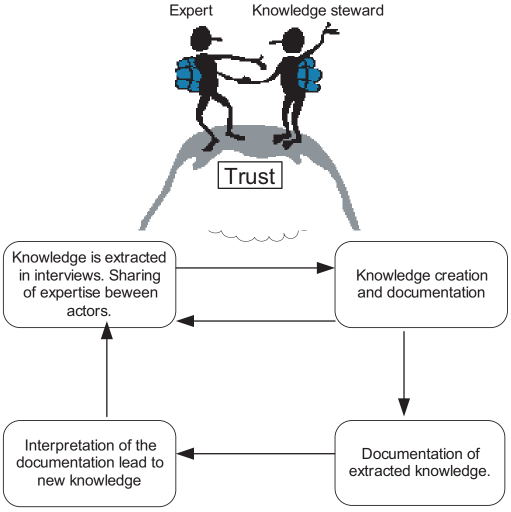

Expert
Knowledge steward
Trust
Knowledge is extracted in interviews. Sharing of expertise beween actors.
Knowledge creation and documentation
Interpretation of the documentation lead to new knowledge
Documentation of extracted knowledge.

Figure 3.1: Knowledge Steward

or receiver) and the edges between the nodes show the relationship and communication between the nodes. The nodes within a network have something they want to share with others. This might include everything from spare time interests (e.g. facebook.com) or in our case expertise. The main driver is that shared communication also means shared interests which again mean shared areas of expertise. This has two implications; by contacting a node with the wanted expertise you may satisfy your need for expertise directly from the node you have contacted, or you can meet a node that do not possess the needed expertise, but knows (has the expert in his social network) the node that is able to provide you with the needed wanted expertise.

Campbell et al. [6] describe how email communication patterns can be used to map social networks and identify what people know of who knows what (who's the expert). Analysis of email traffic show who communicates with each other and what kind of information that is sent. The underlying thought is that people often send email about some subject to the person who is knowledgeable about this subject and this person receive more emails about this subject than everyone else (figure 3.2).

This system works in three stages:

1. Construct clusters of all emails concerning a certain topic.

2. Analyze emails between each sender and receiver to see who is sending information to whom and construct a directed graph that show the information flow between them.

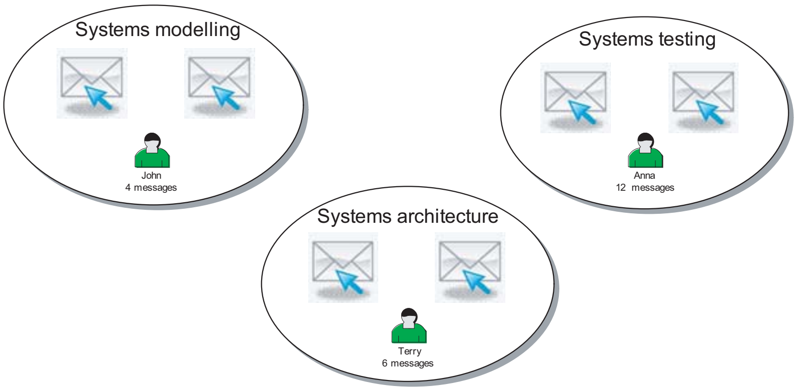


Figure 3.2: Using email communication patterns in expertise finding

3. Analyze the graph to provide a rating for all senders and receivers.

Campbell et al. compared this method with another approach that only mined the email messages without the analysis of communication patterns. The evaluation showed that the approach involving communication patterns achieved the best retrieval results with a precision of 67 percent at 33 percent recall.

Obviously, there are some privacy issues related to Campbell et al.'s approach. Analyzing the employees' professional and private emails is a questionable approach to summarize an organizations expertise.

ReferralWeb [21] initially also used email communication to map social networks in the localization of expertise, but dropped this idea because of the privacy issues involved. Instead, this system use content analysis and social networks to extract adjacent name occurrences from different kind of online sources as an indication of relationship. These name occurrences are found in links in homepages, lists of co-authors in technical papers and citations of papers, exchanges between individuals recorded in news archives and organization charts. The ReferralWeb Expertise finder works as follows:

1. When a user is registered in the ReferralWeb a search engine is used to find documents related to this user. In addition to this are other employees co-occurring in the documents extracted as well. This process is applied recursively in one or two levels, and the results are then entered into a global network model.

2. In the search stage one can - based on the global network model - search for a topic and filter this search based on social criteria, such as "Which

person, who is a colleague of mine, is an expert in building ontologies?"

The Expert Locator prototype developed and described by D'Amore [9] use a variety of sources to find expertise evidence. Project spaces, formal organization spaces and ListServ discussion threads are used to collect expertise evidence based on the assumption that people belonging to these activity spaces actually have knowledge of the topic(s) discussed, and that evidence of expertise is found based on that people signalize their expertise in these spaces' entities (e.g. people), events (e.g. labor) and artifacts (e.g. reports). By signaling these expertise evidence, Expert Locator tries to locate the social context where people communicate within, for instance in discussion threads, and hence deduce communities of practice.

## 3.3 Expertise profile based on electronic evidence

Numerous approaches have been used to automatically construct expertise profiles from available electronic sources. Some approaches are domain specific and some more general. An example of a domain specific approach is Expertise Recommender [27]. Expertise Recommender is a product of a field study carried out in a software company. Expertise Recommender locates experts based on two important rules (heuristics): A "change history rule" that locates information about the programmer who last modified a system module, and a "technical support rule" that uses information about the person who last solved a technical problem. If the need for expertise is related to the system module, the programmer who have made modifications to the system module is regarded as the expert, and if the need for expertise is related to a technical problem the person who solved a similar problem is regarded as the expert. When an expertise seeker issues a query in the Expertise Recommender he may filter the potential experts based on his social network. In this way he is able to choose from expert candidates he is familiar with.

Several other Expertise finders belonging to this category utilize different document storages as a source of expertise evidence. One example is the P@noptic expert finder. This system stores an employee-document (essentially an expertise profile) for each employee in the organization and only these employee-documents are indexed. The employee-document consists of the employees' contact information and concatenated text from all documents mentioning this employee on the intranet and the employees' home page(s). When the results from the expertise query are presented, the top ranked expert is presented with picture, contact information and the documents he is mentioned in, while the other retrieved experts are presented with basic contact information and a link to supporting documents.

Balog et.al [4] tried a probabilistic approach to expert finding. They developed, tested and evaluated two different models using heterogeneous test data from the TREC collection. In the first model all term information from all the documents associated with the expert candidate is collected and then used to represent the candidate. Hence, this model predicts how probable the query topic is to rank the different candidates, i.e. finding expertise based on a candidate view. The second model ranked documents according to the query, and then they determined how likely a candidate was an expert by considering the documents associated with these candidates, i.e. finding expertise based on a document view. One of the evaluation measures used to test these models was R-precision. The evaluation showed that the second model performed better, with an R-precision score of 23.3 percent.

# Chapter 4

# Information retrieval

This chapter is devoted to the field of information retrieval (IR). Expert finder systems often use techniques found in the IR domain in the same manner as other IR systems. After all it is usually plain text that contain the expertise evidence, no matter if the system is a web search engine or an expert finder system. The only thing that usually seperates a general IR system with an expert finder, is the fact that experts found and presented in an expert finder are products of different kind of information, often collected from different sources. You might find the actual expertise evidence (e.g. a name) in one document collection, the contact information in another (e.g. telephone number and e-mail address), and some information that makes it possible for the expert searcher to validate the expert in a third source (e.g. a reference to other texts mentioning the expert). Hence, most of the principles and techniques used are similar.

This chapter begins with a description of the search process. Then some of the techniques used in indexing and pre-processing of documents will be explained. Further some of the main building blocks in IR, the IR models, will be described before this chapter ends with a description of evaluation measures used in expert finder systems as well as in IR systems.

## 4.1 The search process

The search process in an IR system starts with a user issuing a query. The query is processed in a way that enables the system to compare the query with the documents residing in the collection. The collections' documents are usually pre-processed and indexed so that the retrieval process is executed as efficiently as possible. This kind of processing is usually performed offline because of performance issues. In the presentation stage the processed query is compared to the indexed documents, and the most relevant documents are retrieved and presented to the user issuing the query (Figure 4.1).

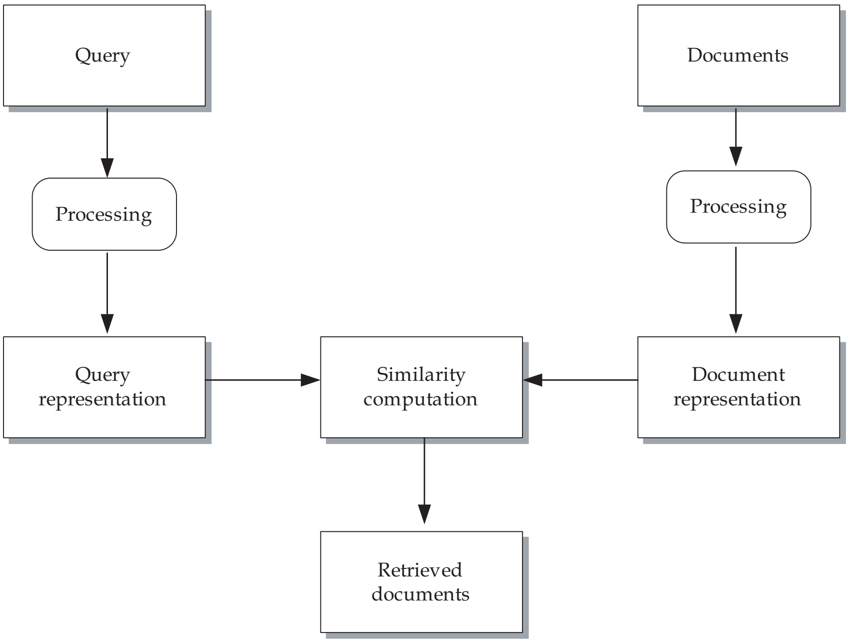


Figure 4.1: The search process

## 4.2 Pre-processing

The purpose with pre-processing is to compress the text before it gets indexed and to improve the systems' relevance judgement. Usually both the user query and the documents in the collection are pre-processed. In this report, two techniques for text compression are described, stop word removal and stemming.

### 4.2.1 Stop word removal

Words that occur too often in the documents in the collection don't have a discriminating effect and should be removed from the text to ensure they are not considered as index terms. A document consists of several words that are not important to the retrieval because they occur to often. Because of this it is common practice to pre-process the documents so that only words that can distinguish documents are taken under consideration in the similarity stage, while the words that occur most frequent are considered stop words and are filtered out prior to the indexing stage. Common stop words are verbs, adverbs, adjectives, prepositions and articles. The stopword removal stage should be carried out with caution. Removing words that might discriminate relevant from non-relevant information could effect the systems performance considerably.

### 4.2.2 Stemming

Stemming is a pre-processing technique that reduces a word to its stem. By doing so it's possible to retrieve words with other inflections than the exact word in the query. The most common approach to stemming is the affix removal technique [3]. This technique removes the ending of the word based on rules in the stemming algorithm, and only the stem is stored in the index. For instance you may issue the query 'process AND technique' and also retrieve documents that mention 'processing techniques' in the text. There is an ongoing debate whether stemming actually improves the retrieval performance, and many search engines do not use this technique because of the inconclusive benefits of stemming [3].

## 4.3 Indexing

Indexing of a document collection implies the construction of data structures that creates a compact version of every document in the collection. By utilizing data structures one avoids a sequential search through all text in the documents. This ensures high performance both with regard to time and storage issues. There are several alternative techniques for indexing, but the ones most used are the inverted file, suffix table and signature file. The inverted file technique is suitable for most applications [3] and will be explained briefly in this report.

An inverted file structure is based on two main components, 1) a vocabulary and 2) a list of occurrences. The vocabulary consists of a list of the different words in the document, while the list of occurrences connects the words in the vocabulary with their position in the document. An example of an inverted file is illustrated in figure 4.2.

## 4.4 IR models

There are two main challenges that need attention when an IR system is to be developed. The first challenge concerns the document representation and the second challenge concerns how to compare similarities between the collections'

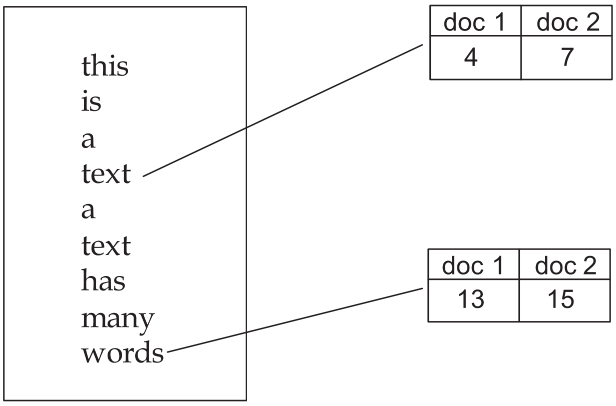


Figure 4.2: Inverted index

documents and the query issued by the user. According to [3], an IR model can be characterized as follows:

An IR-modell is a quadruple $[\mathbf{D}, \mathbf{Q}, \mathcal{F}, R(q_i, d_j)]$ where

1. **D** is a set composed of logical views (or representations) of the documents in the collection.

2. **Q** is a set composed of logical views (or representations) of the user information needs. Such representions are called queries.

3. $\mathcal{F}$ is a framework for modelling document representations, queries, and their relationships.

4. $R(q_i, d_j)$ is a ranking function which associates a real number with a query $q_i \in \mathbf{Q}$ and a document representation $d_j \in \mathbf{D}$. Such a ranking defines an ordering among the documents with regard to the query $q_i$.

When building an IR model the first thing to consider is how the documents are represented (usually by index terms) and how to represent the users' information needs (usually given as a query). Given these representations or logical views, the framework used in modelling the representations is to be decided, and also how the ranking functionality should be carried out. There are essentially three retrieval models to consider at this stage, the Boolean model, the Vector space model and the Probability model [3]. The Boolean Model and the Vector Space Model will be explained in the following as these are the two model relevant for this thesis.

### 4.4.1 Boolean model

This IR model is based on set theory and Boolean algebra. This models main advantage is that it is relatively easy to comprehend as it is based on the fact that the index term either exists in a document or it does not. This is also one of the models' drawbacks as it means there is no ranking of documents - the document is relevant or it is non-relevant as this formula shows:

$$sim(d_j, q) = \left\{\begin{matrix} 1\ if\ \exists \vec{q}_{cc} \mid (\vec{q}_{cc} \in \vec{q}_{dnf}) \wedge (\forall_{k_i}, g_i(\vec{d}_j) = g_i(\vec{q}_{cc})) \\ 0\ otherwise \end{matrix}\right\}\ where$$

$$\begin{aligned} sim(d_j, q) &= \text{The similarity between the document } d_j \text{ and the query } q \\ \vec{q}_{dnf} &= \text{The disjunctive normal form for the query q} \\ \vec{q}_{cc} &= \text{Any of the conjunctive components of } q_{dnf} \end{aligned}$$

*if $sim(d_j, q)$ = 1 then the Boolean model predicts that the document $d_j$ is relevant to the query q. Otherwise, the prediction is that the document is not relevant.*

Another disadvantage following from this is that it does not reduce the information space and thus break one of IRs' main postulates; to reduce the amount of information in the information seeking process. The Boolean models' principle is an exact binary match which often results in either too many or too few hits given a query.

### 4.4.2 Vector Space model

Salton et al introduced the Vector space model in 1975[1]. While the Boolean model has the disadvantage that it uses only binary comparison between terms in the query and the documents, the Vector space model recognizes that a binary comparison between terms in a query and documents not is sufficient and uses non-binary weighting of index terms. The weighting of index terms is made possible by the Term Frequency (TF) - Inverse Document Frequency (IDF) algorithm. The Term frequency measures how well an index term describes the document content by assigning high weight to terms that frequently occur in a document. The Inverse document frequency on the other hand measures how well the index term discriminates between relevant and non-relevant documents in the collection by giving high weights to rare terms. The term weighs are further used to compute the degree of similarity between each document stored in the system

[1] Already in 1968 Salton published an article about the IR system SMART that used similar principles in IR

and the users' query. This degree of similarity can be seen in the vector space as the cosine value on the angle between the query vector and the document vector. By sorting the retrieved documents in descending order based on the degree of similarity the Vector space model considers documents that only partly match the query [3].

$$sim(d_j, q) = \frac{\vec{d_j} \bullet \vec{q}}{|\vec{d_j}| \times |\vec{q}|} = \frac{\sum_{i=1}^{t} w_{i},_{j} \times w_{i},_{q}}{\sqrt{\sum_{i=1}^{t} w_{i}^{2},_{j}} \times \sqrt{\sum_{i=1}^{t} w_{i}^{2},_{q}}}, where \tag{4.1}$$

| | | |
|---|---|---|
| $sim(d_j, q)$ | = | correlation between $d_j$ and $q$ |
| $\vec{d_j}$ | = | the vector to the document j |
| $\vec{q}$ | = | the query vector |
| $t$ | = | total number of indexterms in the collection |
| $w_{i},_{j}$ | = | the weight to the term in document j |
| $w_{i},_{q}$ | = | the weight to the term in the query |

The advantages you get by using the Vector retrieval model is first and foremost that you by using a non-binary weighting scheme on the index terms get a good retrieval performance. And by using the cosine ranking the user issuing the query gets presented with relatively similar documents because of the partial match strategy and results ranked by relevance.

The disadvantages of the Vector space model are that the model does not take into account the relations between terms [25] and that is allegedly only works optimally with short documents because it's difficult to compute similarity measurements based on vectors on long documents [15].

## 4.5 Evaluation of IR systems

During the evaluation of IR systems the focus is on confirming that the relevant documents are retrieved and the non-relevant ones are not. The evaluation metrics often used are recall and precision. A high level of recall denotes that the relevant documents in the collection are retrieved and presented for the user, while a high precision denotes that the documents retrieved and presented actually are relevant. In the same manner recall and precision denotes relevant experts in an expert finder. A high level of recall shows that all the relevant experts are retrieved and a high level of precision shows that the novices and non-experts are not retrieved. The formulas for recall and precision are:

$Recall = \frac{|Ra|}{|R|}$ where |Ra| are the retrieved relevant documents (experts) and |R| denotes the total set of relevant documents (experts)

$Precision = \frac{|Ra|}{|A|}$ where |Ra| are the retrieved relevant documents (experts) and |A| are the retrieved documents (experts) presented to the user.

In practice, it is normal that the higher the recall, the lower the precision is, and vice versa. This is because with high recall the system gathers a large portion of the collection, and some non-relevant documents are also retrieved, with the consequence that the precision is being reduced ([25]).

When evaluating Expertise finders, key personnel are often inquired to manually separate the experts from non-experts in their organization. To illustrate this with an example:

*Tone is the personnel manager in a software development company. She knows the organizations' employees well and can pinpoint who has knowledge about UML modelling. She also knows that the expert in this subject is Lasse. Given the query "use case modelling" in the company's expertise locator, Tone is given the task to evaluate the performance of the system. She immediately recognizes that Lasse is lacking in the retrieved results. This shows that the system has low recall. She also notice that Gabriel - the company accountant and not very knowledgeable about UML- appear on the list. This is a clear indication of low precision in the Expertise finder.*

### 4.5.1 Precision at recall levels

Within the field of expertise finding, precision is often regarded as a better measure than recall [28]. When you look for an expert, you are mainly interested in finding the one expert with the highest level of expertise, not a list of ten potential experts who may or may not possess the needed expertise. Also, because knowing exactly all relevant experts to a query is difficult, it is common to compute the precision at a given recall level. This procedure is illustrated by the following example:

Say your human evaluators have specified these five persons as experts to a given query:

$E_q = \{p_1, p_2, p_3, p_4, p_5\}$ where $E_q$ is the set containing the total number of relevant experts, and $p_n$ are the relevant experts found by the evaluators.

The expert finder system computes this ranked list of experts:

1. $p_1$
2. $p_3$
3. $p_8$
4. $p_{16}$
5. $p_{12}$

Since $p_1$ amount to 20 percent of all relevant experts ($E_q$), the recall is 20 percent and since this is the first hit on the list the precision is 100 percent. So we say that the system on this query has 100 percent precision at 20 percent recall.

Further, $p_8$ has a precision of approximately 66 percent (Two out of three experts are relevant) at 40 percent recall, etc.

### 4.5.2 Interpolated average precision

To find a single summary representative for all queries in a test, it is common to average the single query results into an average precision value. This is achieved by averaging the precision at each recall level as follows:

$\overline{P}(r) = \Sigma_{i=1}^{N_q} \frac{P_i(r)}{N_q}$ where

$\overline{P}(r)$ = average precision at the recall level r,
$N_q$ = the number of queries used,
$P_i(r)$ = the precision at recall level r for the i-th query.

This is usually done with computing the precision at eleven recall levels (0-100 percent), and by interpolating the results so that the interpolated precision at the $j$-th standard recall level is the maximum known precision at any recall level between the $j$-th recall level and the ($j$+1)-th recall level.

$P(r_j) = \max r_j \leq r \leq r_{j+1} P(r)$

This means that the precision at 0, 10, 20 and 30 percent recall level is interpolated to 100 percent precision in our last example.

### 4.5.3 R-precision

R-precision considers the size of the set of relevant experts ($E_q$) when computing the precision. If this set of assumed experts for instance consists of five experts, this means that the precision amongst the five first hits is used to compute R-precision.

E.g. if $E_q = \{p_1, p_2, p_3, p_4, p_5\}$ , and the results computed by the expert finder system presents ten experts, we take the first five proposed experts, and use these to compute R-precision. And, if three of the experts in $E_q$ occur in the results, then we have an R-precision of 0.6 (or sixty percent).

# Part II

# Case study

# Chapter 5

# Urban Development Area

## 5.1 The case

The Urban Development area in the municipality of Trondheim is an important premise provider for the future development and day-to-day maintenance in Trondheim. Urban Development is responsible for areas such as city regulation, building applications, housing administration, environmental issues, fire and rescue services, and general maintenance in the municipality. There are a total of 1100 people employed in eleven different units at Urban Development. This study focuses on six of the eleven units. The reason for doing this excerpt is that these six units are the units that essentially are the decision-making authorities while the other five units are mainly executing or advisory units. The involved units are (Norwegian names in parenthesis):

- Urban zoning office (Byplankontoret)
- Building permits office (Byggesakskontoret)
- Environment Office (Miljøenheten)
- Trondheim Real Estate (Trondheim eiendom)
- Map and surveying office (Kart- og oppmålingskontoret)
- Department of Infrastructure, Environment and Property Management (Trondheim byteknikk)

The five units not contributing to this study are City maintenance (Trondheim bydrift), Chief Municipal Treasurer (Byantikvaren), Chimney Sweeping Service

(Feiervesenet), the Fire and rescue service (Brann- og redningstjenesten) and the Housing Office (Boligenheten).

The overall structure of the municipal organization including the Urban Development area is presented below (5.1) (units contributing to the case study in grey):

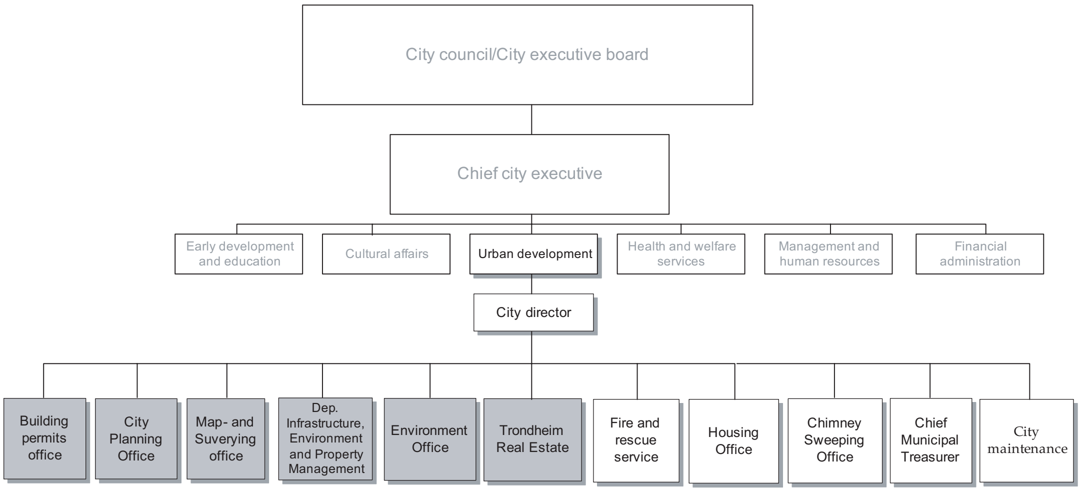


Figure 5.1: Urban Development organization chart

All these municipal units are normally accessed through a public reception called the City reception (Bytorget). This reception handles incoming calls and manual attendance, and distributes these inquiries to the correct destination. Since the Building Permits office, the Map- and Surveying office, and Urban zoning office usually receive more inquiries from the public than the other units, a seperate reception is organized for these three units.

## 5.2 Objectives

This study was carried out to find answers in two main areas. Firstly, the study needed to provide an understanding of Urban Developments responsibilities, internal- and external communication and the different kinds of knowledge, competencies and expertise the units possess. These issues are defined as ***organizational issues***:

- Get a picture of the overall organizational structure
- Find the units' main responsibilities
- Analyze the units' possessed expertise

- Analyze the communication flow (internally and externally) in the units and what kind of expertise that is inquired
- Find the internal and external conditions (Customers, laws and legislations etc.)

Secondly, there was a need to find out what the technical possibilities and terms were. It was important to discover what sources contain potential evidence of expertise, and how to access them. These are defined as ***technical issues***:

- Locate the sources of expertise evidence and figure out how to access these sources
- Find out how to best perform a mapping between the located expertise evidence and the experts
- Explore how to get access to relevant information sources used in the expert presentation (Contact info)

## 5.3 Method

The case study started autumn 2006 with a direct observation period at the City reception. As I at the time was employed at the Urban Development area, I had the opportunity to get direct access to incoming inquiries and how they were treated at their destination. Notes were taken on what inquiries the different units received, how they were processed and distributed, and in some cases what the outcome of the inquiries were.

During the first months of 2007 the interviewing and the other meetings involved in the case study were planned and executed. A total of eleven interviews with the involved units' managers and formal meetings (appointed meetings) with other relevant contributors were held, and also some informal ones (Mainly technical). The structure of the interviews can be characterized as semi-structured as the main topics were predefined, but there was no rigorous questionnaire guiding the interviews. The formal meetings followed more or less the same procedure. All interviews were recorded and transcribed immediately after the interviews were held. When there were inconsistencies found in the transcription stage, these were resolved either by issuing a mail or a telephone call to the unit manager being interviewed.

The same unit managers were also asked to provide sample queries and experts they saw fit to these queries. These queries were used in the evaluation of the expert finder prototype (Described in chapter eigth).

## 5.4 Presentation of the involved units

In this section the different units will be presented. The presentation includes a presentation of the main responsibilities and tasks the unit performs, the positions people are employed in (narrowed down to the five most employed positions), what kind of inquiries the unit receives, and where expertise evidence may be found.

### 5.4.1 Building permits office

This unit consists of 51 employees, where the majority have backgrounds within the fields of engineering, architecture and law (figure 5.2). Additionally, the Building permits office has a large administrative section performing tasks related to personnel and economy, handling in- and out correspondence, invoicing and the City reception.

When it comes to the units' main responsibility - the treatment of building applications - the case handlers are divided into two sections based on geographical location. The case handlers in one section treat building applications concerning the west side of the city, while the others are responsible for the east side. The case handlers have backgrounds mainly in architecture and engineering, but also other backgrounds may be present, e.g. social science.

In addition to the administrative tasks and the treatment of building applications, the unit has a legal department responsible for handling complaints, illegal building activities and this department also performs an advisory function in the treatment of building applications.

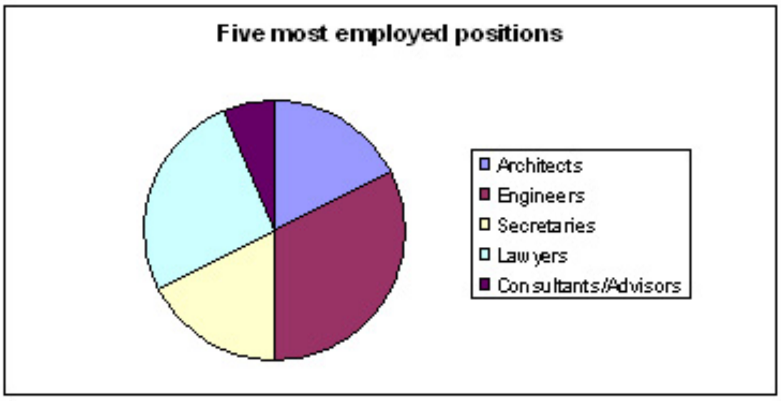


Figure 5.2: Building permits office

### Inquiries and communication

Together with the Urban zoning office and the Map- and surveying office, the Building permits office manages their own City reception that handles both incoming calls and manual attendance. Typical inquiries that concern the Building permits office are related to questions such as what case handler is treating a given building application, what one is allowed to do with ones' house, how the building application is formulated, and how the whole building application process proceeds.

Sometimes the citizens contact the case handlers directly, especially if they are a part of an ongoing building application. In this case the applicant may have received some kind of document containing contact information such as email address or telephone number from the case handler. In most cases, the citizens contact Bytorget, either by telephone, mail, or by attending at the service desk en persona.

### Expertise evidence

As a bureaucratic authority, the main information element and the obvious source containing expertise evidence are the documents produced by the case handlers. The unit mainly produce statutes as a response to building applications. In addition there often is a lot of in- and out correspondence between the applicant and the case handler. For instance, if the application is missing some important documentation, the case handler sends a letter describing what kind of documentation that is missing. If the treatment of the building application results in a refusal, the applicant may issue a complaint. In that case the legal department gets involved and one of the offices' lawyers decides whether the complaint contains some documentation that may alter the case handlers' decision. If not, the complaint is forwarded to the County administration. The County administrations' decision is final, so if this instance upholds the case handlers and lawyers decision, the refusal is final. If the County administrations' decision is contrary to the decision of the case handler and the lawyer, the building process may start. During this complaint-process a lot of correspondence may occur, mainly between the responsible lawyer and the applicant. When the building process is finished and the house is ready to move into, the applicant need to receive a certificate of completion from the case handler before he can move into the house.

Although some of the tasks performed in this unit require tacit knowledge, a lot of the knowledge is externalized into rules, procedures and documents. It therefore seems natural to investigate the document repositories as a source for expertise evidence. Nevertheless, many of the decisions a case handler has to make often require experience. For instance, large building projects are often distributed to the more experienced case handlers, whereas small, simpler projects such as applications regarding sheds or garages are often distributed to case handlers with less experience.

All statutes and all correspondence taken place at the Building permits office is archived in K2000.

### 5.4.2 Urban zoning office

The Urban zoning office employs 51 persons divided into areas such as administration, case handlers and lawyers. This units main responsibilities covers planning and development of the city's physical environment through overall strategies aimed at Urban Development, area- and transport planning and formulating regulating plans at different levels according to the planning- and building law.

The case handlers are as in the Building permits office mostly engineers and architects, but there are also some social scientists among the case handlers (figure 5.3). As with the Building permits office, the knowledge application involves a mixture of explicit knowledge found in laws, rules and procedures on one hand, and the more inexplicable, tacit knowledge on the other hand. There are guidelines to be followed stated in the planning and building law and there are also spesific procedures to follow when performing case handling on incoming proposals to new- or revided area plans. But the hierachi among the experienced and not so experienced case handlers is quite evident, something that is disclosed in the employees' title (Architect, architect II and chief architect). This hierarchi denotes the experience and education of the case handlers and also the magnitude of the cases the case handlers are treating.

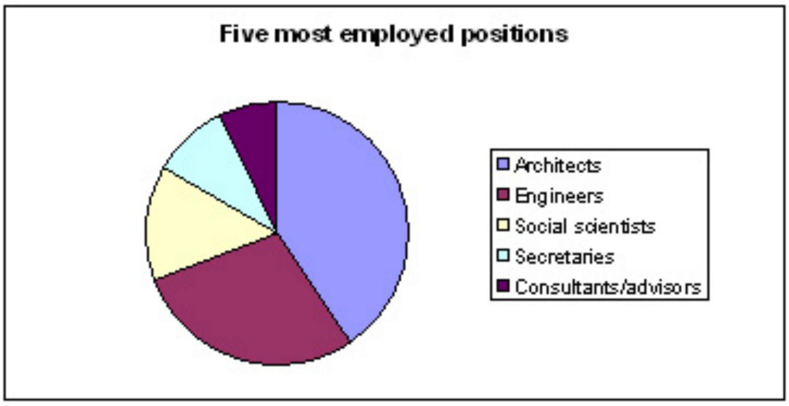


Figure 5.3: Urban zoning office

### Inquiries and communication

As with the Building permits office, the City reception handles much of the external inquiries - involving both incoming calls and manual attendance. Typical inquiries to the Urban zoning office concerns questions about the case handler responsible for e.g. area plans or ground separations, how the process involved in such cases proceeds and questions and complaints related to the regulating of traffic and accident prevention (traffic signs, speed bumps etc.).

The Urban zoning office collaborates with the Building permits office and set the premises on how houses and other building structures are both planned and built by forming regulation plans. Additionally, the Urban zoning office often collaborates with the City technique unit in designing and developing city areas which are currently un-developed and on traffic planning.

### Expertise evidence

This unit publishes documents such as area plans (regulation plans and the superior city plans), statutes related to ground separations, statutes related to traffic issues. Most documents follow the “municipal template structure” and contain all elements necessary to find an employee-to-document mapping such as the name of the case handler, a reference to the geographical location the actual action is taking place in, and a description of the documents’ content. As the Urban zoning unit often treats cases (e.g. regulation plans) concerning larger city areas, some documents contain references to place names rather than addresses. This might affect the expertise retrieval results when a query includes an address. All documents published by this unit are archived in K2000.

## 5.4.3 Map- and surveying office

This units’ main responsibilities concerns geo-referenced information. The Map- and surveying office is responsible for reference marks showing coordinate information, base maps, staking out grounds and property information such as street addressing and the administration of land- and holding number[1]. In addition to these services the unit is responsible for the maintenance and administration of the geo-referenced information in Trondheim, and the maintenance of the geo-referenced information systems in the municipality.

The Map- and surveying office consists of 39 employees organized into the areas map services, surveying assignments, counselling of limit adjustment, and counselling of property information. As figure 5.4 shows, a large percentage of the

[1] The land- and holding number is a unique combination of digits for each lot in the country

employees are engineers. The job descriptions in this unit are somewhat diffuse as there is a separation between engineers and case handlers although the engineers often perform case handler assignments and thus produce statutes and other kind of documentation archived in K2000.

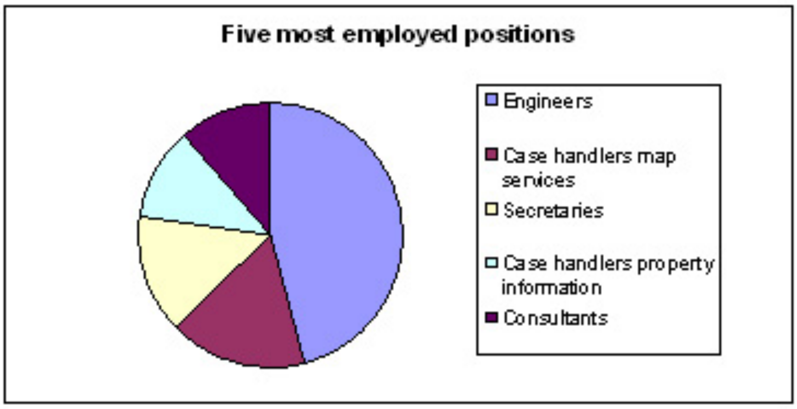


Figure 5.4: Map- and surveying office

### Inquiries and communication

The Map- and surveying office has a tight collaboration both with the Building permits office and the Urban zoning office. In the case of new building projects this unit is often engaged in plotting lots, determining building heights, and when applications on ground separations are to be treated, the Map- and surveying office together with the Urban zoning office is involved. City planning ensures that the regulations in the Planning- and Building law are protected, whereas the Map- and surveying office ensures that the regulations in the Act on Partition law are being attended to.

Typical inquiries to the Map- and surveying office involves questions concerning applications related to sectioning of property, requests for base- and property maps, basis for property taxes etc. Some of these inquiries are handled by the personnel at Bytorget, and some are forwarded to the correct employee at the Map- and surveying office.

**Expertise evidence**

Publications produced at this unit involve statutes related to sectioning, property taxation and ground adjustments (both separation and joining of grounds). As a division of this unit work with geo-reference information, including both manual and electronic maps, some documentation is not suitable for employee-to-document mapping. Most documents produced by the Map- and surveying office are archived in K2000.

### 5.4.4 Department of Infrastructure, Environment and Property Management

This unit consist of 63 employees divided into seven different subject groups. These groups are: Waste, Water and drainage, Geo technique, Road, Habitation and industry, a Legislative section and Green areas. Almost fifty percent of the employees at this unit are engineers (figure 5.5). In addition to the management there is only one person performing administrative duties, some employees dealing with geotechnical testing and the rest of the unit consist of people performing case handling.

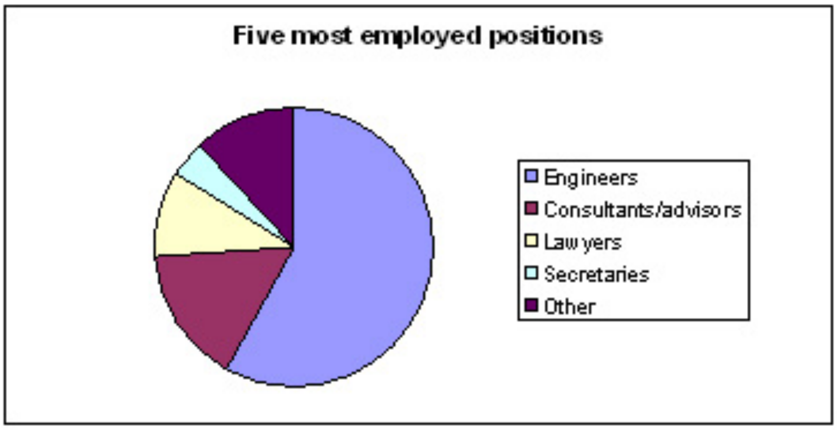


Figure 5.5: Department of Infrastructure, Environment and Property Management

The main responsibilities for this unit can be divided into three areas:

- **Task related to development and investment**: This area administrates a large investment project governing 300 MNOK. This project suggests what kind of services and development that should be focus areas in a budget year. A draft of this investment project is sent to the city council which eventually may decide that there will be built roads for 20 MNOK and that 10 MNOK will be earmarked for park maintenance etc. In the next stage the City technique unit invites tenders from contactors and such, and decides who will perform the actual work. In short, this area of the City technique unit governs and suggests how the city's funds should be invested.
- **Maintenance and administration of municipal services**: This area covers the infrastructure services such as water and drainage, city lights, and a responsibility for administering collective services such as parking and public transportation (bus). To maintain these services the City technique unit lays the framework for the different kinds of services, but other municipal or private units perform the actual work on behalf of City technique. Examples illustrating this are Team Trafikk who runs the city's bus service and City maintenance who perform the needed maintenance on roads.
- **Law management and regulations**: This area deals with the juridical aspects related to the area plans decided by the Urban zoning office, and management of municipal grounds. Some of the more detailed tasks of this area are acquisition of ground - both in terms of expropriation and purchase, contracts related to new building projects, administration of the municipal grounds, and legal matters concerning roads, water and drainage.

### Inquiries and communication

Typical external inquiries to this unit are related to ground acquisitions and sale of municipal property, maintenance of public roads (pedestrian crossings etc.) and leased property (in term of years or for special occasions such as the city's market taking place once a year).

### Expertise evidence

Even though this unit consist mainly of case handlers, a lot of the documents produced are drawings made by engineers and letters related to tenders. This kind of documentation - especially drawings produced in CAD format - is difficult to process in a search system. Other kinds of documents this unit produces are related to ground acquisition and sale, and documents related to investment projects. These usually contain a case handler name and should be a source for expertise evidence. For the most part the documents are archived in K2000.

### 5.4.5 Environment Office

The Environment Office consists of thirty-two employees divided into four subject groups. These groups are Environmental health, Environmental development, Nature management and outdoor life, and Agricultural management. Some of the units' main responsibilities are to provide a good childhood environment in schools and kinder gardens, contribute to environmental handling of contaminated ground, supervision of drinking water and being an advisory unit in fields such as radiation, noise, accident prevention, climate and energy.

The Environment Office consist of employees with very various backgrounds (see 5.6). Engineers, consultants (agronomists, forest technicians and other professions), medical personnel, nature managers and architects are some of the professions performing case handler assignments in this unit. The Environment Office also follow laws, rules and procedures in mostly the same manner as the other units, but since the Environment Office is more of a advisory unit, this may involve more tacit knowledge utilization than what the other case handling units are using.

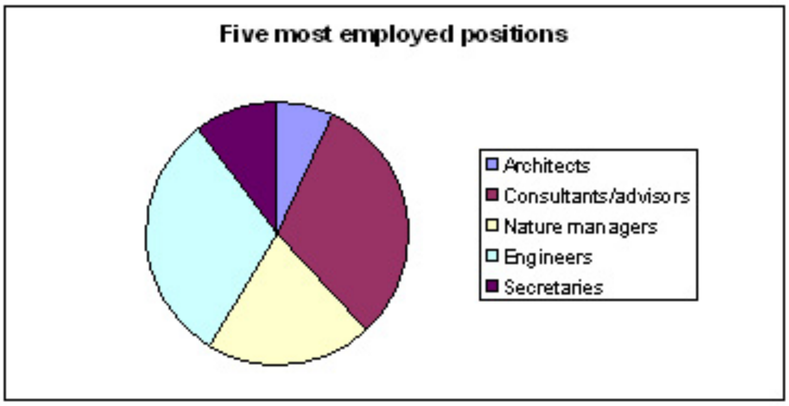


Figure 5.6: Environment Office

#### Inquiries and communication

The Environment Office perform a lot of internal tasks. They often perform advisory tasks on request from other municipal units. For instance, the Department of Infrastructure, Environment and Property Management unit provide economic funds to finance environmental reports written by the Environment Office.

### Expertise evidence

The documents published from this unit are statutes related to environmental issues, environment reports from the different subject groups, emergency plans at schools, legionella emergency plans and a lot of internal documents such as medical environment reports. The majority of the documents produced at this unit are archived in K2000, but some information is also archived in the units' own quality system.

## 5.4.6 Trondheim Real Estate

This unit is the largest unit in the municipality of Trondheim. It employs approximately 600 people divided into five areas:

- **Project and estate development**: This division is responsible for project management concerning investments in municipal buildings, new construction and rehabilitation on old buildings, development, purchase and sale of municipal buildings.
- **Management - Operation - Maintenance (FDV) on schools and kinder gardens**: This division is responsible for the management, operation and maintenance on the municipal schools and kinder gardens. This involves the maintenance of 320 000 square meters of school property and 36 000 square meters of kinder garden property.
- **Management - Operation - Maintenance on nursing homes, culture- and administration buildings**: There is a total of 217 000 square meters of nursing homes, culture- and administrations property that needs day-to-day management, operation and maintenance. This is the responsibility of this division.
- **Management - Operation - Maintenance on houses**: Trondheim Real Estate rents out 177 000 square meters of houses and apartments to citizens of Trondheim. These buildings require major or minor maintenance operations to uphold acceptable standards.
- **Environment services**: This division is responsible for the cleaning services in municipal buildings where municipal employees are located. It has been chosen to exclude all employees belonging to the environmental services division. Environmental services perform cleaning of municipal buildings and produce little or no documentation relevant for this study.

  As figure 5.7 illustrates a major part of the work force consists of maintenance personnel. There are also a lot of craftsmen employed in this unit.

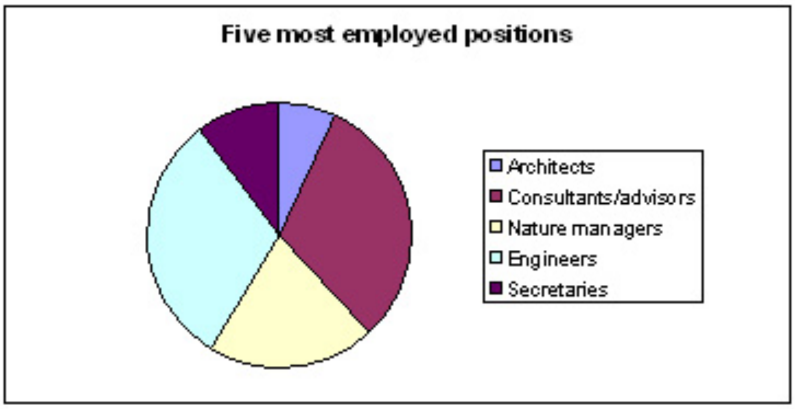


Figure 5.7: Trondheim Real Estate

### Inquiries and communication

Inquiries to this unit are usually sent by post, e-mail or telephone. The unit has its own reception that handles telephones and mail. Typical inquiries related to this unit are people reporting maintenance needs, problems regarding environmental issues in public schools and kinder gardens, and both complaints and offers related to building projects administered by this unit.

### Expertise evidence

This unit do not produce a lot of statutes like the other units. The documentation being produced at this unit is rental contracts, documentation related to building activities, and documentation related to the management and maintenance of the municipal buildings. However, these documents also usually contain the same information as the other units' documents, namely the name of the case handler, either an address or a place name, and a description of the documents' content. As the other units, Trondheim Real Estate use K2000 as its main document repository.

# Part III

# Prototype experiment

# Chapter 6

# Prototype design

This chapter presents the main principles and choices behind the thesis' prototype. The findings from the case study showed that a lot of the work carried out in the Urban development area involved the production of formal documents. These documents are found in the document storage related to the case handling system, K2000, and contained a quite evident mapping between the documents' content (described action, geographical location and document type) and the case handler who has published the document. Another finding from the case study was that much of the knowledge is externalized, meaning that the tacit knowledge possessed by the case handlers is translated into object based explicit knowledge materialized by the documents they produce. These findings suggests an approach based on the creation of expert profiles based on electronic evidence, where the electronic evidence is a) the name occurences of the case handler found in the documents and b) the documents' content.

A framework adapted from [33] is being used to describe this thesis' principles. This framework involves four stages:

1. Expertise extraction (The sources containing expertise evidence)

2. Expertise modelling (How expertise is defined, i.e. what separates experts from non-experts)

3. Expertise matching (How the relevant expertise is compared and ranked)

4. Expertise presentation (How the relevant experts are presented to the expertise seeker)

## 6.1 Prototype architecture

The above framework together with the resources found in the case study suggests this system architecture (figure 6.1). The architecture is explained from chapter 6.2 and on. An UML class diagram illustrating the relations between the modules in this architecture can be seen in appendix B.

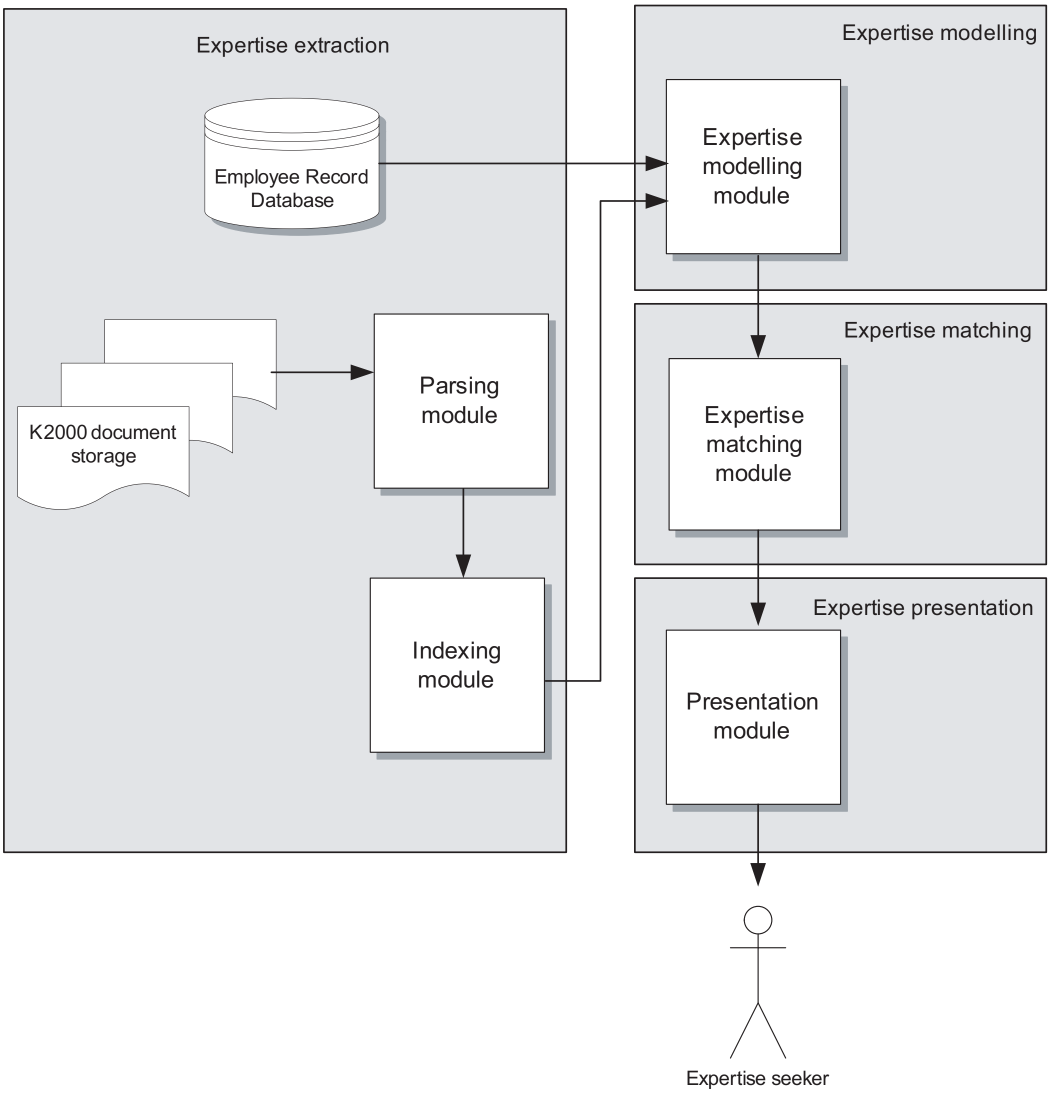


Figure 6.1: Prototype architecture

## 6.2 Expertise extraction

This part of the framework is responsible for providing the system with the sources for expertise evidence. The findings from the case study showed that one of the main tasks and responsibilities of the units is to handle incoming requests, applications and complaints. In response to these inquiries, the unit's case handlers process the inquiries and formulate statutes and other kinds of outgoing correspondence. Thus, a central source of expertise evidence is found in the case handler systems' (k2000) document storage, and this storage should be a main ingredient of the prototype.

To find experts we need employees. The employee database contain personnel information that may be coupled with experience, and by definition, it may be a source of expertise. Especially interesting is the field that specifies how long an employee has been in the current position. Another possibly interesting field is the employees' actual position. It matters - as a validation criterion - whether the person who is the owner of a document has the position secretary or for instance civil engineer, because secretaries do not perform case handler assignments but a civil engineer might.

### 6.2.1 K2000 document storage

All seven units involved in this study heavily use K2000 as their main repository of documents. K2000 is an information system/case handling system developed by IBM used by most municipalities in Norway. The K2000 structure is made up by cases at the superior level and every case contains one or more journals. Each journal consists of one or more documents (figure 6.2).

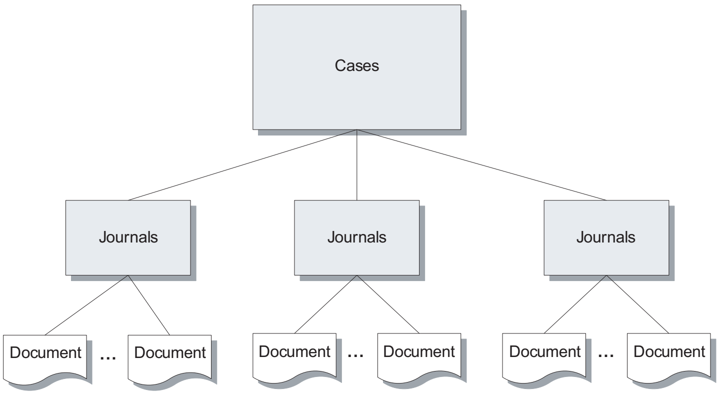
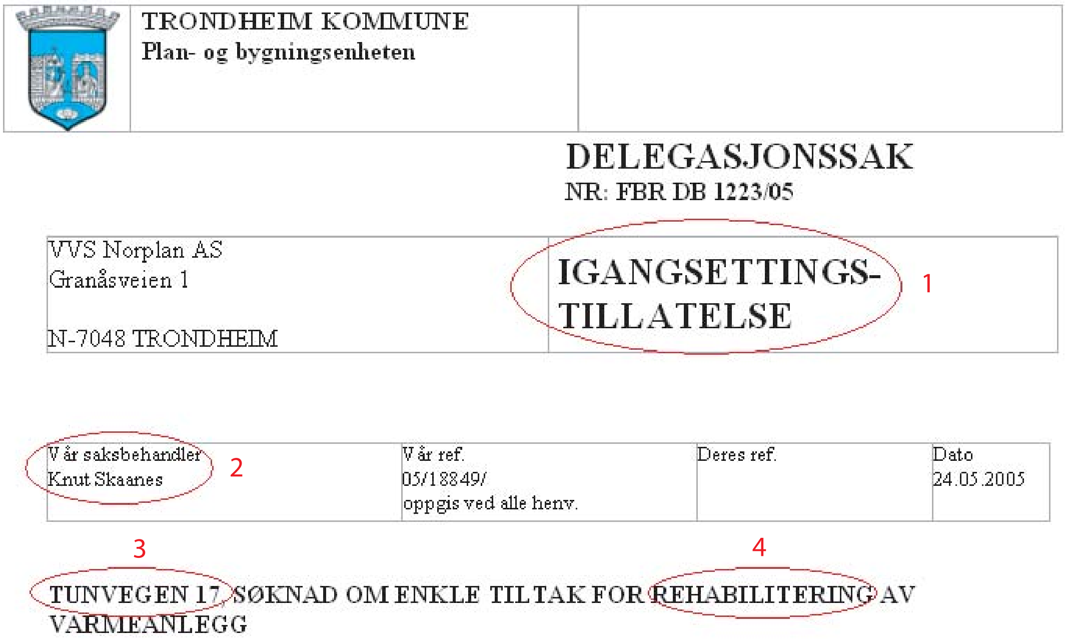


Figure 6.2: The K2000 structure

The documents include statutes, outgoing correspondence from counsellors to external parties (applicants or other relevant parties), incoming and internal correspondance and different kinds of drawings and images. For the prototype, the statutes and other outgoing correspondence are most relevant because these are the documents containing the most evident mapping between the document theme and case handler. The documents archived in K2000 use more or less the same structure decided by document templates.

A part of a typical document archived in K2000 is presented in figure 6.3. This particular document is a statute, but all outgoing correspondence is quite similar to this with regards to title fields, address fields and case handler name position.



TRONDHEIM KOMMUNE
Plan- og bygningsenheten

DELEGASJONSSAK
NR: FBR DB 1223/05

VVS Norplan AS
Granåsveien 1

N-7048 TRONDHEIM

IGANGSETTINGS-
TILLATELSE

| | | | |
|---|---|---|---|
| Vår saksbehandler<br>Knut Skaanes | Vår ref.<br>05/18849/<br>oppgis ved alle henv. | Deres ref. | Dato<br>24.05.2005 |

TUNVEGEN 17, SØKNAD OM ENKLE TILTAK FOR REHABILITERING AV VARMEANLEGG

Figure 6.3: Example K2000 document

There are a number of interesting areas in this document; it states what kind of document this is (1), it says who is the case planner (2), it shows the address the document concerns (3) and it shows what kind of action the document describes (4). In addition to this information, the statutes usually also contain a further description of the action applied for, the case handlers judgement of the action and a more thorough explanation of the statutes' outcome.

All this above mentioned document information should be preserved and made searchable in the prototype (not seen as stop words) as they all are relevant in a typical search situation. As described in chapter II, typical inquiries from the public often involves a geographical reference, a type of action, and a case handlers' name (e.g. who is the case handler working on the building application on the new house in Byåsveien 60).

The productivity (i.e. how many documents produced) of the case handler on a certain subject denotes what kind of experience the case handler has on this subject. A case handler who has published a statute or another outgoing document in a certain case is in most cases the case handler responsible for this actual case, and has a certain amount of expertise on this subject. This is an important aspect of this work. In addition, a case handler who is referenced in more documents on a subject than the other case handlers is the one who implicitly is the employee with highest expertise.

### 6.2.2 The employee database

The database table containing the employee records has the following fields (table 6.1):

| **Database field name** | **Explanation** |
|---|---|
| name | The employees name |
| employmentDate | The date the employee started in this position |
| email | The employees e-mail address |
| telephone | The employees telephone number |
| unit | The unit the employee belong to |
| position | What position the employee has |

Table 6.1: The employee database

Besides providing contact information, these employee records contain information that makes it possible to discriminate expertise by experience. The 'employment date' states how long the employee has been employed in this exact position and the 'position' field could make it possible to perform some validation of this employees' expertise. If the employees' position has the value 'secretary' it is likely that this particular employee has less expertise on the subject 'regulating plans on Byåsen' than employees with the position 'Civil architects'.

In this case this means a representation of the documents found in the K2000 document storage. These documents residing in K2000 are MS Word documents, and a parsing solution is necessary to translate them into plain text as required by the search engine. The parsed documents are then pre-processed to optimize the indexing and searching stage. This pre-processing is carried out by removing the most frequent stop words in the collection. After this the parsed and pre-processed documents are indexed to allow efficient searching by the prototypes' search engine. See figure 6.4.

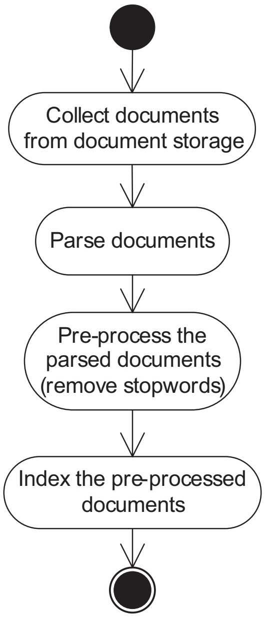

Collect documents from document storage
Parse documents
Pre-process the parsed documents (remove stopwords)
Index the pre-processed documents

Figure 6.4: Expertise extraction

## 6.3 Expertise modelling

The expertise modelling is concerned with how the expertise evidence found in the previous stage is used to separate the experts from the non-experts. In essence, this means to provide a solution that defines expertise. In our case this can be achieved by two different approaches; 1) either by utilizing the actual employee-to-document mapping (name occurrences) found in the K2000 document storage and stating that the more documents the employee is mentioned in, the more of an expert is he. Or, 2) by using the number of years the employee has been employed in his current position as an indication of experience and expertise. One presumption with the latter alternative is that the employee has some affiliation with the query, hence he must be mentioned at least in *one* of the relevant documents, meaning that a document $d$ in the document storage is associated with a expert candidate $ec$, if there is a non-zero association $a(d,ec) > 0$.

The name occurrences are found by utilizing a search engine which in the first stage searches through the entire collection and extracts the relevant documents, caches the relevant documents and then indexes these documents into a new index.

In the second stage, the prototype system searches for name occurrences in this index by performing a new search where all employees in the employee database acts as queries. The result of this search is a data structure consisting of a) employee names and b) a number of occurrences. See figure 6.5

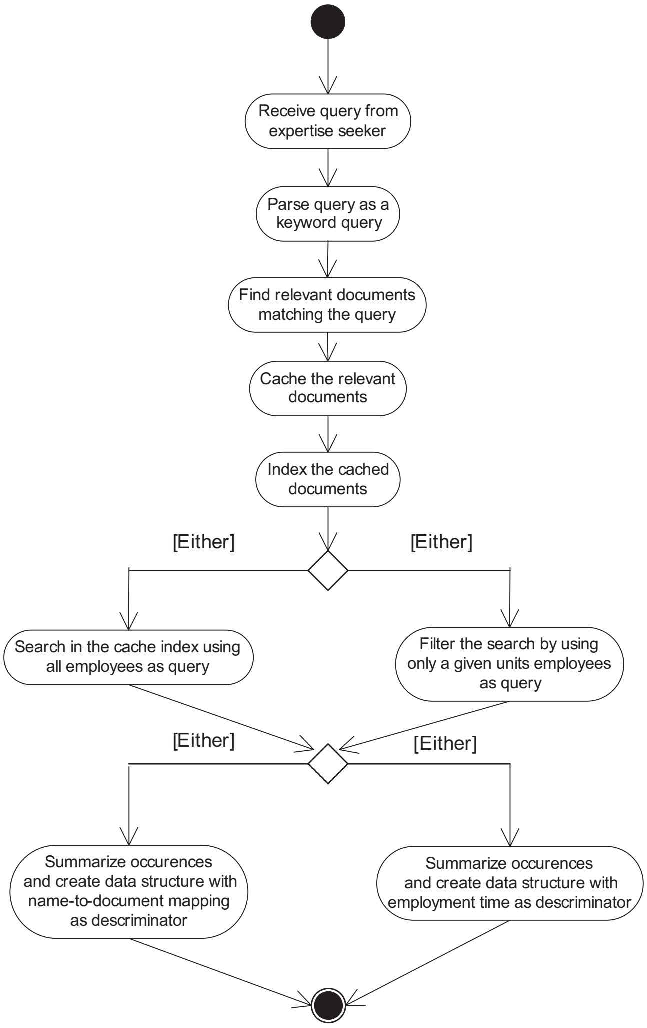

Receive query from expertise seeker
Parse query as a keyword query
Find relevant documents matching the query
Cache the relevant documents
Index the cached documents
[Either]
[Either]
Search in the cache index using all employees as query
Filter the search by using only a given units employees as query
[Either]
[Either]
Summarize occurences and create data structure with name-to-document mapping as descriminator
Summarize occurences and create data structure with employment time as descriminator

Figure 6.5: Expertise modelling

The second alternative, using the number of years the employee has been in his current position, starts similar as the name occurrences alternative. The search engine searches through the entire collection and extracts relevant documents given the expertise query issued by the expertise seeker, caches the relevant documents, and indexes these. Also, as in the first alternative, the system searches for name occurrences, as we need to establish that the nominated experts have some affiliation with the relevant documents. If they do, the employment time is the determinant that decides the ranking of the experts. The same data structures are used in this alternative, but the number of occurrences used in the first alternative is replaced with the number of years the employee has been employed in the current position.

## 6.4 Expertise matching

This stage deals with how to compare the nominated experts to each other, and how to determine the ranking in the list of experts that is presented to the expertise seeker.

In the query stage, the expertise seeker should be given an opportunity to filter the experts based on some proximity measure. In chapter three we saw that the Expertise Recommender [27] gave the expertise seeker an opportunity to limit his search within his social network. As this prototype is not based on social network structures, we won't use the same mechanism, but we can filter the resulting experts based on what unit they are employed in without any further analysis of the social structures. In this way the employee seeker may receive a ranked list of experts who he might be familiar with within his own unit.

The number of times an employee has been mentioned in the relevant documents or the number of years the employee has been employed decides his rank in the final list presented to the expertise seeker.

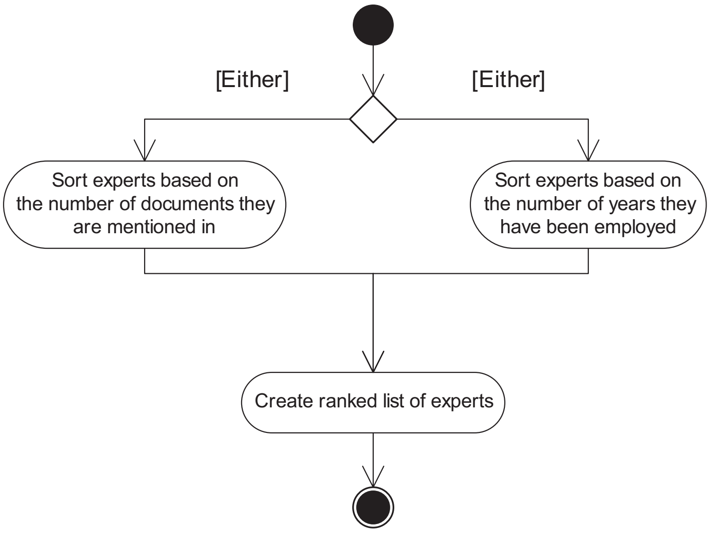

[Either]
[Either]
Sort experts based on the number of documents they are mentioned in
Sort experts based on the number of years they have been employed
Create ranked list of experts

Figure 6.6: Expertise matching

## 6.5 Expertise presentation

The expertise presentation stage concerns how the resulting list of experts is presented to the expertise seeker. This list should satisfy two criterions. Firstly, the returned list of experts should provide information on how the expert can be reached. This means providing contact information such as the experts' name, e-mail address, telephone number, and where the expert is located both physically (e.g. office or floor) and logically (in the organizational structure). Secondly, the expertise seeker should be given information that gives him an opportunity to validate the experts. This validation may be realized by letting the expertise seeker view some of the expertise evidence used by the system to define the experts. In our case this means that the system should not just present the experts in descending order with contact information, but also present the number of mentions the expert has in the relevant documents, the position of the expert (see chapter 7.1), and a link to the documents mentioning the expert. See figure 6.7.

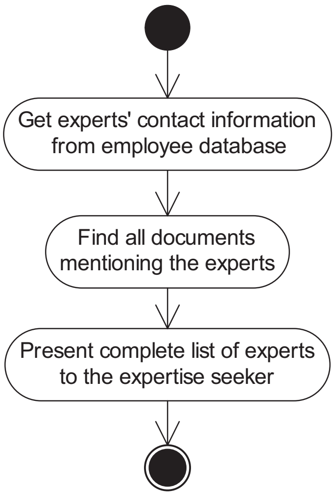

Get experts' contact information from employee database
Find all documents mentioning the experts
Present complete list of experts to the expertise seeker

Figure 6.7: Expertise presentation

# Chapter 7

# Prototype development

This chapter describes the actual implementation of the Expertise finder prototype. The framework presented in the previous chapter is in this chapter used to describe the development of the prototype.

## 7.1 Expertise extraction

The expertise extraction concerns the sources where expertise may be revealed. In this prototype the focus is on the document storages housing documents published in K2000. Before the actual expertise extraction may take place, the documents in the document storages have to be parsed into the required format that the search engine uses, they have to be pre-processed, meaning that stop words have to be removed, and they have to be indexed. The search engine is an adaptation of the open source search engine Lucene.

### 7.1.1 The K2000 document storage

A total of approximately 2200 statutes and other outgoing letters from 2005 and 2006 constitute the test case from K2000. The documents are evenly divided among the six participating units as table 7.1 shows. This table shows the number of documents extracted per unit and in parenthesis the total number of documents published in each unit. As the table illustrates, the total number of documents produced at the different units varies substantially.

All documents extracted from K2000 are in Microsoft Word format. Word documents are not easily parsed as this format is proprietary. However, the POI

| **Unit** | **2005** | **2006** |
|---|---|---|
| Urban zoning office | 200 (711) | 200 (1500) |
| Dep. of infrastructure... | 200 (2622) | 200 (2595) |
| Trondheim Real estate | 200 (2386) | 200 (2474) |
| Building permits office | 200 (3400) | 200 (3221) |
| Map and surveying office | 200 (906) | 200 (959) |
| Environment office | 106 (117) | 107 (117) |

Table 7.1: Document extraction from K2000

project[1], an Apache Jakarta project has some open source solutions which make it possible for others to manipulate - and in this case parse - documents using Microsoft formats. This parsing is necessary because to be able to search through these documents with Lucene, the initial Word documents have to be parsed into plain text files.

### 7.1.2 The employee database

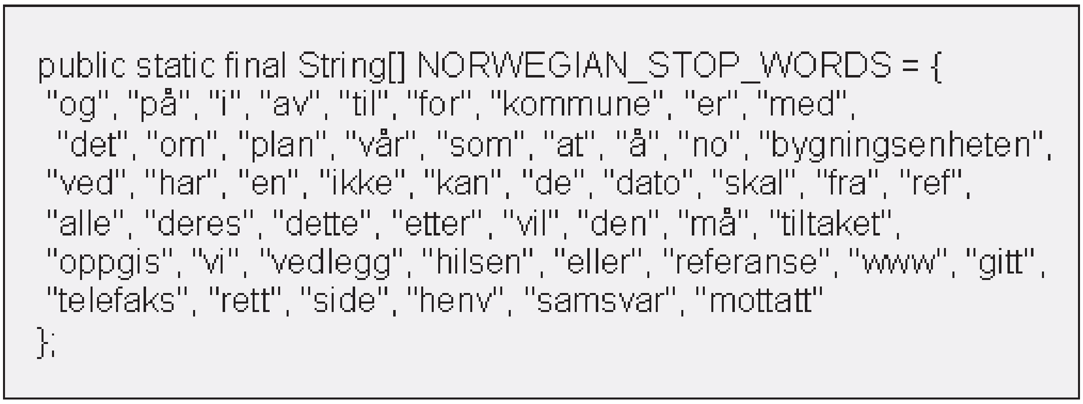
The database was delivered as an MS Excel report, and had to be converted to a relational database. This was achieved by using MS Access as a mediator. The report was first converted into an MS Access database, then - with the help of ODBC - it was converted into the final MySql database. Certain small changes had to be made with the database. As described in chapter four, the Environment service division in the Trondheim Property Service unit consists of mainly cleaners who do not produce any significant documentation relevant for this thesis. All persons employed at this division were removed from the database. Also, there were some inconsistens between the names as they were written in the documents and how they were written in the employee database, thus some minor adjustments had to be made to harmonize the case handler names in these two storages. For instance, in the database the names were written 'Surname Firstname (Middlename)' and in the documents they were represented as 'Firstname (Middlename) Surname'. Java String-methods were applied to take care of this problem in the prototype code.

### 7.1.3 Removing stop words

To optimize both the indexing stage (make sure unwanted stop words are not seen as index terms) and to get rid of noise during the searching, it is common to remove stop words from the documents before they are indexed. After having

[1] See `http://poi.apache.org/`

analyzed all the words in the involved documents, fifty words were considered stop words and removed from the documents before indexing. The stop words removed were:



```
public static final String[] NORWEGIAN_STOP_WORDS = {
 "og", "på", "i", "av", "til", "for", "kommune", "er", "med",
  "det", "om", "plan", "vår", "som", "at", "å", "no", "bygningsenheten",
 "ved", "har", "en", "ikke", "kan", "de", "dato", "skal", "fra", "ref",
 "alle", "deres", "dette", "etter", "vil", "den", "må", "tiltaket",
 "oppgis", "vi", "vedlegg", "hilsen", "eller", "referanse", "www", "gitt",
 "telefaks", "rett", "side", "henv", "samsvar", "mottatt"
};
```

Figure 7.1: Stop words

As figure 7.1 illustrates, many of the stopwords are quite domain specific, and in another setting this list of stop words might appear totally different. An important consideration during the stop word selection was to make sure none of the stop words could be relevant in a expertise search situation (see chapter 6.2.1).

### 7.1.4 Indexing the collection

An inverted index is the result of the indexing stage. This means that the index list, for a term, the documents that contain it. Lucene [26] organizes its indexes into multiple sub-indexes, called segments, and these sub-indexes are stored on disk. Each segment contains the following parts:

- **Fields**: This is the structure stating what information unit is indexed. For instance, the title of the document may be indexed as a title-field, or more common, the content of the document may be indexed as a content-field.
- **Term dictionary**: This is a dictionary containing all terms used in the indexed fields of all the documents that are indexed. The dictionary also stores the number of documents which contain the term and pointers to the terms' frequency and proximity data.
- **Term frequency data**: For each term stored in the dictionary, the numbers of all the documents that contain that term, and the frequency of the term in that document.

- **Term proximity data**: For each term in the dictionary, the position of the term in the document is stored.
- **Normalization factors**: For each field in each document, a value is stored that is multiplied into the score for hits on that field.
- **Term vectors**: For each field in each document, the term vector may be stored. A term vector consists of term text and term frequency.
- **Deleted documents**: An optional file indicating which documents that is deleted from the index.

By using Lucene in the indexing stage, the initial document storage (after parsing) is now reduced by 61 percent, and since Lucene use vector space model principles, the results from a search may be ranked by relevance and not only by the notion that a document is either relevant or not relevant.

## 7.2 Expertise modelling

As chapter 6 described, the modelling or definition of expertise is composed by two alternative approaches. The first one utilizes the name occurrences and the second one uses the number of years the person has been employed at his current position as an indicator of expertise.

This process starts in the same manner with both approaches. It begins with the expertise seeker issuing a query verbalizing the expertise need. Initially, two query formulations were implemented: a keyword query and a boolean query. The Expertise finder prototype then finds all documents relevant to this query and cache each of the relevant documents into another folder on disk. To prepare the actual finding of employees in these documents, the cached documents are indexed. Figure 7.2 shows the folder structure used to organize this process. The 'K2000_documents_doc' folder houses the initial MS Word documents collected from K2000, the 'K2000_documents_txt' folder houses the parsed text documents, the 'Index' folder contains the index segments produced after indexing the parsed text documents, the 'Precache' folder houses the raw and unprocessed text documents that are found relevant in the search for relevant documents, the 'Cache' folder contains the same documents, but now they are processed (tokenized and lowercased, i.e. made search friendly), and the 'Cache_index' folder contains the index segments from the last indexing process.

After the relevant documents found in the expert seekers search are indexed, a new search, using all employees' names in the employee database as queries, finds all employees mentioned in the relevant documents to the initial expertise query

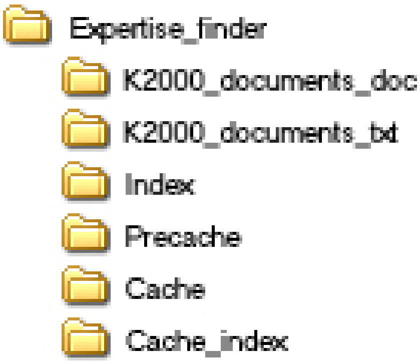


Figure 7.2: Folder structure

issued by the expertise seeker. The code segment below shows the java code that use employees from the employee database as queries.

```
//use all employees as query
for (int i = 0; i < employeeList.length; i++) {
Query queryObject = QueryParser.parse("\"" +
employeeList[i] + "\"",GlobalValues.contentFieldName,
new StandardAnalyzer());
Hits hits = is.search(queryObject);
numHits = hits.length();

//if this employee found in the relevant documents, add
//the employee and the number of documents he is mentioned
//into the treeMap
if (hits.length() > 0) {
counter[i] = hits.length();
tm.put(employeeList[i],occurrence = new Integer(counter[i]));
}
hits = is.search(queryObject);
}
```

## 7.3 Expertise matching

A data structure, a TreeMap (figure 7.3), contains each employee (only those who are mentioned in the relevant documents) together with the number of times he is mentioned in the relevant documents. As TreeMaps can't be sorted by values, we have to transform the keys and values in the TreeMap into a TreeSet (figure 7.4), and then sort the employee-to-occurrences mappings by the number of document mentions.

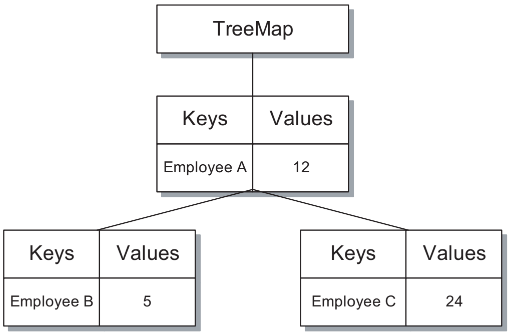


Figure 7.3: TreeMap

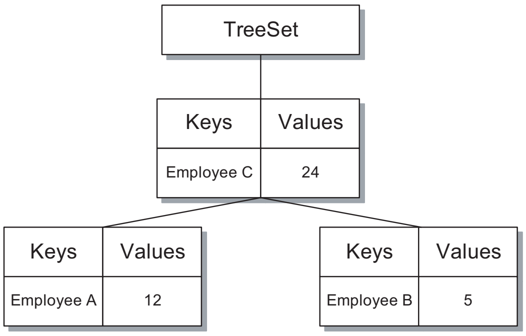


Figure 7.4: TreeSet

The other parameter that may signalize expertise is how much work experience the employee has. The number of years the employee has been employed in his current position reflects the work experience the employee has. Expertise and experience are closely related as chapter two describes. A case handler who has been working with the same issues for twenty years will according to theory (and common sense) often possess more knowledge than a case handler employed for one year. It is of course questionable whether this experience is linear, meaning that the expertise level increases linear with time as the employee has achieved a certain level of competence. Anyway, this parameter deserves testing. The employee database has a field indicating the number of years the employee has been employed in his current position, and the same approach as in the employee-to-occurrences with the TreeMap and TreeSet structures just swapping the number of document mentions with number of years employed.

The filter option is based on the field in the employee database that specifies the logical belonging to the employee. The alternative to the filter option is an expertise-search searching through all employees at the Urban development area (The six participating units). A filter that lets the expertise seeker filter by unit is easily implemented by using a SQL-query. The employee database has a field specifying the unit number connected to each unit (table 7.2).

| Unit | Unit number |
|---|---|
| Building permits office | 523000 |
| Urban zoning office | 522000 |
| Environment office | 510000 |
| Trondheim real estate | 500000 |
| Map and surveying office | 525000 |
| Dep.of infrastructure... | 529000 |

Table 7.2: Unit numbers

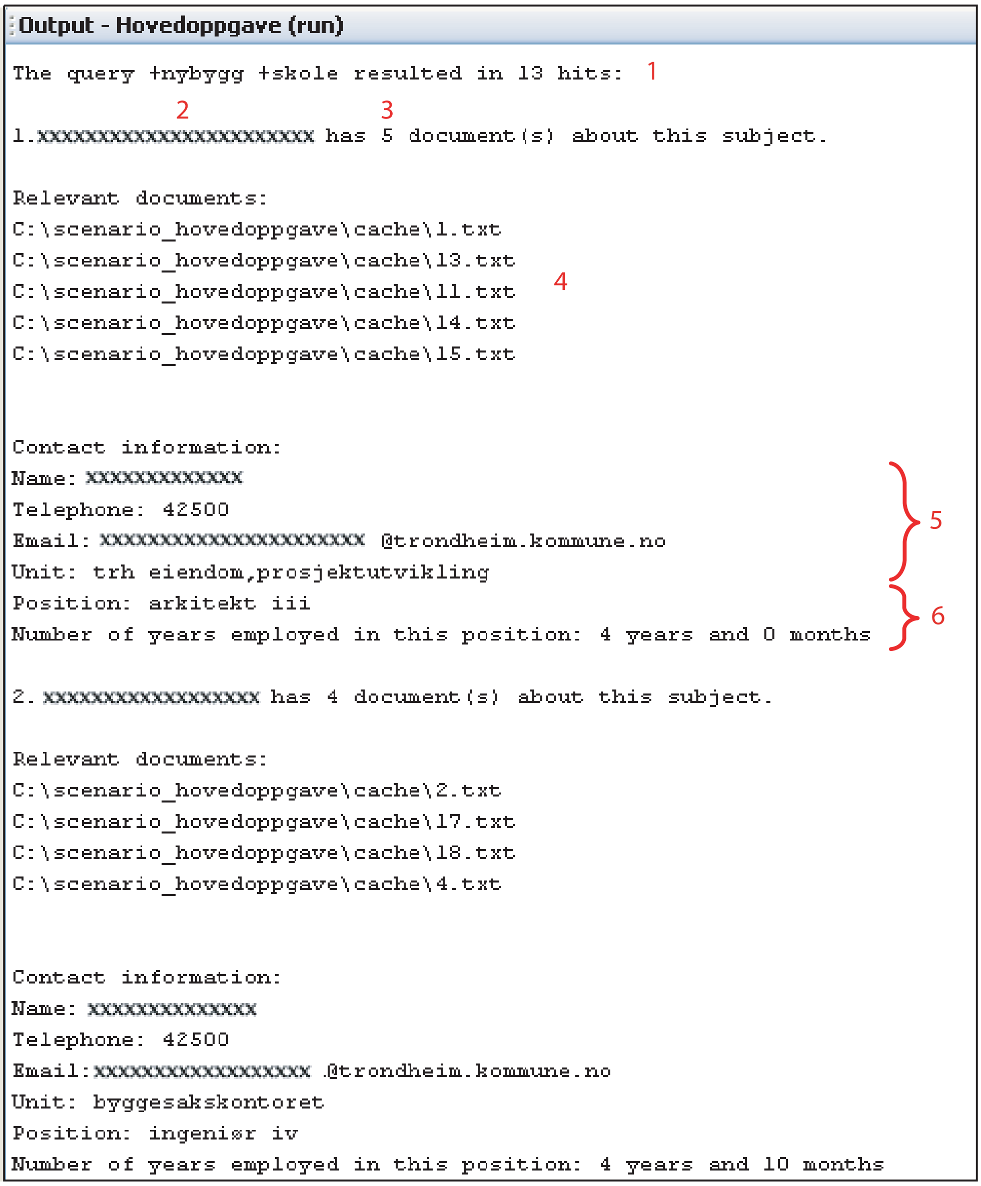
By issuing the following query, only the experts employed at this unit are retrieved:

```
SELECT 'name' FROM 'personal' where 'unit' = 523000;
```

## 7.4 Expertise presentation

After having sorted the relevant experts, either by number of name occurrences or by employment time, and optionally filtered the experts based on the unit they belong to, it's time to present the results to the expertise seeker. Together with the experts' name, we also need to include some contact information such as the telephone number, email address, what unit the expert belongs to, what position he holds and how long he has been employed in his current position. In addition, the expertise seeker should be given an opportunity to validate the experts by reference to the publications the experts are mentioned in.

How the Expertise finder prototype presents experts is illustrated in figure 7.5. The expertise seeker is presented with the issued query and the number of experts found given this query (1). Further, the name of the expert together with how many documents he is mentioned in is presented (2 and 3). The validation information is given by a link to the documents the nominated expert is referenced in (4), the position of the expert and how many years the expert has been employed in this position (6). (5) shows how to access the expert and the experts' logical belonging.



```
The query +nybygg +skole resulted in 13 hits:

1.XXXXXXXXXXXXXXXXXXXXXXXX has 5 document(s) about this subject.

Relevant documents:
C:\scenario_hovedoppgave\cache\1.txt
C:\scenario_hovedoppgave\cache\13.txt
C:\scenario_hovedoppgave\cache\11.txt
C:\scenario_hovedoppgave\cache\14.txt
C:\scenario_hovedoppgave\cache\15.txt


Contact information:
Name: XXXXXXXXXXXXX
Telephone: 42500
Email: XXXXXXXXXXXXXXXXXXXXXXXX @trondheim.kommune.no
Unit: trh eiendom,prosjektutvikling
Position: arkitekt iii
Number of years employed in this position: 4 years and 0 months

2. XXXXXXXXXXXXXXXXXX has 4 document(s) about this subject.

Relevant documents:
C:\scenario_hovedoppgave\cache\2.txt
C:\scenario_hovedoppgave\cache\17.txt
C:\scenario_hovedoppgave\cache\18.txt
C:\scenario_hovedoppgave\cache\4.txt


Contact information:
Name: XXXXXXXXXXXXXX
Telephone: 42500
Email:XXXXXXXXXXXXXXXXXXX .@trondheim.kommune.no
Unit: byggesakskontoret
Position: ingeniør iv
Number of years employed in this position: 4 years and 10 months
```

Figure 7.5: Returned experts from expertise search

# Chapter 8

# Testing and evaluation

This chapter describes the testing of the prototype and performs an evaluation of the results from the tests. The basis for the evaluation is sample queries proposed by the unit managers from the six units participating in the case study. Together with the sample queries, they have also proposed experts they have found relevant as results to the sample queries. By the assumption that these managers are the ones who know the units' employees and their expertise areas- and levels best, this approach contribute to that the evaluation is being handled in a valid and reliable way.

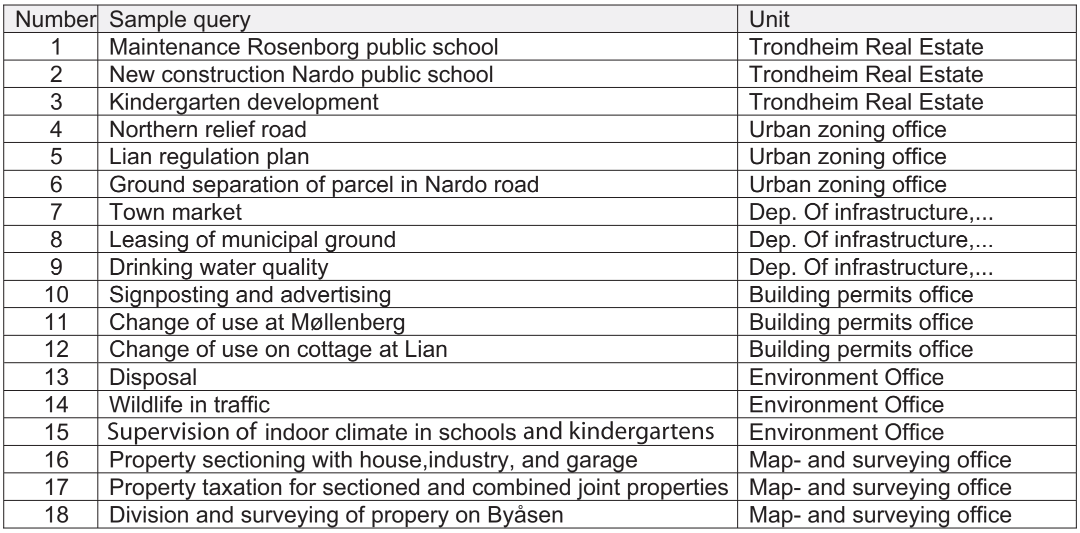
Two alternative approaches will be tested in this chapter, both relevant to the experts' experience. As chapter two concluded, experience has a strong correlation with expertise. Alternative one concerns the number of name occurrences an expert has in documents that deal with the demanded expertise verbalized in the query. Alternative two is based on how long the expert has been employed in his current position. One premise is that in this alternative, the case handler need to have published at least one document relevant to the expertise query. This ensures that the case handler has at least some knowledge with the query given.

The evaluation measures being used are the same measures as those described in chapter 4.5; precision and recall, but since there are a total of eighteen sample queries, some single value summaries are needed to get an overall picture of the evaluation. Two evaluation measures will be applied; the average interpolated precision at recall for all queries and the mean R-precision.

## 8.1 Test data

The tests are run using the K2000 storage containing approximately 2.200 documents. The sample queries together with the proposed experts provided by the unit managers are shown in figure 8.1.

The unit managers have provided 18 sample queries relevant to their unit, and also employees they consider as experts to each query.



| Number | Sample query | Unit |
|---|---|---|
| 1 | Maintenance Rosenborg public school | Trondheim Real Estate |
| 2 | New construction Nardo public school | Trondheim Real Estate |
| 3 | Kindergarten development | Trondheim Real Estate |
| 4 | Northern relief road | Urban zoning office |
| 5 | Lian regulation plan | Urban zoning office |
| 6 | Ground separation of parcel in Nardo road | Urban zoning office |
| 7 | Town market | Dep. Of infrastructure,... |
| 8 | Leasing of municipal ground | Dep. Of infrastructure,... |
| 9 | Drinking water quality | Dep. Of infrastructure,... |
| 10 | Signposting and advertising | Building permits office |
| 11 | Change of use at Møllenberg | Building permits office |
| 12 | Change of use on cottage at Lian | Building permits office |
| 13 | Disposal | Environment Office |
| 14 | Wildlife in traffic | Environment Office |
| 15 | Supervision of indoor climate in schools and kindergartens | Environment Office |
| 16 | Property sectioning with house,industry, and garage | Map- and surveying office |
| 17 | Property taxation for sectioned and combined joint properties | Map- and surveying office |
| 18 | Division and surveying of propery on Byåsen | Map- and surveying office |

Figure 8.1: Sample queries

## 8.2 Deciding the query formulation

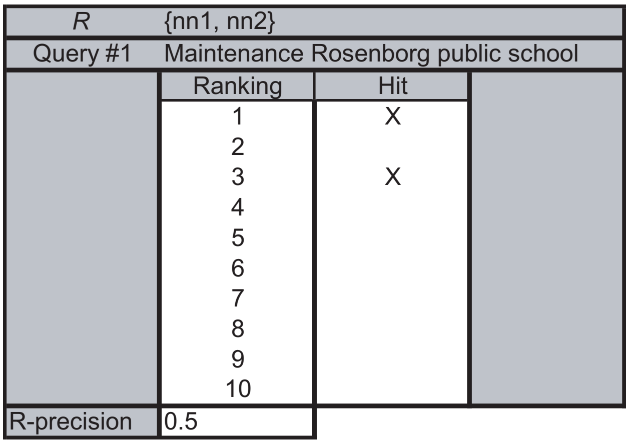
To decide what query formulation method to use when evaluating the prototype, a pre-test was performed. The two query formulation methods tested was a Boolean query and a keyword query. Both alternatives (Based on name occurrences and based on employment time) was tested with the two different query formulations as this aspect might affect the final evaluation results. The Boolean query formulation will in this case use the boolean operator AND to ensure that all query words reside in the relevant documents. The keyword query formulation searches for all the words in the query formulation (except words recognized as stop words). The keyword query, contrary to the boolean query, uses the Term Frequency and Inverse Document Frequency in the Vector Space Model to weight the individual terms in the query, and thereby judge the relevance of the documents in non-binary way (See chapter 4.4.2). The pretest showed that the keyword query formulation performed best[1], and is thereby used in the evaluation.

[1] The keyword query formulation achieved a R-precision of 30 percent when evaluating all queries in all the test sets, whereas the boolean query formulation scored 26 percent

## 8.3 Evaluation based on name occurrences

The tests are run for both each separate unit (the unit where the unit manager has proposed sample queries) and for the entire Urban Development area (the six participating units). Each of these two tests are evaluated based on interpolated precision at recall which provides the evaluation diagram and the R-precision which provides an alternative single summary value.

As an example, a sample query is shown in figure 8.2. Here, the query tested asks for experts having knowledge or competence about maintenance at Rosenborg public school. As the results in figure 8.2 show, two relevant experts have been pre-defined by this units' manager (denoted *nn1* and *nn2* in the figure). The figure also illustrate that both of these predefined experts are found by the prototype system, one at rank 1 and the second expert in rank 3. Since the set $R$ consists of two experts ($R = 2$), the R-precision of this query is 0.5 (or 50 percent).



| | | |
|---|---|---|
| *R* | {nn1, nn2} | |
| Query #1 | Maintenance Rosenborg public school | |
| | Ranking | Hit |
| | 1 | X |
| | 2 | |
| | 3 | X |
| | 4 | |
| | 5 | |
| | 6 | |
| | 7 | |
| | 8 | |
| | 9 | |
| | 10 | |
| R-precision | 0.5 | |

Figure 8.2: Sample query

### 8.3.1 Filtered query

Figure 8.3 shows the interpolated precision at recall scores for this test. As the figure illustrate, this query returned a precision of just below 60 percent at recall levels 0 to 50 percent, about 40 percent at 60 percent recall, and around 35 percent precision at 70-100 percent recall. Another pattern revealed by this figure, is the fact that precision decreases as recall increases. This results reflects the relationship between precision and recall well (See chapter 4.5). The mean R-precision across all queries in this test is 40 percent.

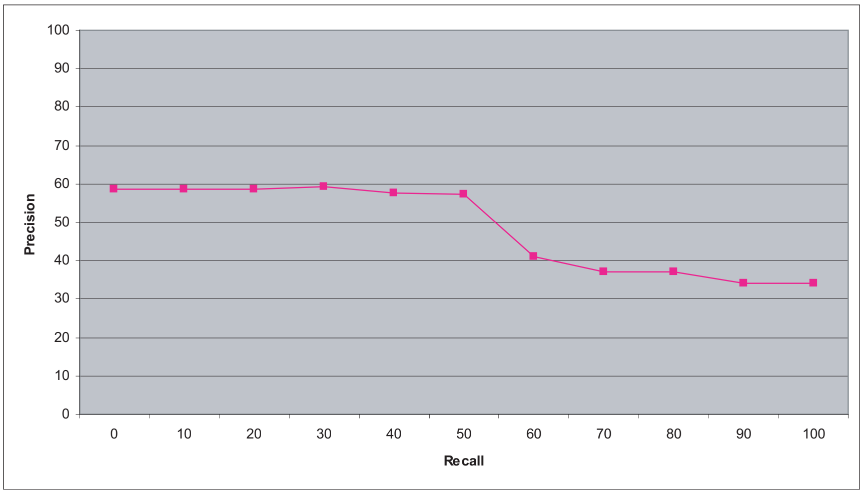


Figure 8.3: Interpolated precision at recall - based on name occurrences in documents and filtering by unit

### 8.3.2 Non-filtered query

When querying using all employees in the Urban development area, the precision at recall curve drops (Figure 8.4). The precision for this query is approximately 40 percent at recall levels 0-40 percent, 35 percent at recall level 50, 25 percent at recall level 60, and then it drops to about 22 percent between recall levels 70-100 percent. This query achieves a mean R-precision score of 32 percent.

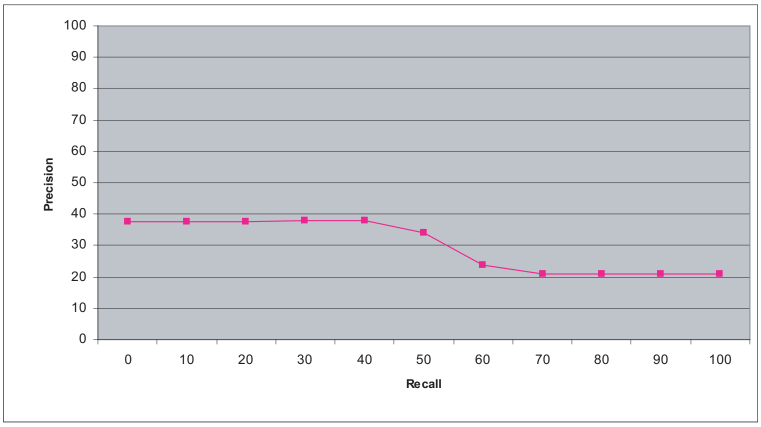


Figure 8.4: Interpolated precision at recall - based on name occurrences in documents and searching for experts using the entire Urban development area

## 8.4 Evaluation based on employment time

This section presents the evaluation results when employment time is used as an indication of expertise.

### 8.4.1 Filtered query

The query based on employment time returns a precision at recall at approximately 45 at 0-20 percent recall, 43 percent at 30 and 50 percent recall, 27 percent at 60-80 percent recall and 25 at 90 and 100 percent recall (Figure 8.5). The mean R-precision score for this test is 26 percent.

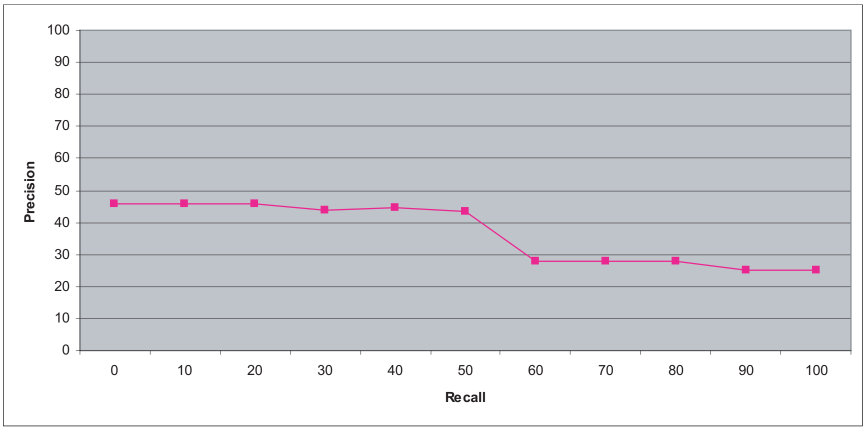


Figure 8.5: Interpolated precision at recall - based on name occurrences and filtering by unit

### 8.4.2 Non-filtered query

When searching for experts in the entire Urban development area using employment time as evidence of expertise, the results show the lowest precision scores of all tests (figure 8.6). This query resulted in a precision at recall of approximately 20 at 0-50 percent recall, 10 at 60 percent recall, and 8 percent precision at 70-100 percent recall. The mean R-precision score reflects the low precision at recall scores with an R-precision of 8 percent.

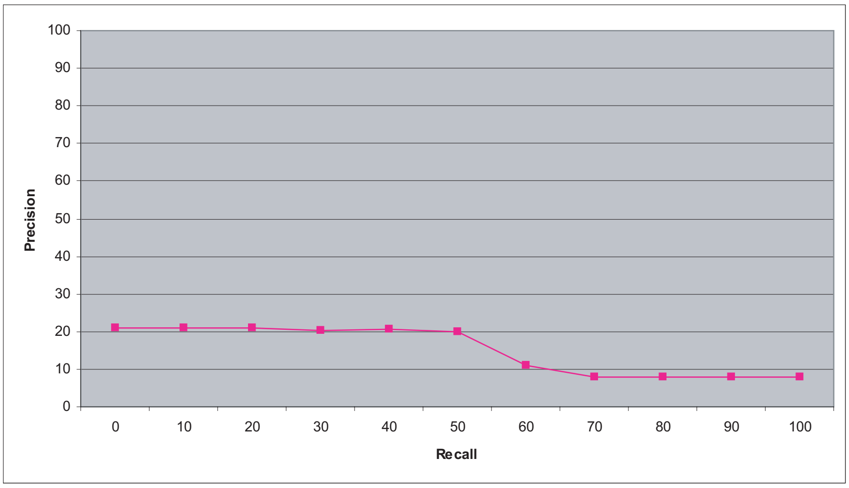


Figure 8.6: Interpolated precision at recall based on employment time and searching for experts using the entire Urban development area

## 8.5 Summarizing the evaluation results

Figure 8.7 shows a summary of the interpolated precision at recall results. We see that the filtered query where the expertise findings are based on name occurrences provide the best score followed by the filtered query where the experts found by employment time. For how the interpolated precision at recall scores are computed for all test sets, see appendix A.2.

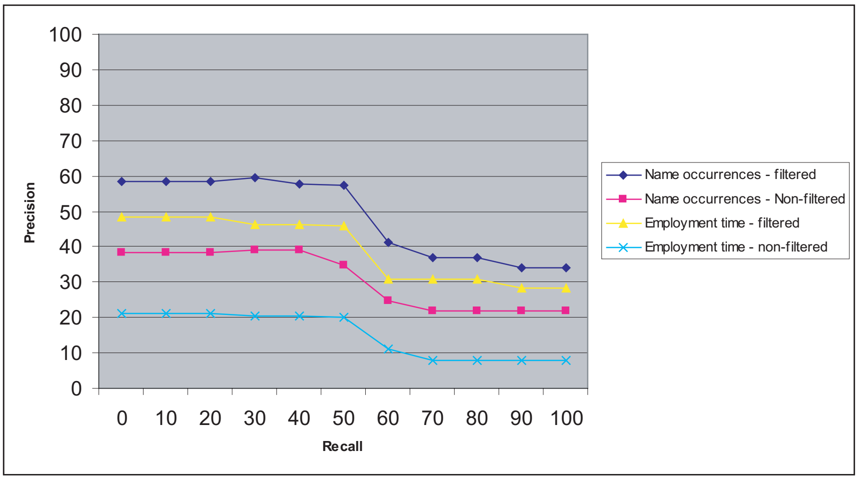


Figure 8.7: Interpolated precision at recall - summary

Figure 8.8 shows a summary of the evaluation results using mean R-precision. For R-precision scores from all tests, see appendix A.1.

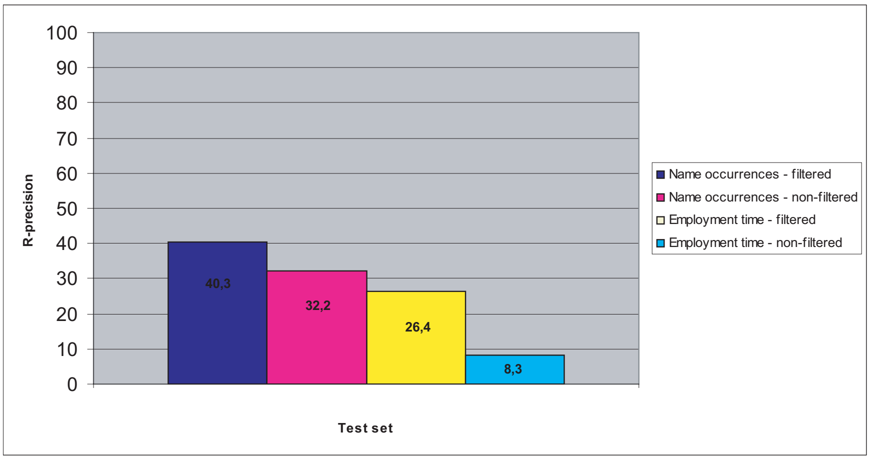


Figure 8.8: R-precision - summary

Curiously, the R-precision results differs from the interpolated precision at recall scores. Here, similarly to the interpolated precision at recall evaluation, the filtered expertise query based on name occurrences achieve the best evaluation score, but the second best R-precision score is the non-filtered query where the experts are ranked based on name occurrences, whereas in the interpolated precision at recall evaluation the filtered query based on employment time achieves the second highest evaluation score. Since R-precision takes the size of the set of predefined relevant experts into consideration, and the fact that these sets in our case consists of rather few experts, this means that in the results from the name occurrences test the relevant experts appear with high rankings.

## 8.6 Comparing the results to related research

To see how this thesis' prototype (and the underlying principles) achieve, this section will compare this thesis' evaluation scores with the evaluation results presented in related research.

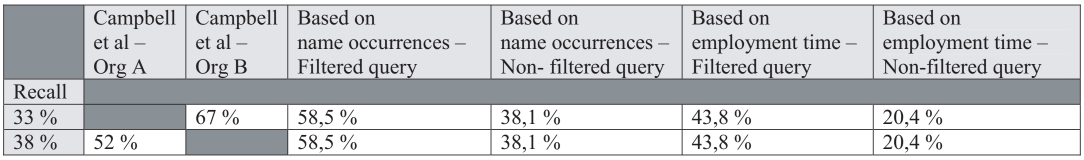
### 8.6.1 Comparing Precision at recall scores

Campbell et al. [6] use email communication to find expertise[2]. They analyzed 13.417 messages from 15 people in one organization and 15.928 messages from 9 people in another organization over a two year period. The top 30 experts on various topics were manually identified. To evaluate the system they used two

[2] Described in chapter 3.2

different approaches. The first approach was a content based approach focusing on email content only. The other approach was a graph based approach that also considered the social networks deduced from email communication patterns. The best evaluation result came from the latter approach where the score for the first organization (OrgA) was 52 percent precision at 38 percent recall and the score for the second organization (OrgB) was 67 percent at 33 percent recall.

Figure 8.9 shows the comparison between Campbell et al.'s system and this thesis' prototype.



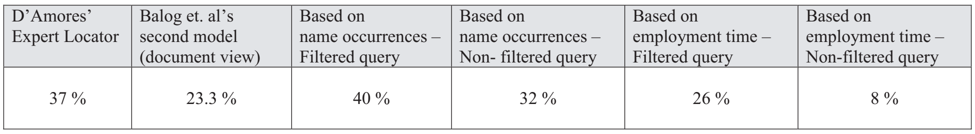
| | Campbell et al – Org A | Campbell et al – Org B | Based on name occurrences – Filtered query | Based on name occurrences – Non- filtered query | Based on employment time – Filtered query | Based on employment time – Non-filtered query |
|---|---|---|---|---|---|---|
| Recall | | | | | | |
| 33 % | | 67 % | 58,5 % | 38,1 % | 43,8 % | 20,4 % |
| 38 % | 52 % | | 58,5 % | 38,1 % | 43,8 % | 20,4 % |

Figure 8.9: Comparison of precision at recall

As the table shows, the evaluation results from Campbell et al.'s orgA achieves the highest precision at recall score followed by the prototypes' test based on name occurrences and the filtered query. The rest of this thesis' test procedures achieve lower than both evaluation scores from Campbell et al.

### 8.6.2 Comparing R-precision scores

Expert Locator by D'Amore [9][3] use the action in different activiy spaces to infer expertise. The evaluation metric used to evaluate the Expert Locator prototype is R-precision, and the mean R-precision across all test queries (32 in total) was 37 percent. Balog et al.'s [4] document-centric model[4] use probabilistic methods to deduce expertise from documents. The evaluation of this model resulted in an R-precision of 23.3 percent. A comparison to this thesis' evaluation is shown in figure 8.10.



| D'Amores' Expert Locator | Balog et. al's second model (document view) | Based on name occurrences – Filtered query | Based on name occurrences – Non- filtered query | Based on employment time – Filtered query | Based on employment time – Non-filtered query |
|---|---|---|---|---|---|
| 37 % | 23.3 % | 40 % | 32 % | 26 % | 8 % |

Figure 8.10: Comparison of R-precision

As the above figure illustrates, according to the mean R-precision scores, the prototype test based on name occurrences and using a filtered query performs better

[3] Presented more thourougly in chapter 3.2

[4] see chapter 3.2

than both D'Amores' Expert Locator and Balog et al.'s model, with 40 percent against 37 percent and 23.3 percent respectively. Expert Locator performs second best, whereas both the test-sets based on name occurrences and non-filtered query and based on employment time and filtered query performs better than Balog et al.'s model.

# Part IV

# Discussion and Conclusion

# Chapter 9

# Discussion

The research approach chosen for this thesis was a combination of a literature review, a case study and a prototype experiment. The results from these three approaches are discussed in the following.

## 9.1 Literature review

The literature review aimed at clarifying the concepts surrounding the notion of expertise and Expertise finders. The literature review section is comprised by three chapters: Expertise, State-of-the-art and Information Retrieval. These chapters are discussed in the following.

### 9.1.1 Expertise

Most of the literature within the fields of expertise and expertise finding embraces a mixture of psychology, sociology, organizational theory and technology. As most articles discussing expertise finding focus on the technological aspects and less on what they actually are developed to find, expertise, I made an effort to clarify the concept of expertise, and to compare it to the related terms of knowledge and competence. This was helpful both in the sense that it gave me a direction to follow during the case study and also an angle as to decide what parameters should be used to find expertise during the prototype development. The term "Expertise finders" seems to be a somewhat misleading name for this kind of computer applications. Expertise, as defined by psychological literature, does not cover the range of qualities and qualifications an Expertise finder is meant to provide. It is not necessarily the expertise, the top competence or the tacit knowledge these applications stride to find, but rather the person who might help the seeker right then and there, no matter if he is an expert or not according to the literatures' description of expertise. This person does not have to be an expert, but he need to possess some important qualities; he must of course possess the element the seeker demands and equally importantly, he must be capable of transferring the necessary assistance to the seeker, whether it is a solution to a problem or filling a function. An Expertise finder system suggesting an expert who merely has the theoretical knowledge of a subject, but who is unable to explicate or practice this knowledge, is worth little or nothing in such a setting. Actually, research show that in some cases the expertise seeker is better off by having someone with intermediate knowledge or skills provide the necessary assistance because the communication between someone fitting the definition as an expert and a novice is difficult to accomplish. First of all it's difficult for the expert to decide the knowledge level of the recipient, something which makes it difficult to decide what granularity level the expertise sharing should take place in. And, the way the expert abstracts his knowledge makes it difficult to explain it to someone who needs a detailed presentation of this knowledge.

The literature review also shows that expertise and experience are two highly correlated concepts. This does not mean that these two concepts are equivalent, but the probability of finding expertise is higher when you encounter an experienced employee than a novice. To become an expert you need a certain amount of experience. This is one of the few features related to expertise that actually might be operationalized and measured. Experience can be found by quantifiable measures, such as the number authorships in documents or the number of mentions in documents about a subject, the number of posts in a discussion forum that concerns a certain subject, the number of emails on a given subject, or by experience based on time of service. This kind of evidence is by [36] called implicit expertise evidence. The explicit expertise evidence consists of self-assessed expertise areas- and levels. I will expand this categorization to include expertise evidence that falls between these two categories, and include information that directly indicates competence and knowledge, but that is not self-assessed or self evaluated. This kind of evidence might be found in curriculum vitas, certificates, competency management reports etc.

### 9.1.2 Expertise finders

The state-of-the-art in chapter three describes how most Expertise finders fit one of three different categories; the database approach, the social network approach or the approach where expertise profiles are made from electronic evidence. The database approach has some severe limitations in that the employees themselves assess their expertise areas and levels. Some of the problems with this approach are that it is difficult to evaluate ones own and others expertise areas- and levels, people might easily overestimate or underestimate their expertise and this self-assessment is a time- and resource consuming effort that effect both the individuals themselves and the organization as a whole. Systems using the social network approach have produced some ok results, but some of the approaches used may not be applicable based on privacy protection issues (e.g. email analysis). The approach based on electronic evidence is least time consuming to maintain, but requires suitable human-to-expertise evidence mappings.

The choice of what approach that will fit the organizations' expertise finding capabilities should depend on the business strategy and from this the knowledge management strategy the organization follows. In an organization utilizing mostly the tacit knowledge embodied within the employees, a personalization strategy is often followed, and the database approach will be best put to use. The social network approach and the approach using electronic evidence to create expert profiles won't be effective in this scenario as there are few or none sources to locate the evidence needed. In organizations focusing on explicit object-based knowledge, the two latter approaches will probably be most suitable (due to the many limitations with the database approach). In this case the documented explicit knowledge will reveal expertise evidence to feed the Expertise finder, either by unveiling social network structures or by unveiling other electronic evidence that map the so-called expertise to humans.

In the two latter approaches, the experts have to be found automatically by mapping expertise evidence to the correct persons; hence some information retrieval approach needs to facilitate this mapping.

### 9.1.3 Information Retrieval

Information Retrievals' main postulate is to reduce the information overflow. This means focusing on the relevant information and cleansing out the non-relevant information. In an Expertise finder setting this means finding the relevant experts and cleansing out the non-relevant ones. Information Retrieval does this by applying techniques such as pre-processing texts with stop word removal and stemming, indexing structures such as inverted files and by using suitable Information Retrieval models such as the Vector Space model, the Boolean Model, the Probabilistic Model or some inferior variant of these. When developing a prototype in this thesis, it was important to be able to relevance-judge the documents mentioning words found in the expertise query, hence the Vector Space Model with its tfidf measuring scheme is a central part of the system.

The evaluation of information retrieval systems is important to assess the reliability of the systems, and should provide measures that reflect how the system is able to retrieve the relevant documents and discard the non-relevant documents.

Common evaluation measures for single-query evaluations are recall and precision, where recall denotes the fraction of the relevant experts which has been retrieved and precision denotes the fraction of the retrieved experts which are relevant. For evaluating several queries however, it is common to average the precision at a certain amount of recall levels (usually eleven) or use R-precision, which takes into account the number of experts in the set of predefined relevant experts assessed by the evaluators. A good Expertise finding system should focus on suggesting the *real* experts as high on the resulting ranking list as possible, thus precision is often prioritized over recall. R-precision takes into account the set of relevant experts when evaluating the system, hence to achieve a high R-precision score, the relevant experts need to be ranked highly by the Expertise finder relatively to the set of predefined experts. Some literature use precision at certain cutoff values (e.g. [29]) where one decides the maximum rank to be considered when computing the precision score, for instance only the 5 or 10 top positions. It is believed that R-precision gives a more reliable answer given the fact that unit managers have proposed a set of experts who should be nominated as experts in our system, and assuming they know their personnel well enough to judge this correctly. Besides, R-precision is regarded as a good overall measure of retrieval performance [2].

## 9.2 Case study

The main findings from the study show that there is a lot of information available for expertise finding. The knowledge created and used at the Urban development area consist mainly of explicit, object-based knowledge. Laws and regulations, rules and procedures guide much of the case handling that is being performed, and as the case handlers use this knowledge to perform case handling, new explicit knowledge is created in the decisions they make. This new knowledge is materialized into statutes and other formal documents stored in the case handling-/information system K2000s' document storage. During the study, it soon became apparent that the expertise evidence is located in two main sources; the documents archived in the document storage related to K2000 and the personal information found in the employee records, and that the approach involving expertise profiles based on electronic evidence was suitable to locate and map the expertise found in this corporate data.

Most of the relevant expertise evidence involves the case handler. It is the case handler who mainly produce the documents residing in K2000 and it is mainly the case handler who is the wanted expertise when people make inquiries. Some information may be found in the intranet and internet, but the employee-to-document mapping may be less consistent in these sources.

Expertise-related inquiries come from both external and internal instances, and these two different inquiries often involved different elements of expertise. The internal inquiries were often focused on knowledge issues, e.g. who possess the knowledge about a given subject to answer a question. Whereas the internal needs was focused on the competence issues, e.g. the need to find someone who possess the skills to fill a function. The citizens issue applications, complaints and other kinds of inquiries which again are followed up by new inquiries, whereas the different units at Urban Development often work together in projects to utilize the different kinds of knowledge existing in each unit. All these aspects require the knowledge of whom to direct the inquiries to. Today, much of the external inquires are handled at the City reception, which acts as a funnel for inquiries. There are a lot of inquiries to re-distribute around the Urban Development area, and when we take into account that there are 1100 people working in this area (Approximately 830 in the six units involved in the case study), and that there at times are novices and temps working at this desk, we understand that there is a need for a solution that may help find the right employee (expert). To illustrate this with an example, I talked to an employee at the unit responsible for employee information in the municipality. She complained that she every once in a while received misplaced calls from City reception. At the day of the meeting she told me she once received a call from a drug addict begging for a methadone refill.

The internal finding of expertise is handled differently. When I interviewed the unit managers, most of them said that when projects were manned, they used their social relations to find the right project composition. In this case we were talking about projects within this actual unit. Ok, but what if you were to put together a multidisciplinary project team from, say 5 different units? How would you go about to gather the right people from four other units you hardly know?

Yimam-Seid and Kobsa [36] emphasize the importance of internal expertise seeking, and especially in cases of large organizations, geographically distributed organizations and organizations with a heterogeneous composition of employees (e.g. strict division borders, different knowledge backgrounds, different histories due to company mergers). With 1100 employees and often very various responsibilities amongst the different units, the Urban development area fits this description well. Another problem with finding expertise through social relations is the validation problem. How can you be sure (in a reasonable amount of time) that the people recommended by these social relations possess the expertise needed?

Another important finding from this study is that there is a difference among the units when it comes to case handling. Some units, for instance the Building permits office, perform a ”mass handling” of applications. Every year several thousand applications are received which all need case handling. The Environment Office on the other hand performs case handling in a totally different way. This unit performs mostly advisory tasks. The consequence of this is at least two-fold. One is that the Building permits office often receives applications involving the same actions (analysis carried out at the Housing office indicates that there is a need for over a thousand new residences each year) and thus require several case handlers with the same competence. The Environment Office on the other hand, receives few applications or inquiries, and has more specialized case handlers. Secondly, the use of explicit vs. tacit knowledge will probably be affected by whether the unit mass-handles applications or if it performs advisory tasks. When treating a building application, the case handler acts on behalf of the planning- and building law, the technical regulations stated by the National Office of Building Technology and Administration, and previous cases concerning the same actions and the use of explicit knowledge. Whereas the advisory tasks produced by the environment unit requires more ”plowing of new ground” and the use of tacit knowledge. This is also reflected in the productivity of written documentation in each unit. While the Building permits office produce over 3000 statutes each year, the Environment office merely produce 100 official K2000 documents, hence probably making it more difficult to extract expertise from the Environment office from the K2000 document storage.

## 9.3 Prototype experiment

The prototype experiment section is made up by the chapters prototype design, prototype development and testing and evaluation. These chapters are discussed below:

### 9.3.1 Prototype design

The findings from the literature review showed that there exist a strong correlation between expertise and experience, hence by finding a person who has major experience on a subject, you often indirectly find either the expert on this subject or a person that is sufficiently involved in this subject that he knows who the real expert is. The case study showed that two central components in the Expertise finder prototype ought to be the K2000 document storage and the employee record database. Both these components contain expertise evidence: the K2000 document storage contain documents that might reveal a mapping between a given subject and the person possessing expertise about this subject. The employee database on the other hand finds an employees’ experience by the employment time database field. From this, two principles in finding expertise is proposed:

1. An expert is found based on the number of relevant documents his name is mentioned in.

2. An expert is found based on how long he has been employed in his current position.

One might of course discuss if the suggested approach gathers the *real* and *all* experts. In general, all Expertise finders based on finding electronic expertise evidence, whether they find the expertise based on analyzing social network structures or create expert profiles by searching through corporate data such as this thesis' prototype, merely find the experts who publish- or who are mentioned in these sources. In this thesis though, the observation and the interviews in the case study revealed that it usually is the employee who publishes documents or who is mentioned in documents who constitute the wanted expertise, hence this should not pose a big threat in our case.

Another possible danger with this approach, is that there are no semantics involved and no prioritizing of the words found in the documents. A document is a "bag-of-words", meaning that it makes no difference whether the words are in the document name, title or content. Often the documents written by the case handlers used standard templates. This means that if we were to find experts given the query *universal design* which is a common term in this domain, we might find suggested expertise based on the fact that this term is used as part of one document template and occur in several documents without being relevant to the actual content of the documents. Hence, the experts proposed might not possess expertise related to the essence of universal design.

This prototype is developed based on the findings from the case study at the Urban development area. With this in mind one might say that this prototype is rather domain specific. On the other hand, written documentation with a similar structure is common in most public offices. Also, other fields emphasize the importance of utilizing the explicit knowledge found in documents, e.g. research organizations or legal offices. In such a system, one major problem is to define the terminology to use when issuing expertise queries, and especially in domain specific Expertise finders the expertise seeker need to know what words and expressions that are being used in the domain. This is an issue that should be taken seriously in this scenario also, as there are many domain specific concepts and words being used that might be unfamiliar to the ”common person". But taking the domain terminology into consideration requires a lot of time and resources, and hence I have not elaborated on this any further.

### 9.3.2 Prototype development

The main objective with the prototype development has been to develop a prototype that makes it possible to test the principles suggested in this thesis, and not to develop a fully fledged Expertise finder system. Less weight has been put areas such as designing a user interface following the rules of usability.

One central component in the prototype is the search engine. The search engine is an adaptation from the open source search engine Lucene. Lucene is developed in Java, and all other development related to the prototype is done in Java, but other programming languages might have been used as Lucene is ported to several other languages such as C++ and Python.

The information retrieval principles have been applied to the extent it has been considered necessary, something which includes all stages presented in figure **??** which present an overview of the search process. The query was initially parsed into two alternative query formulations: a keyword query and a Boolean query. To decide what query formulation to use during the testing, a pre-test was carried out. This pre-test showed that the keyword query resulted in the best evaluation score and is because of this used in the actual prototype testing (See chapter 8.2). The pre-processing stage consisted of stopword-removal. Lucene has a built-in list of stopwords, but this list consists of English stopwords. Obviously, the documents residing in the K2000 document storage are written using the Norwegian language, thus the Lucene stopword list had to be swapped with a Norwegian one. To find these stopwords, a search using the the entire document collection (2213 documents) was performed to find the most occurring words. These words was analyzed to see if any of the fifty most occurring words might be relevant to find relevant documents given the expertise query. After the possibly relevant words were taken out of the top-fifty list, the remaining words were treated as stopwords, and not taken into consideration during the indexing stage. Using a stemmer to improve the systems' relevance judgement is a controversial subject within the information retrieval field [3] and I chose not to include stemming in this prototype.

The prototype includes two indexing stages. The first one indexes the entire document collection. The second indexing stage indexes the documents regarded as relevant based on the query issued by the expertise seeker. This is probably not the most effective way (with regards to performance nor agility) and there are probably better utilizations of the Lucene search engine. One approach that was tried was to use a search-in-search where the results from one search is used as the source of the second search. However, when using this approach, the results conflicted from the double-indexing approach, and was not followed up any further.

### 9.3.3 Testing and evaluation of prototype

Four different test were carried out in the testing and evaluation stage:

- Based on employees' names occurring in relevant documents and filtering the search by the unit whose manager suggested the sample query.
- Based on employees' names occurring in relevant documents and not filtering the search.
- Based on employment time and filtering the search by the unit whose manager suggested the sample query.
- Based on employment time and not filtering the search.

30 sample queries were suggested from the unit managers at the Urban development area. Only 18 of these were used in the actual testing (3 from each unit) due to the fact that some query variation was needed, both in the sense of query length and query subject. These queries are formulated by someone having extensive domain knowledge, and whether the query formulation would be similar in a real setting is questionable. However, Expertise finders are seldom used by external actors, and in the context of the Urban development area, the hosts at the City reception would be ones formulating the queries in this case, also having domain knowledge.

As the evaluation results in chapter 8 show, both of the filtered queries achieve substantially higher scores than the non-filtered queries. One tendency when querying using all employees, was that the mass-producing units - such as the Building permits office - dominated the resulting list of nominated experts. Many of the rather domain specific terms used within the Urban development area are used by several of the units. Thus, the case handler who produce most documents are nominated as the uppermost expert regardless of which unit he belongs to. As an example of this take the query *garbage*. When using the filtered query where only the employees at the Environment office is considered, the expert suggested by the unit manager ranks first in the results, whereas when using the unfiltered search, employees from the Department of Infrastructure, Environment and Property Management and Trondheim Real estate ranks higher, and the nominated expert is ranked as number six. This issue pose a serious weakness with the prototype. Often, when inquiries are issued at the City reception, the hosts employed there might not know what unit the inquiry should be distributed to, hence a non-filtered query would often be most appropriate. Given the low evaluation scores using this approach, this might not be feasible. However, in situations where it is obvious what unit the inquiry should be distributed to, and where the expertise is situated, the filtered query can be used, with seemingly good performance.

The evaluation scores from the testing are compared to evaluation scores found in related research. Both the interpolated precision at recall scores and the R-precision scores seem to perform well compared to other results, and especially when using expertise search based on name occurrences and a filtered approach.

One element of uncertainty is the size of the document collection used. The K2000 document repository consists of approximately 21.000 documents from the years 2005 and 2006, and of this a subset of 2213 documents were extracted manually. It was practically impossible to retrieve all documents residing in the K2000 document repository. A request was made to extract all documents, but somehow the organization managing the document database was not capable of performing this extraction. However, the consequences of this are not severe. In fact, it is reasonable to believe that the evaluation results would be even better with a larger document collection, at least in the filtered queries, as the experts nominated by the unit managers normally seem to reflect the name occurrences found in the relevant K2000 documents.

# Chapter 10

# Conclusion and further work

## 10.1 Conclusion

This thesis has investigated how expertise might be found without using costly and labour-intensive self-assessments of expertise. This investigation concludes with that the expertise found in expert finders not necessarily match the definitions of expertise used by e.g. the psychological literature. Expert finders main objective is to provide a user with someone who might help right there and then, which not necessarily requires an expert, but someone with sufficient knowledge or competence to provide a solution to some problem or to fill a function. A central indication of expertise is found by using experience parameters as an indicator. Literature show that experience is highly correlated with expertise, and that experience indicators are quite easily revealed for instance in human resource records or in other electronic sources.

An automatic approach to expert finding is proposed. This approach use information retrieval principles to find nominated experts in formal correspondence at the Urban development area in the Municipality of Trondheim. A case study was carried out to investigate what kind of expertise was utilized and possessed in this area, what kind of expertise was sought after, and how this expertise could be located automatically. Findings from this study showed that a central "expertise hub" was the document repository connected to the information system, K2000. Here, all statutes and other outgoing, formal documents are archived, and a lot of explicit, object based knowledge resides in this repository.

A prototype is developed, which uses a search engine to find expertise in documents collected from this repository. Two main experience parameters are used to find expertise evidence: employees' names occurring in documents, and how long the employee has been employed in his current position. The prototype is evaluated based on the information retrieval evaluation measures R-precision and Interpolated precision at recall. The evaluation shows that the approach based on name occurrences achieves best, with scores that compete with, and in some cases achieve better than related research.

## 10.2 Further work

The work carried out in this thesis could be continued into many interesting directions. First of all it would be interesting to see how the proposed principles would achieve with a complete collection of the documents residing in the K2000 repository. Having used only a subset of the total amount of documents archived in the repository, it is some uncertainty how the results would have been in a more “genuine” setting.

Secondly, testing this prototype in another context would assess how domain specific this prototype really is. The documents used as test data in this thesis' experiment contain a rather rigorous structure, and it is not for certain that similar evaluation results would be achieved in another setting with a different kind of documents.

Thirdly, an organization usually possess several heterogenous repositories that could possibly contain expertise evidence. It would be quite easy to expand this thesis' prototype to include sources such as an organizations' Intranet and Internet pages, or other kinds of document formats.

# Bibliography

[1] Maryam Alavi, John Cook, Lucy Cook, and Dorothy Leidner. Knowledge Management and Knowledge Management Systems. *MIS Quarterly*, 25(1), 2001.

[2] Javed A. Aslam, Virgiliu Pavlu, and Emine Yilmaz. A Sampling Technique for Efficiently Estimating Measures of Query Retrieval Performance Using Incomplete Judgements. In *Proceedings of the 22nd International Conference on Machine Learning*, Bonn, Germany, 2005.

[3] Ricardo Baeza-Yates and Berthier Ribeiro-Neto. *Modern Information Retrieval.* ACM Press, New York, 1999.

[4] Kristian Balog, Leif Azzopardi, and Maarten de Rijke. Formal Models for Expert Finding in Enterprise Corpora. In *SIGIR'06*, pages 43–50, Seattle, Washington, USA, 2006.

[5] Rainer Bromme, Riklef Rambow, and Matthias Nuckles. Expertise and Estimating What Other People Know: The Influence of Professional Experience and Type of Knowledge. *Journal of Experimental Psychology: Applied*, 7(4):317–330, 2001.

[6] Christopher S. Campbell, Paul P. Maglio, Alex Cozzi, and Byron Dom. Expertise Identification using Email Communications. In *Proceedings of the twelfth international conference on Information and knowledge management*, pages 528–531, New Orleans, USA, 2003.

[7] Michelene T. H. Chi. Two Approaches to the Study of Experts' Characteristics. In *The Cambridge Handbook of Expertise and Expert Performance.* Cambridge University Press, New York, 2006.

[8] Chun Wei Choo, Brian Detlor, and Don Turnbull. The Structure and Dynamics of Organizational Knowledge. In *Web Work: Information Seeking and Knowledge Work on the World Wide Web*. Kluwer Academic Publishers, 2000.

[9] Raymond D'Amore. Expertise Community Detection. In *SIGIR '04*, pages 498–499, Sheffield, South Yorkshire, UK, 2004.

[10] Thomas H. Davenport and Lawrence Prusak. What Do We Talk about When We Talk about Knowledge? In *Working Knowledge: How Organizations Manage What They Know*. Harvard Business School Press, Boston, 2000.

[11] Kevin C. Desouza. Barriers to Effective Use of Knowledge Management Systems in Software Engineering. *Communications of the ACM*, 46(1):99–101, 2003.

[12] Torgeir Dingsøyr, Hans Karim Djarraya, and Emil Røyrvik. Practical Knowledge Management Tool Use in a Software Consulting Company. *Communications of the ACM*, 48(12):96–100, 2005.

[13] K. Anders Ericsson. An Introduction to The Cambridge Handbook of Expertise and Expert Performance. In *The Cambridge Handbook of Expertise and Expert Performance*. Cambridge University Press, New York, 2006.

[14] P.J. Feltovich, M.J. Prietula, and K.A. Ericsson. Studies of Expertise from Psychological Perspectives. In *The Cambridge Handbook of Expertise and Expert Performance*. Cambridge University Press, New York, 2006.

[15] E. Garcia. Description, Advantages and Limitations of the Classic Vector Space Model, 2007. `http://www.miislita.com/term-vector/term-vector-3.html`.

[16] Jungpil Hahn and Mani R. Subramani. A Framework of Knowledge Management Systems: Issues and Challenges for Theory and Practice. In *ICIS '00: Proceedings of the twenty first international conference on Information systems*, pages 302–312, Atlanta, GA, USA, 2000. Association for Information Systems.

[17] Morten T. Hansen, Nitin Nohria, and Thomas Tierney. What's Your Strategy for Managing Knowledge. *Harvard Business Review*, March-April, 1999.

[18] Pamela J. Hinds and Jeffrey Pfeffer. Why Organizations Don't "Know What They Know": Cognitive and Motivational Factors Affecting the Transfer of Expertise. In *Sharing Expertise - Beyond Knowledge Management*. MIT Press, Cambridge, 2003.

[19] Earl Hunt. Expertise, Talent and Social Encouragement. In *The Cambridge Handbook of Expertise and Expert Performance*. Cambridge University Press, New York, 2006.

[20] Katja Karhu. Expertise Cycle - an Advanced Method for Sharing Expertise. *Journal of Intellectual Capital*, 3(4):430–446, 2002.

[21] Henry Kautz, Bart Selman, and Mehul Shah. Referral Web: Combining Social Networks and Collaborative Filtering. *Communications of the ACM*, 40(3):63–65, 1997.

[22] Linda Lai. *Strategisk Kompetansestyring.* Fagbokforlaget, Bergen, 2. utg. edition, 2003.

[23] Hubert L.Dreyfus and Stuart E. Dreyfus. Expertise in Real World Contexts. *Organization Studies*, 26(5):779–792, 2005.

[24] Tobias Ley. *Organizational Competency Management - A Competence Performance Approach.* PhD thesis, University of Graz, Graz, Austria, 2006.

[25] Guojun Lu. In *Multimedia Database Management Systems*, pages 82–83. Artech House Inc., Boston, 1999.

[26] Apache Lucene. Apache lucene - Index File Formats, 2001. `http://lucene.apache.org/java/docs/fileformats.html`.

[27] David W. McDonald and Mark S. Ackerman. Expertise Recommender: a Flexible Recommendation System and Architecture. In *Proceedings of the 2000 ACM conference on Computer supported cooperative work*, pages 231–240, Philadelphia, Pennsylvania, United States, 2000.

[28] Alistair McLean, Anne-Marie Vercoustre, and Mingfang Wu. Combining Evidence from Web Pages and Corporate Data. In *Proceedings of the 8th Australasian Document Computing Symposium*, Canberra, Australia, 2003.

[29] Alistair McLean, Anne-Marie Vercoustre, and Mingfang Wu. Combining Structured Corporate Data and Document Content to Improve Expertise Finding. *ArXiv Computer Science e-prints*, (September), 2005.

[30] Ikujiro Nonaka and Hirotaka Takeuchi. Theory of Organizational Knowledge Creation. In *The Knowledge Creating Company.* Oxford University Press, 1995.

[31] P. L. Powell, J. H. Klein, and N. A. D. Connell. Experts and Expertise - The Social Context of Expertise. In *Proceedings of the 1993 conference on Computer personnel research*, pages 362 – 368, St Louis, Missouri, United States, 1993. ACM.

[32] A. Rebecca Reuber, Lorraine S. Dyke, and Eileen M. Fisher. Using a Tacit Knowledge Methology to Define Expertise. *ACM*, 1990.

[33] Yee-Wai Sim and Richard Crowder. Evaluation of an Approach to Expertise Finding. In *Proceedings of 5th International Conference on Practical Aspects of Knowledge Management*, pages 141 – 152, Vienna, Austria, 2004.

[34] Tove Thagaard. *Systematikk og Innlevelse : En Innføring i Kvalitativ Metode*. Fagbokforl., Bergen, 2. utg. edition, 2003.

[35] Ilkka Tuomi. Data is more than knowledge: implications of the reversed knowledge hierarchy for knowledge management and organizational memory. *J. Manage. Inf. Syst.*, 16(3):103–117, 1999.

[36] Dawit Yimam-Seid and Alfred Kobsa. Expert-finding Systems for Organizations. In *Sharing Expertise - Beyond Knowledge Management*. MIT Press, Cambridge, Massachusetts, 2003.

# Part V

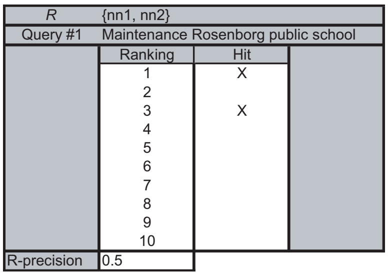
# Appendix

# Appendix A

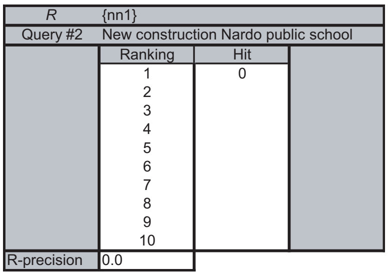
# Test documentation

## A.1 Expertise query results

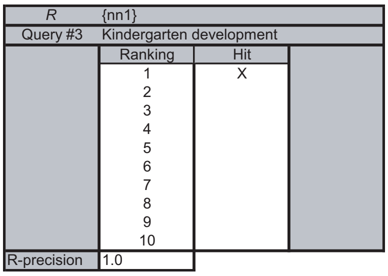
### A.1.1 Based on name occurrences - Filtered query



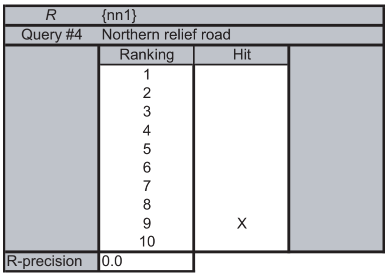
| | | |
|---|---|---|
| *R* | {nn1, nn2} | |
| Query #1 | Maintenance Rosenborg public school | |
| | Ranking | Hit |
| | 1 | X |
| | 2 | |
| | 3 | X |
| | 4 | |
| | 5 | |
| | 6 | |
| | 7 | |
| | 8 | |
| | 9 | |
| | 10 | |
| R-precision | 0.5 | |



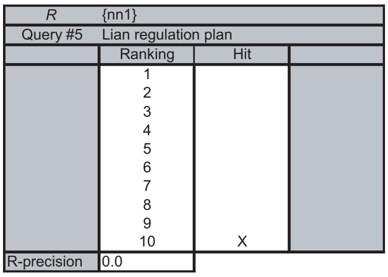
| | | |
|---|---|---|
| *R* | {nn1} | |
| Query #2 | New construction Nardo public school | |
| | Ranking | Hit |
| | 1 | 0 |
| | 2 | |
| | 3 | |
| | 4 | |
| | 5 | |
| | 6 | |
| | 7 | |
| | 8 | |
| | 9 | |
| | 10 | |
| R-precision | 0.0 | |



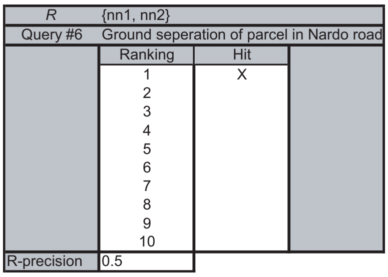
| | | |
|---|---|---|
| *R* | {nn1} | |
| Query #3 | Kindergarten development | |
| | Ranking | Hit |
| | 1 | X |
| | 2 | |
| | 3 | |
| | 4 | |
| | 5 | |
| | 6 | |
| | 7 | |
| | 8 | |
| | 9 | |
| | 10 | |
| R-precision | 1.0 | |



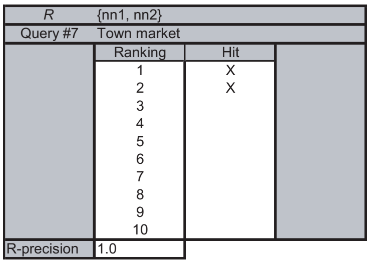
| | | |
|---|---|---|
| *R* | {nn1} | |
| Query #4 | Northern relief road | |
| | Ranking | Hit |
| | 1 | |
| | 2 | |
| | 3 | |
| | 4 | |
| | 5 | |
| | 6 | |
| | 7 | |
| | 8 | |
| | 9 | X |
| | 10 | |
| R-precision | 0.0 | |



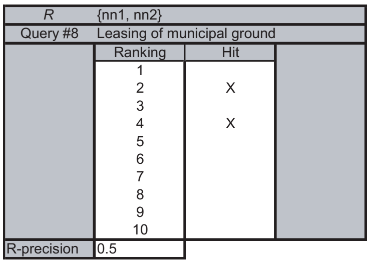
| | | |
|---|---|---|
| *R* | {nn1} | |
| Query #5 | Lian regulation plan | |
| | Ranking | Hit |
| | 1 | |
| | 2 | |
| | 3 | |
| | 4 | |
| | 5 | |
| | 6 | |
| | 7 | |
| | 8 | |
| | 9 | |
| | 10 | X |
| R-precision | 0.0 | |



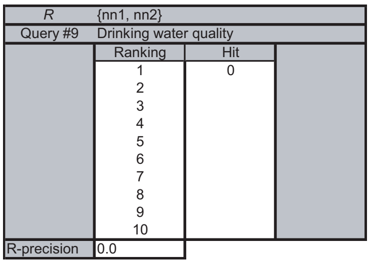
| | | |
|---|---|---|
| *R* | {nn1, nn2} | |
| Query #6 | Ground seperation of parcel in Nardo road | |
| | Ranking | Hit |
| | 1 | X |
| | 2 | |
| | 3 | |
| | 4 | |
| | 5 | |
| | 6 | |
| | 7 | |
| | 8 | |
| | 9 | |
| | 10 | |
| R-precision | 0.5 | |



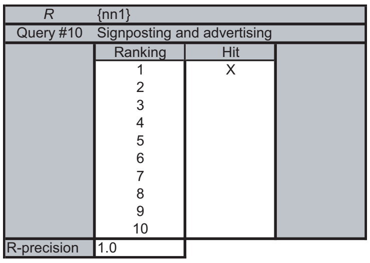
| | | |
|---|---|---|
| *R* | {nn1, nn2} | |
| Query #7 | Town market | |
| | Ranking | Hit |
| | 1 | X |
| | 2 | X |
| | 3 | |
| | 4 | |
| | 5 | |
| | 6 | |
| | 7 | |
| | 8 | |
| | 9 | |
| | 10 | |
| R-precision | 1.0 | |



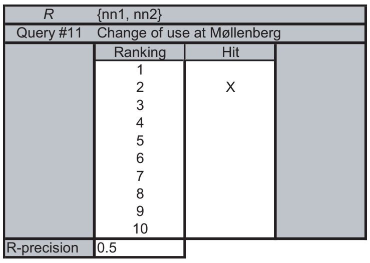
| | | |
|---|---|---|
| *R* | {nn1, nn2} | |
| Query #8 | Leasing of municipal ground | |
| | Ranking | Hit |
| | 1 | |
| | 2 | X |
| | 3 | |
| | 4 | X |
| | 5 | |
| | 6 | |
| | 7 | |
| | 8 | |
| | 9 | |
| | 10 | |
| R-precision | 0.5 | |



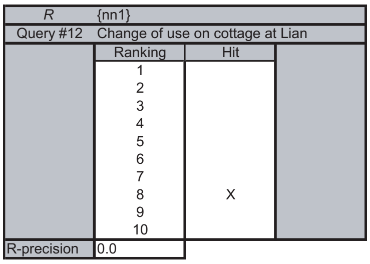
| | | |
|---|---|---|
| *R* | {nn1, nn2} | |
| Query #9 | Drinking water quality | |
| | Ranking | Hit |
| | 1 | 0 |
| | 2 | |
| | 3 | |
| | 4 | |
| | 5 | |
| | 6 | |
| | 7 | |
| | 8 | |
| | 9 | |
| | 10 | |
| R-precision | 0.0 | |



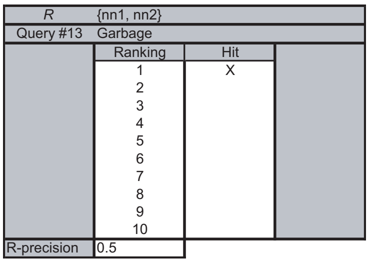
| | | |
|---|---|---|
| *R* | {nn1} | |
| Query #10 | Signposting and advertising | |
| | Ranking | Hit |
| | 1 | X |
| | 2 | |
| | 3 | |
| | 4 | |
| | 5 | |
| | 6 | |
| | 7 | |
| | 8 | |
| | 9 | |
| | 10 | |
| R-precision | 1.0 | |



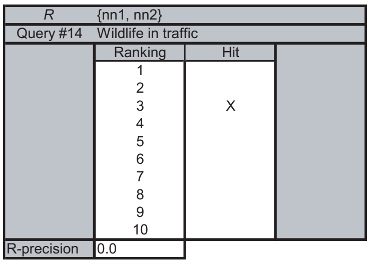
| | | |
|---|---|---|
| *R* | {nn1, nn2} | |
| Query #11 | Change of use at Møllenberg | |
| | Ranking | Hit |
| | 1 | |
| | 2 | X |
| | 3 | |
| | 4 | |
| | 5 | |
| | 6 | |
| | 7 | |
| | 8 | |
| | 9 | |
| | 10 | |
| R-precision | 0.5 | |



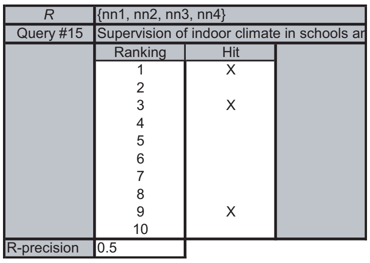
| | | |
|---|---|---|
| *R* | {nn1} | |
| Query #12 | Change of use on cottage at Lian | |
| | Ranking | Hit |
| | 1 | |
| | 2 | |
| | 3 | |
| | 4 | |
| | 5 | |
| | 6 | |
| | 7 | |
| | 8 | X |
| | 9 | |
| | 10 | |
| R-precision | 0.0 | |



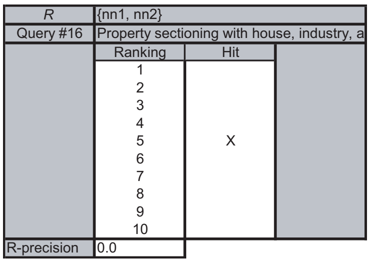
| | | |
|---|---|---|
| *R* | {nn1, nn2} | |
| Query #13 | Garbage | |
| | Ranking | Hit |
| | 1 | X |
| | 2 | |
| | 3 | |
| | 4 | |
| | 5 | |
| | 6 | |
| | 7 | |
| | 8 | |
| | 9 | |
| | 10 | |
| R-precision | 0.5 | |



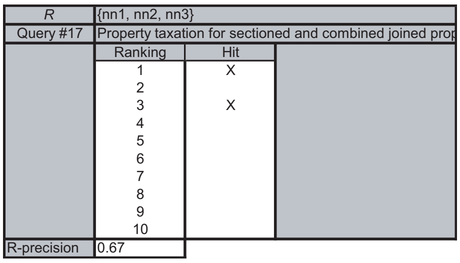
| | | |
|---|---|---|
| *R* | {nn1, nn2} | |
| Query #14 | Wildlife in traffic | |
| | Ranking | Hit |
| | 1 | |
| | 2 | |
| | 3 | X |
| | 4 | |
| | 5 | |
| | 6 | |
| | 7 | |
| | 8 | |
| | 9 | |
| | 10 | |
| R-precision | 0.0 | |



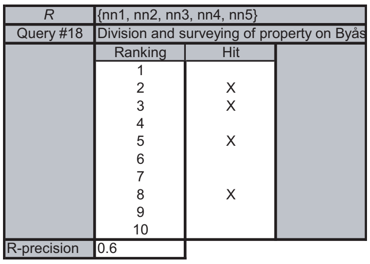
| | | |
|---|---|---|
| *R* | {nn1, nn2, nn3, nn4} | |
| Query #15 | Supervision of indoor climate in schools ar | |
| | Ranking | Hit |
| | 1 | X |
| | 2 | |
| | 3 | X |
| | 4 | |
| | 5 | |
| | 6 | |
| | 7 | |
| | 8 | |
| | 9 | X |
| | 10 | |
| R-precision | 0.5 | |



| | | |
|---|---|---|
| *R* | {nn1, nn2} | |
| Query #16 | Property sectioning with house, industry, a | |
| | Ranking | Hit |
| | 1 | |
| | 2 | |
| | 3 | |
| | 4 | |
| | 5 | X |
| | 6 | |
| | 7 | |
| | 8 | |
| | 9 | |
| | 10 | |
| R-precision | 0.0 | |



| | | |
|---|---|---|
| *R* | {nn1, nn2, nn3} | |
| Query #17 | Property taxation for sectioned and combined joined prop | |
| | Ranking | Hit |
| | 1 | X |
| | 2 | |
| | 3 | X |
| | 4 | |
| | 5 | |
| | 6 | |
| | 7 | |
| | 8 | |
| | 9 | |
| | 10 | |
| R-precision | 0.67 | |



| | | |
|---|---|---|
| *R* | {nn1, nn2, nn3, nn4, nn5} | |
| Query #18 | Division and surveying of property on Byås | |
| | Ranking | Hit |
| | 1 | |
| | 2 | X |
| | 3 | X |
| | 4 | |
| | 5 | X |
| | 6 | |
| | 7 | |
| | 8 | X |
| | 9 | |
| | 10 | |
| R-precision | 0.6 | |

### A.1.2 Based on name occurrences - Non-filtered query

| | | |
|---|---|---|
| *R* | {nn1, nn2} | |
| Query #1 | Maintenance Rosenborg public school | |
| | Ranking | Hit |
| | 1 | X |
| | 2 | |
| | 3 | X |
| | 4 | |
| | 5 | |
| | 6 | |
| | 7 | |
| | 8 | |
| | 9 | |
| | 10 | |
| R-precision | 0.5 | |

| | | |
|---|---|---|
| *R* | {nn1} | |
| Query #2 | New construction Nardo public school | |
| | Ranking | Hit |
| | 1 | 0 |
| | 2 | |
| | 3 | |
| | 4 | |
| | 5 | |
| | 6 | |
| | 7 | |
| | 8 | |
| | 9 | |
| | 10 | |
| R-precision | 0.0 | |

| | | |
|---|---|---|
| *R* | {nn1} | |
| Query #3 | Kindergarten development | |
| | Ranking | Hit |
| | 1 | X |
| | 2 | |
| | 3 | |
| | 4 | |
| | 5 | |
| | 6 | |
| | 7 | |
| | 8 | |
| | 9 | |
| | 10 | |
| R-precision | 1.0 | |

| | | |
|---|---|---|
| *R* | {nn1} | |
| Query #4 | Northern relief road | |
| | Ranking | Hit |
| | 1 | 0 |
| | 2 | |
| | 3 | |
| | 4 | |
| | 5 | |
| | 6 | |
| | 7 | |
| | 8 | |
| | 9 | |
| | 10 | |
| R-precision | 0.0 | |

| | | |
|---|---|---|
| *R* | {nn1} | |
| Query #5 | Lian regulation plan | |
| | Ranking | Hit |
| | 1 | 0 |
| | 2 | |
| | 3 | |
| | 4 | |
| | 5 | |
| | 6 | |
| | 7 | |
| | 8 | |
| | 9 | |
| | 10 | |
| R-precision | 0.0 | |

| | | |
|---|---|---|
| *R* | {nn1, nn2} | |
| Query #6 | Ground seperation of parcel in Nardo road | |
| | Ranking | Hit |
| | 1 | X |
| | 2 | |
| | 3 | |
| | 4 | |
| | 5 | |
| | 6 | |
| | 7 | |
| | 8 | |
| | 9 | |
| | 10 | |
| R-precision | 0.5 | |

| | | |
|---|---|---|
| *R* | {nn1, nn2} | |
| Query #7 | Town market | |
| | Ranking | Hit |
| | 1 | X |
| | 2 | X |
| | 3 | |
| | 4 | |
| | 5 | |
| | 6 | |
| | 7 | |
| | 8 | |
| | 9 | |
| | 10 | |
| R-precision | 1.0 | |

| | | |
|---|---|---|
| *R* | {nn1, nn2} | |
| Query #8 | Leasing of municipal ground | |
| | Ranking | Hit |
| | 1 | |
| | 2 | |
| | 3 | |
| | 4 | |
| | 5 | |
| | 6 | X |
| | 7 | |
| | 8 | |
| | 9 | |
| | 10 | |
| R-precision | 0.0 | |

| | | |
|---|---|---|
| *R* | {nn1, nn2} | |
| Query #9 | Drinking water quality | |
| | Ranking | Hit |
| | 1 | 0 |
| | 2 | |
| | 3 | |
| | 4 | |
| | 5 | |
| | 6 | |
| | 7 | |
| | 8 | |
| | 9 | |
| | 10 | |
| R-precision | 0.0 | |

| | | |
|---|---|---|
| *R* | {nn1} | |
| Query #10 | Signposting and advertising | |
| | Ranking | Hit |
| | 1 | X |
| | 2 | |
| | 3 | |
| | 4 | |
| | 5 | |
| | 6 | |
| | 7 | |
| | 8 | |
| | 9 | |
| | 10 | |
| R-precision | 1.0 | |

| | | |
|---|---|---|
| *R* | {nn1, nn2} | |
| Query #11 | Change of use at Møllenberg | |
| | Ranking | Hit |
| | 1 | |
| | 2 | X |
| | 3 | |
| | 4 | |
| | 5 | |
| | 6 | |
| | 7 | |
| | 8 | |
| | 9 | |
| | 10 | |
| R-precision | 0.5 | |

| | | |
|---|---|---|
| *R* | {nn1} | |
| Query #12 | Change of use on cottage at Lian | |
| | Ranking | Hit |
| | 1 | |
| | 2 | |
| | 3 | |
| | 4 | |
| | 5 | |
| | 6 | |
| | 7 | |
| | 8 | X |
| | 9 | |
| | 10 | |
| R-precision | 0.0 | |

| | | |
|---|---|---|
| *R* | {nn1, nn2} | |
| Query #13 | Garbage | |
| | Ranking | Hit |
| | 1 | |
| | 2 | |
| | 3 | |
| | 4 | |
| | 5 | |
| | 6 | X |
| | 7 | |
| | 8 | |
| | 9 | |
| | 10 | |
| R-precision | 0.0 | |

| | | |
|---|---|---|
| *R* | {nn1, nn2} | |
| Query #14 | Wildlife in traffic | |
| | Ranking | Hit |
| | 1 | 0 |
| | 2 | |
| | 3 | |
| | 4 | |
| | 5 | |
| | 6 | |
| | 7 | |
| | 8 | |
| | 9 | |
| | 10 | |
| R-precision | 0.0 | |

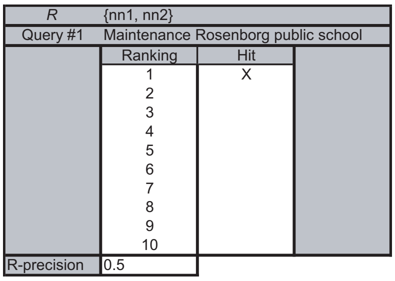
| | | |
|---|---|---|
| *R* | {nn1, nn2, nn3, nn4} | |
| Query #15 | Supervision of indoor climate in schools ar | |
| | Ranking | Hit |
| | 1 | 0 |
| | 2 | |
| | 3 | |
| | 4 | |
| | 5 | |
| | 6 | |
| | 7 | |
| | 8 | |
| | 9 | |
| | 10 | |
| R-precision | 0.0 | |

| | | |
|---|---|---|
| *R* | {nn1, nn2} | |
| Query #16 | Property sectioning with house, industry, a | |
| | Ranking | Hit |
| | 1 | 0 |
| | 2 | |
| | 3 | |
| | 4 | |
| | 5 | |
| | 6 | |
| | 7 | |
| | 8 | |
| | 9 | |
| | 10 | |
| R-precision | 0.0 | |

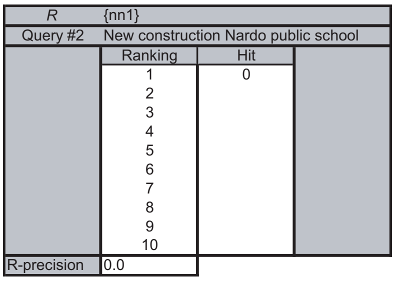
| | | |
|---|---|---|
| *R* | {nn1, nn2, nn3} | |
| Query #17 | Property taxation for sectioned and combined joined prop | |
| | Ranking | Hit |
| | 1 | X |
| | 2 | |
| | 3 | |
| | 4 | X |
| | 5 | |
| | 6 | |
| | 7 | |
| | 8 | |
| | 9 | |
| | 10 | |
| R-precision | 0.33 | |

| | | |
|---|---|---|
| *R* | {nn1, nn2, nn3, nn4, nn5} | |
| Query #18 | Division and surveying of property on Byås | |
| | Ranking | Hit |
| | 1 | |
| | 2 | |
| | 3 | |
| | 4 | |
| | 5 | |
| | 6 | |
| | 7 | X |
| | 8 | X |
| | 9 | |
| | 10 | |
| R-precision | 0.0 | |

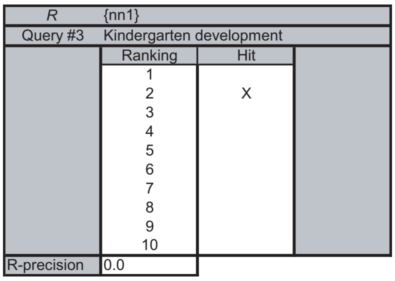
### A.1.3 Based on employment time - Filtered query



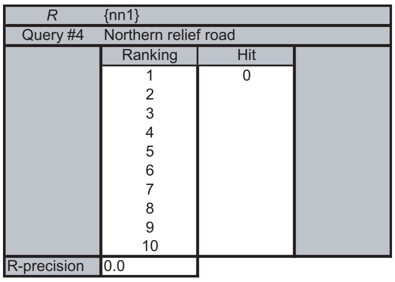
| *R* | {nn1, nn2} | |
|---|---|---|
| Query #1 | Maintenance Rosenborg public school | |
| | Ranking | Hit |
| | 1 | X |
| | 2 | |
| | 3 | |
| | 4 | |
| | 5 | |
| | 6 | |
| | 7 | |
| | 8 | |
| | 9 | |
| | 10 | |
| R-precision | 0.5 | |



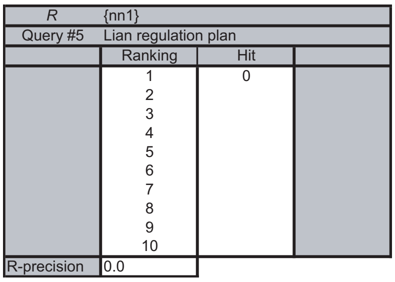
| *R* | {nn1} | |
|---|---|---|
| Query #2 | New construction Nardo public school | |
| | Ranking | Hit |
| | 1 | 0 |
| | 2 | |
| | 3 | |
| | 4 | |
| | 5 | |
| | 6 | |
| | 7 | |
| | 8 | |
| | 9 | |
| | 10 | |
| R-precision | 0.0 | |



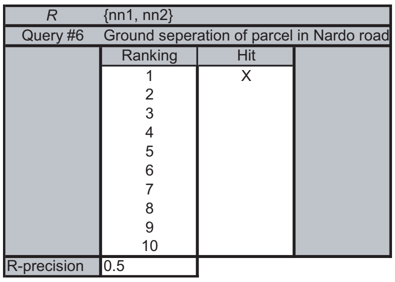
| *R* | {nn1} | |
|---|---|---|
| Query #3 | Kindergarten development | |
| | Ranking | Hit |
| | 1 | |
| | 2 | X |
| | 3 | |
| | 4 | |
| | 5 | |
| | 6 | |
| | 7 | |
| | 8 | |
| | 9 | |
| | 10 | |
| R-precision | 0.0 | |



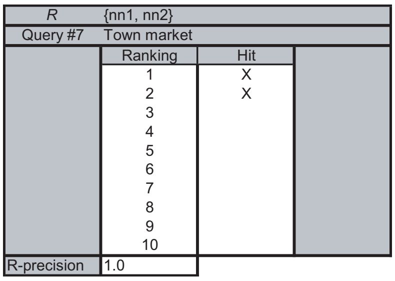
| *R* | {nn1} | |
|---|---|---|
| Query #4 | Northern relief road | |
| | Ranking | Hit |
| | 1 | 0 |
| | 2 | |
| | 3 | |
| | 4 | |
| | 5 | |
| | 6 | |
| | 7 | |
| | 8 | |
| | 9 | |
| | 10 | |
| R-precision | 0.0 | |



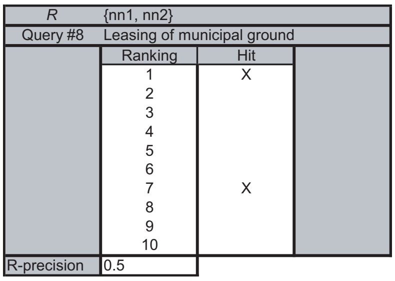
| *R* | {nn1} | |
|---|---|---|
| Query #5 | Lian regulation plan | |
| | Ranking | Hit |
| | 1 | 0 |
| | 2 | |
| | 3 | |
| | 4 | |
| | 5 | |
| | 6 | |
| | 7 | |
| | 8 | |
| | 9 | |
| | 10 | |
| R-precision | 0.0 | |



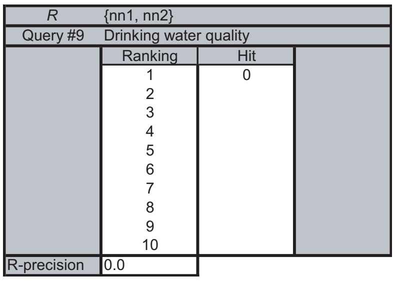
| *R* | {nn1, nn2} | |
|---|---|---|
| Query #6 | Ground seperation of parcel in Nardo road | |
| | Ranking | Hit |
| | 1 | X |
| | 2 | |
| | 3 | |
| | 4 | |
| | 5 | |
| | 6 | |
| | 7 | |
| | 8 | |
| | 9 | |
| | 10 | |
| R-precision | 0.5 | |



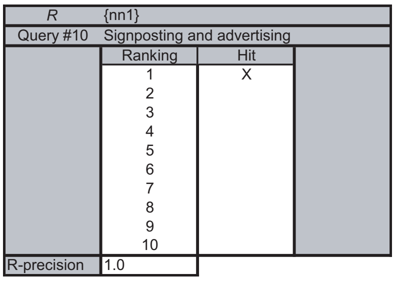
| *R* | {nn1, nn2} | |
|---|---|---|
| Query #7 | Town market | |
| | Ranking | Hit |
| | 1 | X |
| | 2 | X |
| | 3 | |
| | 4 | |
| | 5 | |
| | 6 | |
| | 7 | |
| | 8 | |
| | 9 | |
| | 10 | |
| R-precision | 1.0 | |



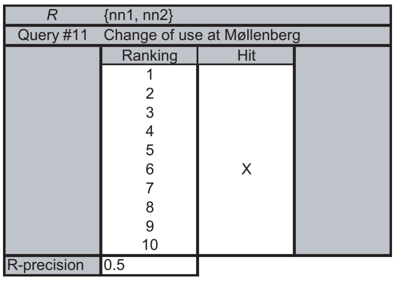
| *R* | {nn1, nn2} | |
|---|---|---|
| Query #8 | Leasing of municipal ground | |
| | Ranking | Hit |
| | 1 | X |
| | 2 | |
| | 3 | |
| | 4 | |
| | 5 | |
| | 6 | |
| | 7 | X |
| | 8 | |
| | 9 | |
| | 10 | |
| R-precision | 0.5 | |



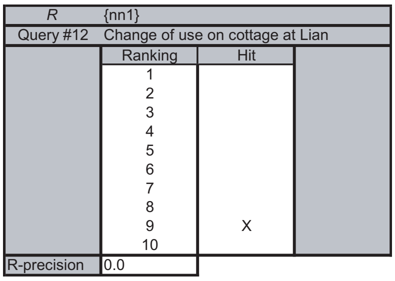
| *R* | {nn1, nn2} | |
|---|---|---|
| Query #9 | Drinking water quality | |
| | Ranking | Hit |
| | 1 | 0 |
| | 2 | |
| | 3 | |
| | 4 | |
| | 5 | |
| | 6 | |
| | 7 | |
| | 8 | |
| | 9 | |
| | 10 | |
| R-precision | 0.0 | |



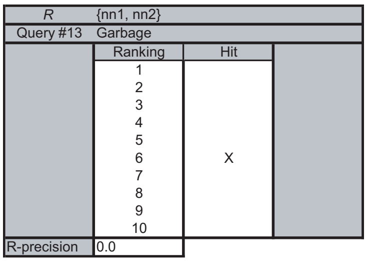
| *R* | {nn1} | |
|---|---|---|
| Query #10 | Signposting and advertising | |
| | Ranking | Hit |
| | 1 | X |
| | 2 | |
| | 3 | |
| | 4 | |
| | 5 | |
| | 6 | |
| | 7 | |
| | 8 | |
| | 9 | |
| | 10 | |
| R-precision | 1.0 | |



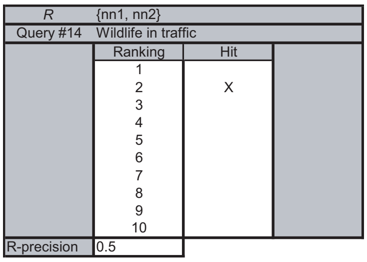
| *R* | {nn1, nn2} | |
|---|---|---|
| Query #11 | Change of use at Møllenberg | |
| | Ranking | Hit |
| | 1 | |
| | 2 | |
| | 3 | |
| | 4 | |
| | 5 | |
| | 6 | X |
| | 7 | |
| | 8 | |
| | 9 | |
| | 10 | |
| R-precision | 0.5 | |



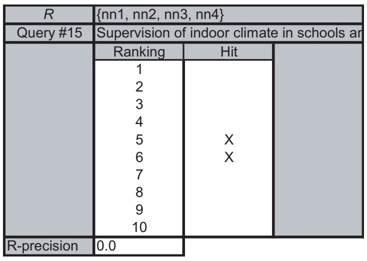
| *R* | {nn1} | |
|---|---|---|
| Query #12 | Change of use on cottage at Lian | |
| | Ranking | Hit |
| | 1 | |
| | 2 | |
| | 3 | |
| | 4 | |
| | 5 | |
| | 6 | |
| | 7 | |
| | 8 | |
| | 9 | X |
| | 10 | |
| R-precision | 0.0 | |



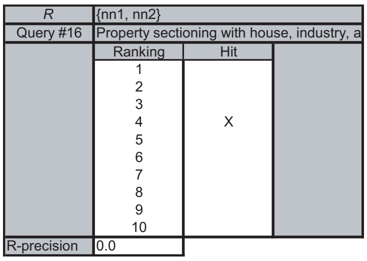
| *R* | {nn1, nn2} | |
|---|---|---|
| Query #13 | Garbage | |
| | Ranking | Hit |
| | 1 | |
| | 2 | |
| | 3 | |
| | 4 | |
| | 5 | |
| | 6 | X |
| | 7 | |
| | 8 | |
| | 9 | |
| | 10 | |
| R-precision | 0.0 | |



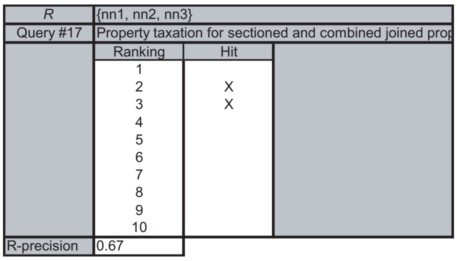
| *R* | {nn1, nn2} | |
|---|---|---|
| Query #14 | Wildlife in traffic | |
| | Ranking | Hit |
| | 1 | |
| | 2 | X |
| | 3 | |
| | 4 | |
| | 5 | |
| | 6 | |
| | 7 | |
| | 8 | |
| | 9 | |
| | 10 | |
| R-precision | 0.5 | |



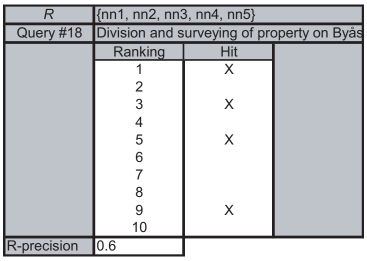
| *R* | {nn1, nn2, nn3, nn4} | |
|---|---|---|
| Query #15 | Supervision of indoor climate in schools ar | |
| | Ranking | Hit |
| | 1 | |
| | 2 | |
| | 3 | |
| | 4 | |
| | 5 | X |
| | 6 | X |
| | 7 | |
| | 8 | |
| | 9 | |
| | 10 | |
| R-precision | 0.0 | |



| *R* | {nn1, nn2} | |
|---|---|---|
| Query #16 | Property sectioning with house, industry, a | |
| | Ranking | Hit |
| | 1 | |
| | 2 | |
| | 3 | |
| | 4 | X |
| | 5 | |
| | 6 | |
| | 7 | |
| | 8 | |
| | 9 | |
| | 10 | |
| R-precision | 0.0 | |



| *R* | {nn1, nn2, nn3} | |
|---|---|---|
| Query #17 | Property taxation for sectioned and combined joined prop | |
| | Ranking | Hit |
| | 1 | |
| | 2 | X |
| | 3 | X |
| | 4 | |
| | 5 | |
| | 6 | |
| | 7 | |
| | 8 | |
| | 9 | |
| | 10 | |
| R-precision | 0.67 | |



| *R* | {nn1, nn2, nn3, nn4, nn5} | |
|---|---|---|
| Query #18 | Division and surveying of property on Byås | |
| | Ranking | Hit |
| | 1 | X |
| | 2 | |
| | 3 | X |
| | 4 | |
| | 5 | X |
| | 6 | |
| | 7 | |
| | 8 | |
| | 9 | X |
| | 10 | |
| R-precision | 0.6 | |

### A.1.4 Based on employment time - Non-filtered query

| | | |
|---|---|---|
| *R* | {nn1, nn2} | |
| Query #1 | Maintenance Rosenborg public school | |
| | Ranking | Hit |
| | 1 | |
| | 2 | |
| | 3 | |
| | 4 | |
| | 5 | |
| | 6 | |
| | 7 | |
| | 8 | |
| | 9 | X |
| | 10 | |
| R-precision | 0.0 | |

| | | |
|---|---|---|
| *R* | {nn1} | |
| Query #2 | New construction Nardo public school | |
| | Ranking | Hit |
| | 1 | 0 |
| | 2 | |
| | 3 | |
| | 4 | |
| | 5 | |
| | 6 | |
| | 7 | |
| | 8 | |
| | 9 | |
| | 10 | |
| R-precision | 0.0 | |

| | | |
|---|---|---|
| *R* | {nn1} | |
| Query #3 | Kindergarten development | |
| | Ranking | Hit |
| | 1 | |
| | 2 | |
| | 3 | |
| | 4 | X |
| | 5 | |
| | 6 | |
| | 7 | |
| | 8 | |
| | 9 | |
| | 10 | |
| R-precision | 0.0 | |

| | | |
|---|---|---|
| *R* | {nn1} | |
| Query #4 | Northern relief road | |
| | Ranking | Hit |
| | 1 | 0 |
| | 2 | |
| | 3 | |
| | 4 | |
| | 5 | |
| | 6 | |
| | 7 | |
| | 8 | |
| | 9 | |
| | 10 | |
| R-precision | 0.0 | |

| | | |
|---|---|---|
| *R* | {nn1} | |
| Query #5 | Lian regulation plan | |
| | Ranking | Hit |
| | 1 | 0 |
| | 2 | |
| | 3 | |
| | 4 | |
| | 5 | |
| | 6 | |
| | 7 | |
| | 8 | |
| | 9 | |
| | 10 | |
| R-precision | 0.0 | |

| | | |
|---|---|---|
| *R* | {nn1, nn2} | |
| Query #6 | Ground seperation of parcel in Nardo road | |
| | Ranking | Hit |
| | 1 | |
| | 2 | |
| | 3 | |
| | 4 | X |
| | 5 | |
| | 6 | |
| | 7 | |
| | 8 | |
| | 9 | |
| | 10 | |
| R-precision | 0.0 | |

| | | |
|---|---|---|
| *R* | {nn1, nn2} | |
| Query #7 | Town market | |
| | Ranking | Hit |
| | 1 | X |
| | 2 | X |
| | 3 | |
| | 4 | |
| | 5 | |
| | 6 | |
| | 7 | |
| | 8 | |
| | 9 | |
| | 10 | |
| R-precision | 1.0 | |

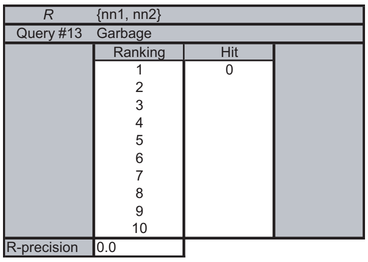
| | | |
|---|---|---|
| *R* | {nn1, nn2} | |
| Query #8 | Leasing of municipal ground | |
| | Ranking | Hit |
| | 1 | X |
| | 2 | |
| | 3 | |
| | 4 | |
| | 5 | |
| | 6 | |
| | 7 | |
| | 8 | |
| | 9 | |
| | 10 | |
| R-precision | 0.5 | |

| | | |
|---|---|---|
| *R* | {nn1, nn2} | |
| Query #9 | Drinking water quality | |
| | Ranking | Hit |
| | 1 | 0 |
| | 2 | |
| | 3 | |
| | 4 | |
| | 5 | |
| | 6 | |
| | 7 | |
| | 8 | |
| | 9 | |
| | 10 | |
| R-precision | 0.0 | |

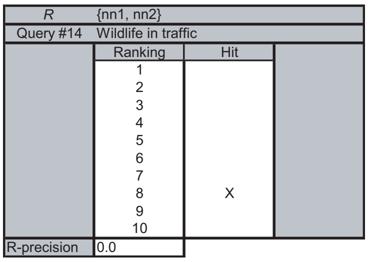
| | | |
|---|---|---|
| *R* | {nn1} | |
| Query #10 | Signposting and advertising | |
| | Ranking | Hit |
| | 1 | |
| | 2 | |
| | 3 | |
| | 4 | |
| | 5 | X |
| | 6 | |
| | 7 | |
| | 8 | |
| | 9 | |
| | 10 | |
| R-precision | 0.0 | |

| | | |
|---|---|---|
| *R* | {nn1, nn2} | |
| Query #11 | Change of use at Møllenberg | |
| | Ranking | Hit |
| | 1 | 0 |
| | 2 | |
| | 3 | |
| | 4 | |
| | 5 | |
| | 6 | |
| | 7 | |
| | 8 | |
| | 9 | |
| | 10 | |
| R-precision | 0.0 | |

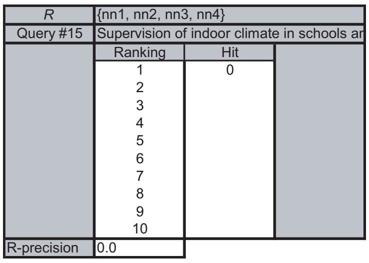
| | | |
|---|---|---|
| *R* | {nn1} | |
| Query #12 | Change of use on cottage at Lian | |
| | Ranking | Hit |
| | 1 | 0 |
| | 2 | |
| | 3 | |
| | 4 | |
| | 5 | |
| | 6 | |
| | 7 | |
| | 8 | |
| | 9 | |
| | 10 | |
| R-precision | 0.0 | |



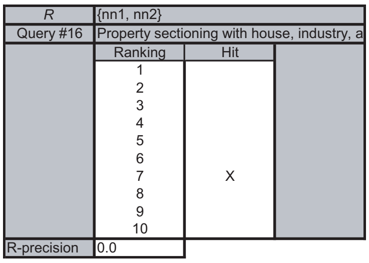
| | | |
|---|---|---|
| *R* | {nn1, nn2} | |
| Query #13 | Garbage | |
| | Ranking | Hit |
| | 1 | 0 |
| | 2 | |
| | 3 | |
| | 4 | |
| | 5 | |
| | 6 | |
| | 7 | |
| | 8 | |
| | 9 | |
| | 10 | |
| R-precision | 0.0 | |



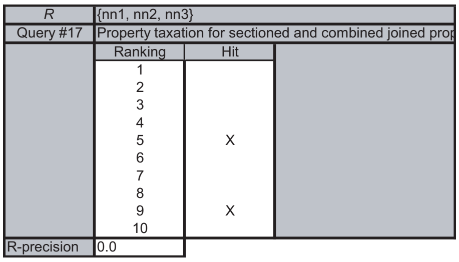
| | | |
|---|---|---|
| *R* | {nn1, nn2} | |
| Query #14 | Wildlife in traffic | |
| | Ranking | Hit |
| | 1 | |
| | 2 | |
| | 3 | |
| | 4 | |
| | 5 | |
| | 6 | |
| | 7 | |
| | 8 | X |
| | 9 | |
| | 10 | |
| R-precision | 0.0 | |



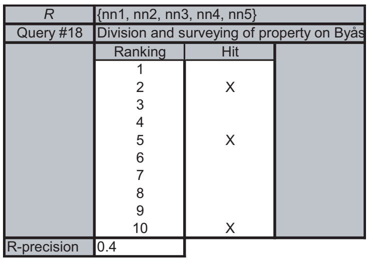
| | | |
|---|---|---|
| *R* | {nn1, nn2, nn3, nn4} | |
| Query #15 | Supervision of indoor climate in schools ar | |
| | Ranking | Hit |
| | 1 | 0 |
| | 2 | |
| | 3 | |
| | 4 | |
| | 5 | |
| | 6 | |
| | 7 | |
| | 8 | |
| | 9 | |
| | 10 | |
| R-precision | 0.0 | |



| | | |
|---|---|---|
| *R* | {nn1, nn2} | |
| Query #16 | Property sectioning with house, industry, a | |
| | Ranking | Hit |
| | 1 | |
| | 2 | |
| | 3 | |
| | 4 | |
| | 5 | |
| | 6 | |
| | 7 | X |
| | 8 | |
| | 9 | |
| | 10 | |
| R-precision | 0.0 | |



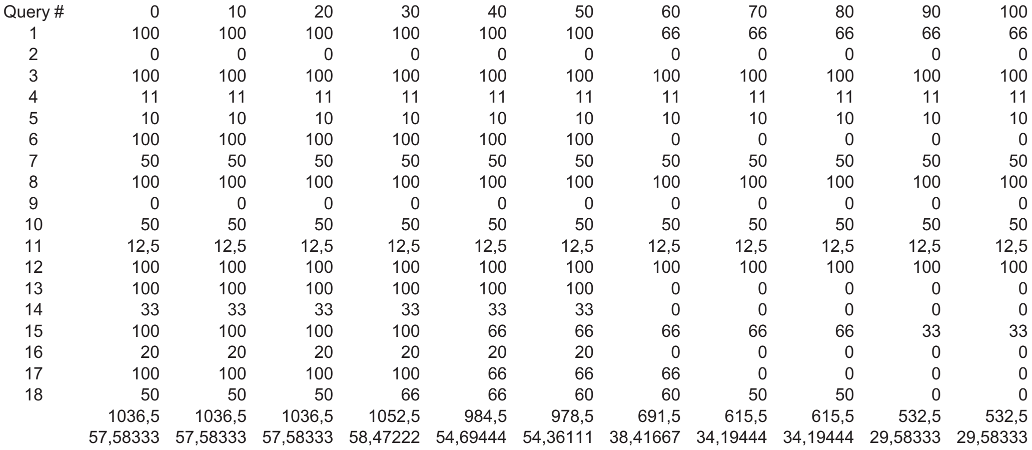
| | | |
|---|---|---|
| *R* | {nn1, nn2, nn3} | |
| Query #17 | Property taxation for sectioned and combined joined prop | |
| | Ranking | Hit |
| | 1 | |
| | 2 | |
| | 3 | |
| | 4 | |
| | 5 | X |
| | 6 | |
| | 7 | |
| | 8 | |
| | 9 | X |
| | 10 | |
| R-precision | 0.0 | |



| | | |
|---|---|---|
| *R* | {nn1, nn2, nn3, nn4, nn5} | |
| Query #18 | Division and surveying of property on Byås | |
| | Ranking | Hit |
| | 1 | |
| | 2 | X |
| | 3 | |
| | 4 | |
| | 5 | X |
| | 6 | |
| | 7 | |
| | 8 | |
| | 9 | |
| | 10 | X |
| R-precision | 0.4 | |

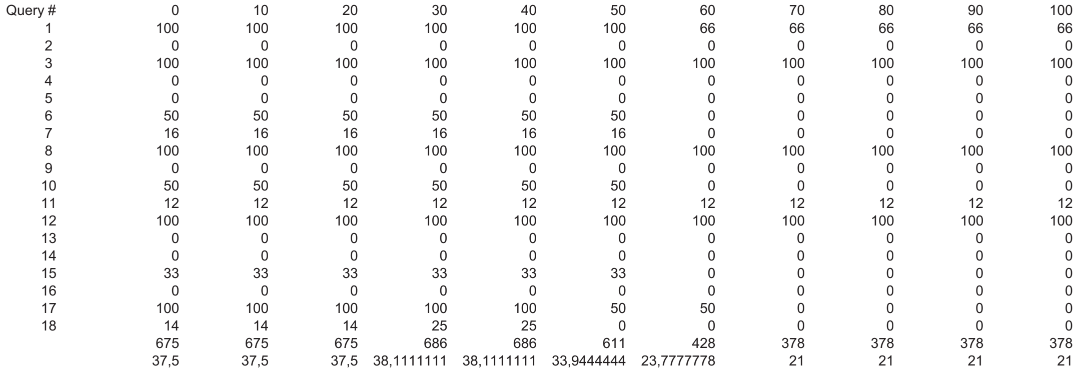
# A.2 Computing interpolated precision@recall

## A.2.1 Based on name occurrences - Filtered query



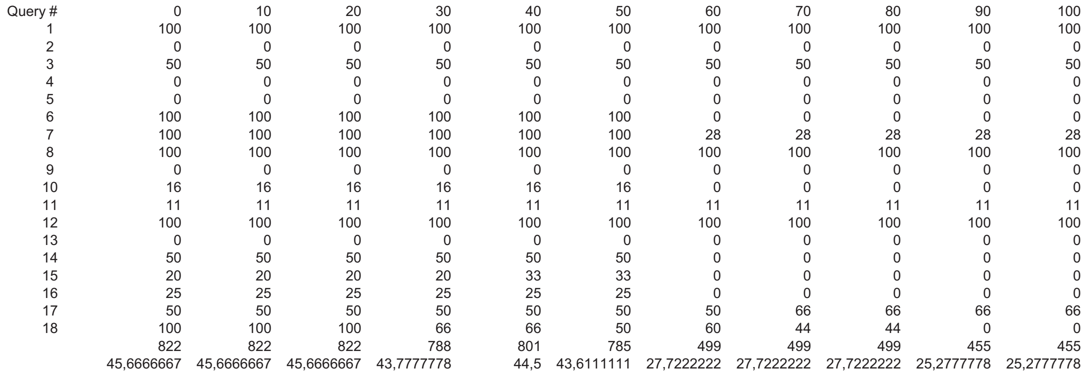
| Query # | 0 | 10 | 20 | 30 | 40 | 50 | 60 | 70 | 80 | 90 | 100 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 100 | 100 | 100 | 100 | 100 | 100 | 66 | 66 | 66 | 66 | 66 |
| 2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 3 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 |
| 4 | 11 | 11 | 11 | 11 | 11 | 11 | 11 | 11 | 11 | 11 | 11 |
| 5 | 10 | 10 | 10 | 10 | 10 | 10 | 10 | 10 | 10 | 10 | 10 |
| 6 | 100 | 100 | 100 | 100 | 100 | 100 | 0 | 0 | 0 | 0 | 0 |
| 7 | 50 | 50 | 50 | 50 | 50 | 50 | 50 | 50 | 50 | 50 | 50 |
| 8 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 |
| 9 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 10 | 50 | 50 | 50 | 50 | 50 | 50 | 50 | 50 | 50 | 50 | 50 |
| 11 | 12,5 | 12,5 | 12,5 | 12,5 | 12,5 | 12,5 | 12,5 | 12,5 | 12,5 | 12,5 | 12,5 |
| 12 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 |
| 13 | 100 | 100 | 100 | 100 | 100 | 100 | 0 | 0 | 0 | 0 | 0 |
| 14 | 33 | 33 | 33 | 33 | 33 | 33 | 0 | 0 | 0 | 0 | 0 |
| 15 | 100 | 100 | 100 | 100 | 66 | 66 | 66 | 66 | 66 | 33 | 33 |
| 16 | 20 | 20 | 20 | 20 | 20 | 20 | 0 | 0 | 0 | 0 | 0 |
| 17 | 100 | 100 | 100 | 100 | 66 | 66 | 66 | 0 | 0 | 0 | 0 |
| 18 | 50 | 50 | 50 | 66 | 66 | 60 | 60 | 50 | 50 | 0 | 0 |
|  | 1036,5 | 1036,5 | 1036,5 | 1052,5 | 984,5 | 978,5 | 691,5 | 615,5 | 615,5 | 532,5 | 532,5 |
|  | 57,58333 | 57,58333 | 57,58333 | 58,47222 | 54,69444 | 54,36111 | 38,41667 | 34,19444 | 34,19444 | 29,58333 | 29,58333 |

## A.2.2 Based on name occurrences - Non-filtered query



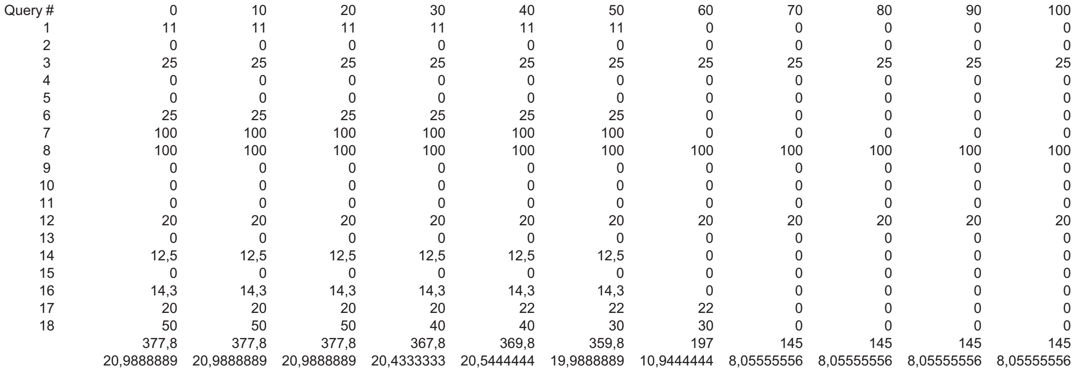
| Query # | 0 | 10 | 20 | 30 | 40 | 50 | 60 | 70 | 80 | 90 | 100 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 100 | 100 | 100 | 100 | 100 | 100 | 66 | 66 | 66 | 66 | 66 |
| 2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 3 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 |
| 4 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 5 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 6 | 50 | 50 | 50 | 50 | 50 | 50 | 0 | 0 | 0 | 0 | 0 |
| 7 | 16 | 16 | 16 | 16 | 16 | 16 | 0 | 0 | 0 | 0 | 0 |
| 8 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 |
| 9 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 10 | 50 | 50 | 50 | 50 | 50 | 50 | 0 | 0 | 0 | 0 | 0 |
| 11 | 12 | 12 | 12 | 12 | 12 | 12 | 12 | 12 | 12 | 12 | 12 |
| 12 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 |
| 13 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 14 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 15 | 33 | 33 | 33 | 33 | 33 | 33 | 0 | 0 | 0 | 0 | 0 |
| 16 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 17 | 100 | 100 | 100 | 100 | 100 | 50 | 50 | 0 | 0 | 0 | 0 |
| 18 | 14 | 14 | 14 | 25 | 25 | 0 | 0 | 0 | 0 | 0 | 0 |
|  | 675 | 675 | 675 | 686 | 686 | 611 | 428 | 378 | 378 | 378 | 378 |
|  | 37,5 | 37,5 | 37,5 | 38,1111111 | 38,1111111 | 33,9444444 | 23,7777778 | 21 | 21 | 21 | 21 |

### A.2.3 Based on employment time - Filtered query



| Query # | 0 | 10 | 20 | 30 | 40 | 50 | 60 | 70 | 80 | 90 | 100 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 |
| 2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 3 | 50 | 50 | 50 | 50 | 50 | 50 | 50 | 50 | 50 | 50 | 50 |
| 4 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 5 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 6 | 100 | 100 | 100 | 100 | 100 | 100 | 0 | 0 | 0 | 0 | 0 |
| 7 | 100 | 100 | 100 | 100 | 100 | 100 | 28 | 28 | 28 | 28 | 28 |
| 8 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 |
| 9 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 10 | 16 | 16 | 16 | 16 | 16 | 16 | 0 | 0 | 0 | 0 | 0 |
| 11 | 11 | 11 | 11 | 11 | 11 | 11 | 11 | 11 | 11 | 11 | 11 |
| 12 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 |
| 13 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 14 | 50 | 50 | 50 | 50 | 50 | 50 | 0 | 0 | 0 | 0 | 0 |
| 15 | 20 | 20 | 20 | 20 | 33 | 33 | 0 | 0 | 0 | 0 | 0 |
| 16 | 25 | 25 | 25 | 25 | 25 | 25 | 0 | 0 | 0 | 0 | 0 |
| 17 | 50 | 50 | 50 | 50 | 50 | 50 | 50 | 66 | 66 | 66 | 66 |
| 18 | 100 | 100 | 100 | 66 | 66 | 50 | 60 | 44 | 44 | 0 | 0 |
|  | 822 | 822 | 822 | 788 | 801 | 785 | 499 | 499 | 499 | 455 | 455 |
|  | 45,6666667 | 45,6666667 | 45,6666667 | 43,7777778 | 44,5 | 43,6111111 | 27,7222222 | 27,7222222 | 27,7222222 | 25,2777778 | 25,2777778 |

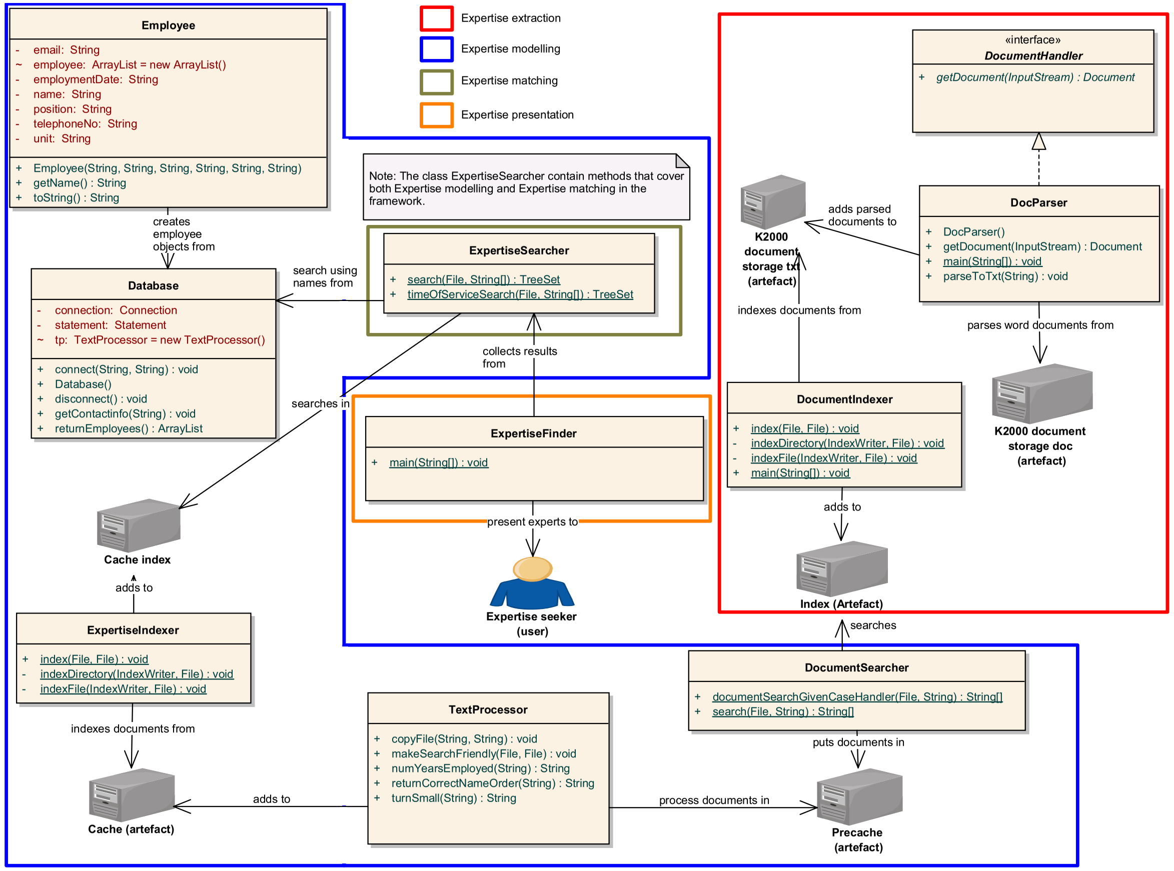
### A.2.4 Based on employment time - Non-filtered query



| Query # | 0 | 10 | 20 | 30 | 40 | 50 | 60 | 70 | 80 | 90 | 100 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 11 | 11 | 11 | 11 | 11 | 11 | 0 | 0 | 0 | 0 | 0 |
| 2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 3 | 25 | 25 | 25 | 25 | 25 | 25 | 25 | 25 | 25 | 25 | 25 |
| 4 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 5 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 6 | 25 | 25 | 25 | 25 | 25 | 25 | 0 | 0 | 0 | 0 | 0 |
| 7 | 100 | 100 | 100 | 100 | 100 | 100 | 0 | 0 | 0 | 0 | 0 |
| 8 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 |
| 9 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 10 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 11 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 12 | 20 | 20 | 20 | 20 | 20 | 20 | 20 | 20 | 20 | 20 | 20 |
| 13 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 14 | 12,5 | 12,5 | 12,5 | 12,5 | 12,5 | 12,5 | 0 | 0 | 0 | 0 | 0 |
| 15 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 16 | 14,3 | 14,3 | 14,3 | 14,3 | 14,3 | 14,3 | 0 | 0 | 0 | 0 | 0 |
| 17 | 20 | 20 | 20 | 20 | 22 | 22 | 22 | 0 | 0 | 0 | 0 |
| 18 | 50 | 50 | 50 | 40 | 40 | 30 | 30 | 0 | 0 | 0 | 0 |
|  | 377,8 | 377,8 | 377,8 | 367,8 | 369,8 | 359,8 | 197 | 145 | 145 | 145 | 145 |
|  | 20,9888889 | 20,9888889 | 20,9888889 | 20,4333333 | 20,5444444 | 19,9888889 | 10,9444444 | 8,05555556 | 8,05555556 | 8,05555556 | 8,05555556 |

# Appendix B

# Class diagram


Expertise extraction
Expertise modelling
Expertise matching
Expertise presentation
Employee
- email: String
~ employee: ArrayList = new ArrayList()
- employmentDate: String
- name: String
- position: String
- telephoneNo: String
- unit: String
+ Employee(String, String, String, String, String, String)
+ getName() : String
+ toString() : String
creates employee objects from
Database
- connection: Connection
- statement: Statement
~ tp: TextProcessor = new TextProcessor()
+ connect(String, String) : void
+ Database()
+ disconnect() : void
+ getContactinfo(String) : void
+ returnEmployees() : ArrayList
search using names from
Note: The class ExpertiseSearcher contain methods that cover both Expertise modelling and Expertise matching in the framework.
ExpertiseSearcher
+ search(File, String[]) : TreeSet
+ timeOfServiceSearch(File, String[]) : TreeSet
collects results from
searches in
ExpertiseFinder
+ main(String[]) : void
present experts to
Expertise seeker (user)
Cache index
adds to
ExpertiseIndexer
+ index(File, File) : void
- indexDirectory(IndexWriter, File) : void
- indexFile(IndexWriter, File) : void
indexes documents from
Cache (artefact)
adds to
TextProcessor
+ copyFile(String, String) : void
+ makeSearchFriendly(File, File) : void
+ numYearsEmployed(String) : String
+ returnCorrectNameOrder(String) : String
+ turnSmall(String) : String
process documents in
Precache (artefact)
puts documents in
DocumentSearcher
+ documentSearchGivenCaseHandler(File, String) : String[]
+ search(File, String) : String[]
searches
«interface»
DocumentHandler
+ getDocument(InputStream) : Document
DocParser
+ DocParser()
+ getDocument(InputStream) : Document
+ main(String[]) : void
+ parseToTxt(String) : void
adds parsed documents to
K2000 document storage txt (artefact)
indexes documents from
parses word documents from
K2000 document storage doc (artefact)
DocumentIndexer
+ index(File, File) : void
- indexDirectory(IndexWriter, File) : void
- indexFile(IndexWriter, File) : void
+ main(String[]) : void
adds to
Index (Artefact)