Domain adaptive extraction of topical hierarchies for Expertise Mining

Georgeta Bordea
Digital Enterprise Research Institute
National University of Ireland, Galway

under supervision of dr. Paul Buitelaar

2013 September

Abstract

In this age of pervasive internet access we have become accustomed to rely on web search for our most basic information needs. But complex queries in knowledge-intensive organisations, as well as in the academic environment, are still best answered by direct interaction with domain experts. Experts produce large amounts of text in their daily activities that can be analysed to automatically map expertise and provide services that allow users to search for experts instead of documents. Current approaches for expert finding are based on keyphrase search, relying on exact string matches to identify experts. What is needed instead is support for exploratory search and discovery of expertise topics and experts, and in-depth measures of expertise, that can be provided by extracting expertise topics and the relations between them.

This dissertation examines methods for extracting knowledge structures from text and their application to expert search. Towards this goal, we introduce a novel methodology called Expertise Mining, that provides solutions for expertise topic extraction, expert profiling and expert finding through text analysis. In particular, we propose a term extraction approach that considers the level of specificity of a term within a domain, as a solution for expertise topic extraction. We investigate relations between expertise topics, proposing a high-coverage method for topical hierarchy construction based on a global generality measure and a graph-based algorithm. We show that topical hierarchies can be used to improve expert finding, by measuring how well an individual covers the subtopics of a field. Additionally, automatically extracted expertise topics are used to construct expert profiles that provide context to the expertise of a person.

This work has been part of the Saffron project, at the Digital Enterprise Research Institute (DERI), NUI Galway. The Saffron system currently provides insight into different Computer Science domains and was deployed at several conferences as a tool for finding collaborators.

Contents

List of Figures

LIST OF FIGURES

List of Tables

1

Introduction

1.1 Motivation

Leveraging the knowledge of experts is essential in organisations, scientific research and online communities, in a multitude of scenarios and settings. These scenarios include finding consultants and collaborators inside or outside an organisation or community, locating topical experts for requests from the media [HBBdR10], discovering solvers in open innovation platforms [SJL11], and finding qualified reviewers to assess the quality of research submissions [MM07a, RB08], to name just a few. Identifying, measuring, and representing expertise has the potential to encourage interaction and collaboration, and ultimately knowledge creation, by constructing a web of connections between experts and the knowledge that they create. These connections allow individuals to access knowledge beyond their tightly-knit social networks, where members have access to similar information [RSN96]. Additionally, this web of connections can accelerate expertise development by providing valuable insight to outsiders and novice members of a community.

Academia and industry separately devised ways to encourage collaboration beyond narrow confines of increasing specialisation. In academic communities, formal events such as conferences are organised to allow more frequent interactions among the members of a research community. The purpose of these regular meetings is to enhance knowledge creation by fostering a shared common language and a shared understanding. Conferences are seen as platforms for finding like-minded researchers and collaborators through direct interaction. But scientific publications describe mostly focused results, providing little context about the general expertise of their authors. Gather-

ing and summarising expertise information from previous editions of a conference as well as from other contributions can enable weak ties and reinforce close ties between participants. At a national scale, research funding agencies are interested in keeping track of experts and expertise areas, to increase knowledge transfer from academia to industry. In both cases, a system that provides facilities for search and exploration of experts and expertise and brief summaries of a person's interests and knowledge in the form of expert profiles can prove to be an adequate solution.

In the enterprise environment, the broad area of knowledge management addresses the need for expert search solutions through competency management. Organisations have an interest in providing access to tacit knowledge acquired by individuals, bringing together experts to ensure knowledge is shared and increased. The success of an organisation depends not only on the individual skills and competencies of their members but also on the way they collaborate and take advantage of their different areas of expertise. A system that allows users to identify experts can increase efficiency, competitiveness, and innovation within the company. Competency management is concerned with profiling human resources by relying on self-provided assessments about expertise and performance information collected from supervisors and peers. This resulted in the creation of several international specifications for competency description (e.g., IEEE RCD, 2004; HR-XML, 2006). The solution proposed in this thesis can be used as a tool for extending these standards to other domains or with other competencies.

Constructing and applying knowledge structures to a real-world application such as Expertise Mining, touches on research problems of broader interest. The Semantic Web uses formal ontologies as a key instrument in order to add structure to data, but building domain-specific ontologies is still a difficult, time consuming and error-prone process since most information is currently available as free-text. Therefore, the development of fast and cheap solutions for ontology learning from text is a central concern for the Semantic Web community. Recent years have seen a massive growth in the number and scale of semantic datasets openly available on the Linked Open Data cloud, but to cite a recent work [Nav12], "much work is still needed to prove that a proper injection of semantics into real-world applications is always beneficial".

Ontologies can not be simply evaluated through user studies, because they are relatively complex structures. Usually only a limited number of concepts and relatively simple structures can be evaluated through user studies due to inherent time

constraints. But domain ontologies are usually designed with a specific application in mind, and can be indirectly evaluated by the improvements they bring for a given task. An application-driven approach to ontology evaluation mitigates the limitations of manual evaluation performed by domain experts, including high costs and subjectivity. Ontology development is primarily concerned with the definition of concepts and relations between them. Therefore, two fundamental research problems in ontology learning from text are the extraction of concepts, and the automatic construction of domain taxonomies, which form the backbone of any ontology. Several directions for ontology learning from text have been explored, but large scale success has been hindered by a lack of common evaluation standards and evaluation datasets. In this thesis, we assess the status of existing methods for concept extraction and taxonomy construction through their application to Expertise Mining. This is an application area which was not previously considered for an application-based evaluation of ontologies.

1.2 Challenges

Information about experts can be gathered during the employment process and through self-assessment, but a person's skills, knowledge and capabilities frequently change over time. To add to the problem, manually collected expert profiles are time-consuming and expensive to build, and they require frequent updates. The solution is to automatically identify experts and expertise topics to reduce the human effort required for competency management. Relatively inexpensive solutions emerged that regard user generated content as a reliable indicator of interest or expertise. In their daily activities, individuals regularly produce large amounts of text both online and offline. This content can be used to automate the extraction of expert profiles and to provide search functionalities similar to the ones available for document search.

Although automatic approaches based on by-products of expert activity have clear advantages in dynamic environments over static expert profiles, they are still restricted in several ways. First, these solutions are geared towards finding experts for well defined information needs, through **keyphrase search**, and less towards **discovery** and **exploration** of expertise. Keyphrase-based search may be sufficient for users that have a clear information need and a good domain understanding, but it is less appropriate for novice users or for users outside the community. Another limitation is that keyword

search is useful for searches that have an obvious solution, but it is less likely that new knowledge can be acquired in this way. Providing users with comprehensive lists of main expertise topics, auto-complete functionality and hierarchical structures, is a possible solution to these problems, as we will see in the following chapters.

Second, most data-driven approaches for expert profiling still require additional knowledge about a domain in the form of **controlled vocabularies** of expertise topics [BdR07, STV+11]. This limits their applicability to domains where such a controlled vocabulary is available or in cases where sufficient resources can be allocated to construct one.

Third, data-driven approaches focus mainly on relating an individual to a given domain of expertise, and not on assessing their **proficiency level** in a particular domain. Working with explicit representations about users and hand-crafted domain models, knowledge repositories overcome these problems by several means. On one hand, they provide users with **hierarchical structures** that allow them to browse through various knowledge areas and to get brief overviews of a domain. In this way, taxonomies and knowledge maps enable users to rapidly locate the knowledge or the individual that has the needed knowledge. On the other hand, expertise topics and expert profiles are explicitly gathered and represented. Also, knowledge-based solutions make use of several **performance indicators**, including frequency, scope, autonomy, complexity, or context of use [Paq07] to assess expertise.

The challenge is to automatically construct knowledge structures from text and to use them to inform existing data-driven approaches, combining the advantages of both approaches and enabling **exploratory search** of experts and expertise. In this way, information access is improved by providing multiple ways to access expertise information.

1.3 Expertise Mining

An **expert** is generally defined as a person that has a great deal of knowledge or skill in a particular field. This knowledge can be acquired as part of a formal training as well as through practical experience. This definition restricts the notion of expertise to a specific domain, because no expertise is universally applicable. Also, the exact amount of knowledge that is required to be qualified as an expert varies depending on

the subject area. In this work, we start from the assumption that expertise is relative to a given domain, but we focus on methods that are portable across domains.

Following [Paq07], we define **expertise** as the highest performance level that a person can achieve for a given competency. Quantifying the expertise of a person is not an easy task, but usually expertise is analysed in the context of an organisation or community. In this setting, an expert is the person that has the most knowledge about a topic among its peers. Expertise can also be measured in extrinsic ways, through the judgement of peers. This can be gathered either directly through interviews or indirectly through citations, for example. In this thesis, we take the intrinsic approach, mainly analysing the documents produced by a person. Our assumption is that documents authored by a person are a representative reflection of their knowledge and thus their expertise. People can also convey their knowledge by phone or through direct interaction but we will only consider knowledge derived from written documents. Therefore, the models proposed in this thesis are an approximation of expertise, as documents partially capture the knowledge of an individual.

Expertise topics are a subset of set of terms in a domain, which have an appropriate level of specificity. The specificity level of a term, also called generality [Zha98, AW02] or abstractness [BKH+11], is a relevant characteristic of terms for expert search, as we will see in Section 3.1. Existing methods for term extraction are adapted for identifying terms of an intermediate level of specificity in Chapter 3. We address the problem of extracting terminology of an intermediate level of specificity, targeting terms that are specific to a domain but broad enough to be usable for summarisation or classification. This category of terms is particularly relevant for constructing expertise profiles that are concise but that have a good coverage of the main areas of expertise covered by a person.

The work presented in this thesis is motivated by several studies from the field of **cognitive psychology**. Research in this area studies information processing abilities of experts by comparing them with novices on tasks such as chess, symbolic logic, or algebra like puzzles [NS+72]. One of the conclusions of this study is that the main difference between chess masters and novice chess players does not come from pure memory capacity but rather from perceptual abilities and knowledge organisation. Experts tend to use more frequently concepts that are positioned at lower levels in a domain hierarchy than novices [TT91]. While it is common for a novice to refer to a research field

using a generic name, an expert would rather mention specific areas of interest. For example, a novice would generically refer to different technologies and techniques from computational linguistics as "natural language processing", while an expert would mention subtopics such as "lexical semantics", "word sense disambiguation", or "semantic parsing". This finding motivates our assumption that term specificity is an indicator of a person's expertise in a domain. But high-coverage domain hierarchies are not always available and are expensive to construct, therefore automatic approaches that derive such structures from text are needed.

The first approaches to expert search, based on knowledge repositories, considered the construction of **knowledge taxonomies** as a central issue in the development of an expert search system [BF00]. In this way, users were able to self-assess their skills more easily against a predefined hierarchy and to browse through a large number of knowledge areas. Data-driven expert finding techniques, that exploit expertise information from documents, benefit as well from using taxonomical relations between expertise topics [CAMA$^+$10, BBA$^+$07]. This type of structured knowledge can be used to measure expertise at different levels of granularity, through inexact matches of expertise. At the same time, taxonomies can inform the construction of complete and concise expert profiles, at the right level of specificity.

In order to address the challenges related to identifying and measuring expertise, we propose an approach that extracts expertise topics and the relations between them from domain corpora. We call this approach **Expertise Mining**, because we rely on text mining techniques to extract information about expertise from text. The information about expertise topics and their relations is further used to derive novel expertise measures for expert finding and to construct expert profiles without the need for controlled vocabularies. In this way, we build upon relevance-based assessment of expertise by making explicit knowledge structures of individuals from the documents they author. While information retrieval approaches for expert finding take a document centric approach, the Expertise Mining approach is topic centric, emphasising the identification of expertise topics.

To sum up, in this work we investigate the automatic construction of concept hierarchies with a focus on application-driven evaluation, in the context of Expertise Mining tasks such as expertise topic extraction, expert profiling, and expert finding.

1.4 Scope of research

Several broad categories of expertise can be identified including physical expertise, cognitive expertise, and social expertise [FDW06]. Physical expertise refers to psychomotor skills related to practical activities, for example the type of expertise needed to play golf. Cognitive expertise refers to knowledge of a domain, while social expertise is knowledge of a social network. Cognitive expertise can be derived from documents produced by individuals, while social expertise can be mined from social networks, including citation networks, co-authorship networks, or email networks [ZTL07]. Depending on the domain of activity, organisations have interest in one or more types of expertise. This thesis investigates expertise in knowledge-intensive organisations through text analysis, therefore we limit ourselves to expertise about domain knowledge.

Our study focuses on two main aspects: the extraction of expertise topics, and the construction and application of topical hierarchies to Expertise Mining. With respect to the extraction of expertise topics, we investigate methods to identify appropriate descriptors of expertise in a domain, at the right level of specificity. The appropriate level of specificity depends on the level of expertise of the person that is searching for an expert as well, but this direction of research is beyond the scope of the work discussed here. It is assumed that expertise topics are specific enough to have a single sense within the domain, therefore sense disambiguation was not regarded as a main concern.

With respect to the investigation of relations between expertise topics, we limit ourselves to the analysis of hierarchical relations, without addressing other types of relations as available in a full fledged ontology. As such, this work is a first step towards applying knowledge structures for mapping expertise. We mainly focus on humans as consumers of the extracted information about expertise, but scenarios where machines directly consume this information can also be envisioned. This has obvious implications on how expertise is represented, but we leave this direction of research for future work.

In a real-life application, it is not enough to provide accurate algorithms for expert finding and expert profiling. A prerequisite of Expertise Mining is the use of unique and unambiguous identifiers for individuals. Several solutions are already put in place in most organisations, using emails and URIs, therefore we do not attempt to solve this problem here. Once the experts are identified, additional information about how

to contact the expert has to be provided. Also, for an expert search tool to enable collaboration without detrimental effects on productivity, several other constraints have to be accommodated. These constraints are related to the availability of an expert at a given time, their workload, their ability and willingness to share information, as well as other factors such as possible conflicts of interest. This work addresses Expertise Mining in abstraction of these restrictions. The proposed techniques are intended to be useful tools for knowledge management that have to be further integrated in existing information systems.

1.5 Research questions

The main research question that guides this thesis is the following:

RQ 1 *Can domain knowledge be modelled and extracted from text to effectively mine expertise?*

In this work, domain knowledge is represented as expertise topics and hierarchical relations between these topics, which are modelled as a topical hierarchy. This general research question can be detailed with several more specific research questions, related to the extraction of expertise topics, the construction of topical hierarchies, and the application of topical hierarchies for expert finding and expert profiling. The following research questions related to expertise topic extraction are investigated in Chapter 3:

RQ 1.1. *What are effective ways to automatically construct a domain model?*

Domain knowledge can be modelled with various degrees of sophistication, from simple vectors of domain-specific words to complex ontologies. In this thesis we investigate methods to automatically construct a domain model from text, in the form of a vector of words. To answer the research question **RQ1.1** we introduce the notion of domain coherence and we propose a novel method to construct a domain model in Section 3.3. We compare this approach against several other approaches based on subsumption, traditional term extraction techniques, and latent semantics (Section 3.7.1).

Our assumption is that such a domain model can be used to inform term extraction, leading us to the following research question.

RQ 1.2. *How can existing term extraction techniques be adapted for expertise topic extraction using a domain model?*

Not every term in a domain is an appropriate expertise topic, therefore existing term extraction techniques have to be adapted for Expertise Mining. We identify specificity level as a main distinction between terms and expertise topics. Ideally, the proposed approaches should be applicable to a large number of domains, which is the topic of the next research question.

RQ 1.3 *Is the expertise topic extraction approach portable over different domains?*

We consider several domains from areas as varied as Computer Science, Food and Agriculture, and Biomedicine, to investigate weather domain modelling can be applied to term extraction across multiple domains. Research questions **RQ 1.2** and **RQ 1.3** are answered through a series of experiments that compare existing term extraction approaches with the extraction approach proposed in this work in Sections 3.4 and 3.7, respectively.

In Chapter 4 we answer research questions related to the construction of topical hierarchies and their application to Expertise Mining. The main research question addressed in this chapter is:

RQ 2 *Are automatically-extracted topical hierarchies useful for Expertise Mining?*

Although recent approaches for building taxonomies from text can construct a domain taxonomy from scratch [KH10, NVF11], these methods have a low coverage of technical terms, which are central in Expertise Mining. For this reason, we consider instead a more relaxed definition of hierarchical relations, in the form of topical hierarchies. We define a topical hierarchy as a hierarchy of expertise topics, where links between nodes represent *broader-narrower* relations between expertise topics. We show that topical hierarchies are more appropriate in this application context, and we propose an algorithm for constructing topical hierarchies from text.

More specifically, we answer the following questions:

RQ 2.1 *Are term extraction techniques suitable to extract expertise topics?*

This research question is addressed in Section 4.4.1, by making use of a dataset of workshop descriptions which are manually annotated with expertise topics.

RQ 2.2 *How can expert profiles be extracted from text, without relying on controlled vocabularies?*

Previous approaches for expert profiling [BdR07] require a controlled vocabulary of expertise topics, but this is not always available and is expensive to construct and main-

tain by hand. In Section 4.2.1, we show that expert profiles can be constructed using automatically extracted expertise topics, avoiding the need for a predefined controlled vocabulary.

RQ 2.3 *Are topical hierarchies useful for improving expert finding results?*

Taxonomical relations between expertise topics have been used for expert finding [CAMA+10, BBA+07], but previous work relied on manually constructed hierarchies. In this thesis we investigate whether automatically-constructed topical hierarchies can be used to improve expert finding results. The dataset of workshop descriptions is again used to analyse the RQ 2.2 and RQ 2.3 questions in Sections 4.4.3 and 4.4.4.

1.6 Main contributions

The main contribution of this work is a novel approach for expert search that makes use of knowledge extracted from text to find and profile experts within a domain. This approach is exploited to develop Saffron [1], an expert search system that allows exploratory search and discovery of expertise. Towards this goal, we investigate the following Expertise Mining tasks: expertise topic extraction, expert profiling and expert finding. In particular, our contributions include three main contributions and several supporting contributions, as follows:

1. **Term extraction considering the level of specificity of terms**, focusing on terms that concisely describe and summarise expertise areas;

    - a method to construct a domain model that is further used to measure the coherence of a term within a domain, in Chapter 3;
    - application-based evaluation of term extraction systems across multiple domains, reusing datasets gathered for keyphrase extraction, index term extraction, and semantic corpus annotation, in Chapter 3.

2. **Graph-based algorithm for constructing topical hierarchies**, that can be used to construct hierarchies from scratch, relying on domain corpora alone, and that can be applied in technical domains;

[1] Saffron:`http://saffron.deri.ie/`

- a global generality measure that makes use of co-occurrence information with other domain terms and that allows us to compare the generality of any pair of terms, in Chapter 4.

3. **Novel measure of expertise based on a topical hierarchy**, that estimates the knowledge of an expert based on how well they cover subordinate expertise topics;

   - an evaluation approach based on information about workshops and committee members, that exploits human-assessed expertise in a peer-review setting, in Chapter 4;
   - a user interface for exploratory search of experts and expertise using topical hierarchies, in Chapter 5.

1.7 Thesis outline

The rest of the thesis is organised as follows. Chapter 2 gives a broad overview of existing work in several research areas that contribute to Expertise Mining from different directions. These areas include knowledge management, information retrieval and natural language processing. We discuss the challenges that knowledge management approaches and information retrieval approaches face and we discuss how natural language processing, and in particular ontology learning from text, can solve them.

The next two chapters describe the core contributions of this thesis, starting with the extraction of expertise topics in Chapter 3, and continuing with the construction and use of topical hierarchies for Expertise Mining in Chapter 4. Each of these core chapters is organised around a first part that describes the problem at hand and the proposed approach, followed by two sections that discuss the experimental setting and results. In Chapter 3 we give several considerations on term specificity, then we introduce and define domain models and we propose a method to extract them from domain corpora using domain coherence. We propose two methods that make use of a domain model for term extraction, a method that uses a domain model to measure the coherence of a term within a domain and a measure that constructs extraction patterns. Using DBpedia as an entry point in the Linked Open Data Cloud, we semantically ground expertise

topics by associating them with a concept URI. Through a series of experiments we evaluate domain models and their application in term extraction.

Chapter 4 investigates hierarchical relations between expertise topics and their application to Expertise Mining. We discuss why available taxonomies are not appropriate for the task at hand and we propose a global generality measure based on some of the insights from Chapter 3. Additionally, we extend an existing algorithm for taxonomy construction to build topical hierarchies using the global generality measure. Extracted expertise topics are used to construct expert profiles making a trade-off between the quality of an expertise topic and its relevance for an individual. We introduce Area Coverage, a measure of expertise based on a topical hierarchy. These approaches are evaluated in comparison with a language modelling approach for expert finding.

In Chapter 5 we describe the Saffron system and we present several applications of the techniques introduced in this thesis. We show how topical hierarchies can be applied to investigate a multidisciplinary research community, to enrich Information Retrieval experimental data, and how Expertise Mining can be integrated in a system for Enterprise Content Management. Finally, in Chapter 6 we recall the main findings of this research, we discuss some of the limitations of our work and we give several directions for future work.

2

Background

This chapter is concerned with the review, interpretation and synthesis of several research areas relevant to Expertise Mining including knowledge management, information retrieval, natural language processing, and ontology learning. The chapter is organised in two main sections, a first section discussing previous work on expert search (Section 2.1) and a second section describing text mining approaches that can be applied to identify expertise from text (Section 2.2). In the second part of this chapter we present several text mining techniques for term extraction (Section 2.2.1) and keyphrase extraction (Section 2.2.2) and we discuss how they can be applied to automatically discover expertise topics. Furthermore, in Section 2.2.3, we describe previous work on discovering and organising relations between concepts, that can be applied to model expertise structure. This chapter is concluded with a summary in Section 2.3.

2.1 Expert search

In this section we provide an overview of previous approaches for expert search. Historically, skill management systems, described in Section 2.1.1, were the first solutions for expert search, designed as database applications that allow a user to search database entries about expertise [YK00]. The main focus of these approaches is to integrate databases spread across the organisation, using dissimilar schemas into one data warehouse that can be mined for information about experts and expertise. One of the main challenges that these systems have to face is that skills databases are expensive to maintain [BF00] and people tend not to update their profiles, which become quickly outdated. Additionally, user requirements tend to be fine-grained and specific, while

descriptions of expertise are more generic, rendering the skill registries unsuitable for the type of requests that an expertise search system has to answer [KS96].

A solution to these problems is to adopt a data-driven approach, developed by the information retrieval community for document search, which avoids entirely the need for identifying expertise topics (Section 2.1.2). Although this works well enough for the task of expert finding, i.e., ranking people using a given topic as a query, the same cannot be said about the complementary task of building summaries of people's knowledge, the expert profiling task. An alternative direction is to model content using topic models which are discussed in Section 2.1.3. Recent advances in expert search either focus on algorithmic aspects of locating experts and maintaining expert profiles, or investigate the more complex problems of expertise management within the area of knowledge management, through ontology-based methods for competence and expertise modelling, as we will see in Section 2.1.4.

2.1.1 Competency management

While expertise refers to proficiency at its highest level, competence is a more generic term that identifies any skill or capability, regardless of the performance level. For example, Paquette [Paq07] proposes four performance levels for competencies that include awareness, familiarity, mastery, and **expertise**, while other works [FZFY12] consider levels such as: novice, advanced beginner, competent, proficient, and **expert**. In both of these frameworks expertise is cited as the highest level on the competency scale. Our main interest in the area of competency management is related to the extensive analysis of performance indicators that can be applied to objectively measure expertise based on content. Several performance indicators are quantified in Chapter 4, where we discuss content-based measures of expertise such as experience, relevance, and area coverage.

Competency management is a research topic from the area of knowledge management that is concerned with identifying and modelling competencies in an organisation to meet personnel selection needs according to strategies and organizational priorities [DM06]. Providing overviews of competencies, as well as competence profiles for people, groups, and communities, is considered beneficial when developing strategies to bridge the gap between existing competencies and competency requirements. Competency management is regarded as a useful tool in human resource management across

all functional domains [PS11]. Identifying required competencies is essential in staffing to guide job descriptions, as well as for personnel development and for planning career paths. At the same time, objective performance metrics play an important role in performance management and compensation.

Competency models are typically constructed using static skill dictionaries and by identifying and verifying competencies through surveys, interviews, or focus groups. This approach has obvious limitations as competencies evolve with time and profiles becomes quickly outdated due to lack of time or commitment to update them. Therefore, there is a need for automatic ways to identify competencies from organisation repositories, web pages, or scientific publications.

A first solution in this direction is the work presented in [ZGU+05], where relations between people and skills are extracted as a network of associations, both people and competencies being handled as entities. The web pages of an organisation are crawled and analysed by a named entity recognition system, and then co-occurrence and frequency data is processed to assign the relation strength between named entities. Similar to data-driven approaches presented in the following section, this method is based on user-provided keyphrases and does not address the task of expert profiling. The closest approach to our method for expertise topic extraction presented in Chapter 3 is the approach discussed in [BE08]. The authors rely on a small number of manually defined linguistic patterns, that are specific to scientific publications and are not easily adapted to other domains. We propose an algorithm to automatically extract a domain model that can be used to construct extraction patterns. In this way, our approach is more portable, without requiring any human intervention.

2.1.2 Information retrieval approaches

In an organisational setting, Intranet documents can be used as evidence of expertise, assigning to each expert an aggregated document that includes all their associated documents [CHVW01]. The proposed system matches the queries submitted by users against this representation and retrieves the associated expert. Typically, it is only an organisation's documents that are analysed, but an aggregation model that assembles various kinds of information such as personal descriptions, related documents and similar people is described in [ZCMZ09]. This work defines a multinomial probability

distribution over the different sources of information. Other sources of expertise evidence can be considered such as the web (e.g. web search, news search, academic search) [SRH08], online encyclopedias [DFI+08], social networks [CMCD03], forums [ZAAN07] or the semantic desktop [DN08]. Going beyond the enterprise search use case, expert finding can be applied in academia for assigning reviewers to papers [MM07b] or on the blogosphere for finding credible bloggers [NZI+09, BdRW08].

Although the expert finding task is quite different from traditional document search, expert finding can be modelled as an information retrieval task by using the topics of interest as a query and performing a full text search for experts instead of documents. The goal of the search is to create a ranking of people who are experts in a given topic, instead of ranking relevant documents. Information retrieval approaches for expert search, generally called expertise retrieval, circumvent the need for assessing the knowledge of an individual by relying instead on the notion of relevance as defined for document search. In this scenario, the system does not identify beforehand all knowledge areas. Therefore the users are not able to browse through different knowledge areas, they are only able to search for experts for a specified topic.

A large body of work is encouraged by the Text REtrieval Conference (TREC)[1] with the introduction of the expert finding task [CdVS06, SdVC07, BCdVS07], providing common evaluation procedures, as well as comparisons and benchmarking of a large number of systems and techniques. The TREC conference supports research in the information retrieval community by annually organising worldwide experimental evaluations since 1992. One of the challenges specific to the datasets proposed by the TREC conference is to infer the association between a person and a supporting document using the occurrence of a person name in text. Although this is a relevant issue in a practical setting, the disambiguation of personal names is beyond the scope of this work. Instead, we assume that the authors of a document are unambiguously identified.

In this context, a lot of effort has been put into formally modelling the expert finding task, with the majority of the participants opting for statistical language modelling. Language modelling techniques were first proposed for speech recognition, to predict the utterance of a word given the previously uttered words. This approach was later successfully applied to ad-hoc document retrieval [PC98], and has become a widely

[1] http://trec.nist.gov/

accepted retrieval model. Instead of measuring relevance, as in traditional information retrieval, language modelling answers the question of how likely it is that a document would produce a query.

The methods proposed by the participants mainly fall in two categories, the candidate model, also called the query-independent approach in [PC06], and the document model, which takes a query-dependent approach. The candidate model builds a textual representation of each candidate expert and ranks them based on the given query using traditional retrieval models. This model is equivalent to the approach used in [CHVW01]. The alternative approach, i.e., the document model, starts by searching relevant documents and then locates people that are associated with them. Both approaches are formalised and compared in [BdRA06], with the document model outperforming the candidate model. The document model is preferable not only for performance reasons, but also for practical reasons as it does not require a separate document index and it can be implemented using a standard search engine. A person-centric model that combines features from both approaches is described as well in [SH08]. As an alternative to the principle of summing relevance scores applied by the two most popular models for expertise retrieval described above, the work presented in [MO06] analyses a number of data fusion techniques including summing reciprocal ranks of documents or using the number of documents that mention the person at the beginning of the document.

Although information retrieval methods use well founded theoretical models and a principled evaluation on common datasets they rely on a simple string based identification of people and they analyse only unstructured data to find exact matches in text documents. Our approach addresses these limitations by using context to identify domain-specific expertise topics as discussed in Chapter 3, and by building a knowledge map of expertise topics presented in Chapter 4.

Building expert profiles

A user interested in contacting an expert needs much more information than a ranked list of people names to make an informed decision about who is the most suitable person to answer their questions. More contextual information is needed, as well as evidence of expertise and explanations that make the recommendations of the system more transparent and trustworthy. Topical profiles should be considered among these

sources of contextual information, as they provide a concise description of areas in which the people are knowledgeable, putting expertise in context. Therefore, expert profiling complements expert finding by summarising the areas of skills and knowledge of a person. Competency management approaches explicitly identify and store expertise topics and expert profiles, but information retrieval approaches mainly focus on the expert finding task.

With a few exceptions [BdR07, STV+11], expert profiling received considerably less attention from data-driven approaches for expert search compared to the expert finding task. The importance of expert profiling in developing an expert finding system is discussed in [BdR07], without addressing the problem of discovery and identification of knowledge areas. Two profiling methods are proposed for measuring a person's competency when the expertise areas are known in advance. The first method, which achieves the best results according to their experiments, represents a person's skills as a score over documents that are relevant to a knowledge area. The second approach estimates the profiling scores by considering both expert candidates and knowledge areas as queries and measuring their similarity.

An extensive analysis of expert profiling is presented in [STV+11], where the language model used in [BdR07] is considered as one of the features used in a machine learning approach. Other features include a more simple binary model of relevance and the frequency of an expertise topic in expert profiles from the training set. Expertise topics, called tags in this work, are assumed to be known in advance, similar to [BdR07], and are collected in a controlled vocabulary through self assessment. It is worth noting that the authors observe that the quality of expertise topics is more important than the relevance to a particular person. In their experiments, the most important feature with respect to its performance contribution is the frequency of the expertise topic, a feature that is independent of the particular employee. We build on this work by using a quality related measure of expertise topics together with relevance based measures for expert profiling in Chapter 4.

Other interesting findings that motivate our work include the observation that abstract tags such as *customer satisfaction* and *best practices* are rarely used in text and are more challenging to assign in profiles. The authors mention as well that the relation between tags should be considered, because it is unlikely that a person knows about *machine learning* but has no knowledge about *data mining*. These issues can only be

addressed if we make use of a generality measure to identify abstract expertise topics, and if we consider the relations between expertise topics. We propose a solution in this direction by introducing an algorithm for automatically extracting topical hierarchies from domain corpora in Chapter 4.

2.1.3 Topic modelling approaches

One of the main concerns of the expert finding task is to model the expertise of a person based on documents. In this section we discuss a more expressive model than the one used by language modelling approaches, i.e., topic modelling. Statistical topic models [BNJ03] represent documents as mixtures of the themes that run through them, i.e., topics, which are further represented as distributions over the words in the corpus. A topic model adds an additional level of representational power compared to language models, which directly estimate the likelihood of a query using a distribution derived from document words. Topic models are applied and evaluated for document retrieval tasks in [WC06], showing that interpolations between Dirichlet smoothed language models and topic models significantly improve the retrieval performance compared to language models alone.

A first model that adapts topic models to include metadata about document authors is the author-topic model introduced by [RZCG$^+$10]. One of the goals of this model is to learn which author is responsible for a given word in a document. This model is simplified for the expert finding task in [MM07b] by duplicating documents that have more than one author. A more sophisticated topic model is the Author-Persona-Topic model [MM07b], that considers that authors often write about several subject areas and have expertise in a combination of several topics, not just a single topic area. Each author's papers are clustered in different personas that group together papers with similar topical combinations. A hybrid model that combines language modelling and topic-based modelling for expert finding is investigated in [DKLK08].

Topic models usually contain single-word terms, but most technical concepts appear in text as multiword expressions. Multiword topics are tagged in a pre-processing step in the model proposed in [JR10], showing that embedding multiword expressions in the author-topic model achieves better performance than using only words. Topic models go further in the direction of explicitly representing expertise topics than language models, but typically topics are displayed by listing the most frequent words of a topic.

This approach has to be combined with methods that identify topic labels to provide an effective exploratory structure for a document collection.

The work proposed in this thesis identifies expertise topics using term extraction techniques that are specifically designed for multiword terms, avoiding the need for topic labelling algorithms. We introduce the idea of constructing a domain model, which is similar to identifying a single representative topic for a whole domain. Additionally, we take into consideration the relations between terms in the form of topical hierarchies, measuring how well a person knows the specialised vocabulary used in a specific area of a domain, as we will see in Chapter 4.

2.1.4 Ontology-based approaches

Systems that automatically discover expertise information from secondary sources such as articles, email communications, and forums, have to deal with issues related to accessing heterogeneous sources of information and to ensure interoperability. Ontologies, i.e., machine processable models that provide a shared common understanding of information structure, have emerged as a common solution to these problems. Integrating disparate data sources about people and expertise requires laborious manual construction of semantic mappings. More recent approaches for expert finding represent domain knowledge and skills in an ontology, facilitating interoperability between different systems and allowing an approximate matching of skills. But developing a competence ontology is an expensive, time-consuming, and complex process that requires consensus across communities which might have radically different views on a domain. A practical solution is to develop and merge lightweight ontologies or reuse existing formal ontologies developed by consortia and standards organisations.

Paquette [Paq07] introduces a competence ontology that describes generic skills, including cognitive and socio-affective skills, that can be applied to any domain-specific knowledge model. The authors make a distinction between generic skills and domain-specific knowledge, and define several performance indicators such as frequency, scope, autonomy, complexity and context of use. A competence ontology is applied in the computer science domain by integrating an existing computer ontology with different classifications available online [PH04]. Instead of building an ontology for expert finding from scratch, an alternative solution is to reuse, combine, and extend existing vocabularies such as FOAF, SIOC and SKOS [AMBB$^+$07].

The Linked Open Data (LOD) cloud can also be used to collect data for expert profiles such as contact information, affiliation, and academic track [LAT10]. Several LOD metrics for expertise that rely on already assigned expertise topics are proposed in [SJL11], using an off-the-shelf concept extraction service to enrich data sources with topics extracted from textual data. Compared to this work, we adapt existing methods for term extraction specifically for the extraction of expertise topics from technical text. In our experiments, we rely mainly on scientific publications as a source of evidence, but the methods proposed here can be applied to annotate textual data from the LOD cloud as well.

Collaborative competence management is another solution for incomplete and outdated competence data [SB08]. This approach combines bottom-up community tagging processes used for collecting opinions about individual competencies with top-down organisational processes to construct and legitimise a shared competence vocabulary. The authors rely on the ontology maturing process model, assuming that ontologies are formalized in multiple phases, each phase corresponding to a different stage of formality. Several levels of formality of a competence ontology are proposed, such as topic tags, competence types, competencies with levels, and competency relationships. In this thesis, we represent expertise topics through topic tags and we extract the relations between expertise topics in the form of a topical hierarchy proposed in Chapter 4.

We address the need for automatic approaches for ontology construction by extracting expertise topics from text and by grounding expertise topics on the Linked Open Data (LOD) cloud, as described in Chapter 3. At the same time we investigate the relations between expertise topics and organise them in a topical hierarchy in Chapter 4. Most knowledge management approaches for expert finding opt for a user study evaluation, which makes results difficult to compare. Instead, we introduce a benchmark dataset, based on data about workshop committee members, that can be used to compare various approaches for expert search.

2.2 Text mining

In this section we present work from two research areas that address relevant issues for expertise topic extraction: term extraction and keyphrase extraction, then we discuss existing work on extracting relations between terms.

2.2.1 Term extraction

Terms play a central role in modelling a knowledge domain under the assumption that terms unambiguously identify domain-specific concepts. Expertise topics are a subset of the terms in a domain that can be efficiently used to describe a person's expertise. In Section 3.1 we discuss that the distinction between a term and an expertise topic is related to their level of specificity. While terms cover all the concepts in a domain regardless of how general or how specific they are, expertise topics should have an intermediate level of specificity.

Automatic term extraction methods received a lot of attention, as manually built thesauri and other lexical resources are expensive to construct and typically suffer from low domain coverage. Term extraction plays an important role in a wide range of applications including information retrieval [YJZY05], keyphrase extraction [LR10], information extraction [YGTH00], domain ontology construction [KVM00], text classification [BMP02], knowledge mining [MAM06], and machine translation [Gau98].

Two conflicting requirements arise from the need to extract domain-specific terms on one side (i.e., the domain-specificity requirement), and the need for domain-independent approaches on the other (i.e, the domain-independence requirement). For instance, an approach tuned to identify Latin terms could be very useful in biology, but would be of limited use when applied to finance or sport. Current approaches address the domain specificity requirement by making use of contrastive corpora, either general purpose corpora or corpora from other domains, and by comparing term distributions across domains. Although they perform well for domains that use a large number of specialised terms, such as the biomedical domain, the extraction of more general or abstract terms, which play an important role in keyphrase extraction or information retrieval, is still a challenge.

With respect to the domain-independence requirement, it is commonly considered that a term extraction approach is domain-independent if it makes use of domain-independent syntactic patterns and domain-independent ranking and filtering features. But domain-independence is rarely evaluated across multiple domains, as term extraction evaluation is a challenging and time consuming process that requires the involvement of domain experts. One of the few studies that attempt to evaluate term extraction approaches across multiple domains shows that term extraction performance

depends on the domain [ZIBC08]. This study provides valuable insights into the dependence of term extraction systems on the targeted domain, but it has its limitations as it considers only two domains. Another limitation of this study is that one of these domains is a small general knowledge corpus of Wikipedia articles.

Existing approaches for term extraction focus on linguistic methods, statistical methods or a combination of both. The first category of approaches, the linguistic methods, rely on syntactic patterns [JK95] or grammar rules [Bou92] to extract noun phrases or noun-noun compounds. These methods can be successfully applied to small sized corpora because all terms are considered regardless of their frequency but they suffer from a lack of coverage and are difficult to apply to other domains than the one they were developed on. For each new domain, a new set of syntactic patterns has to be manually developed.

The second type of approaches, which analyse the frequency of terms, are more adequate for large corpora. Frequency alone is not sufficient to distinguish true terms as the most frequent words in a language have only a functional role, but good results are reported for the simple technique of discarding the most common and seldom occurring words [Dam90]. Multi-word units have received closer attention in this category of approaches as their structural stability and unity [KU96] distinguishes them from random combination of words. The statistical measures used to characterise these properties of terms include mutual information [Dam93], T-test [CGHH91] and log-likelihood [Dun94]. The third category of term extraction techniques, hybrid methods, take advantage of both directions, making use of syntactic analysis to select candidate terms, and of statistical measures to further refine the final result set [Dai96, FAM00].

Methods that use only corpus statistics are faced with the challenge of distinguishing general language expressions (e.g., *last week*) from terminological expressions. A solution to this problem is to use contrastive corpora [Hui86]. Contrastive corpora approaches for term extraction define domain relevance measures by comparison with general purpose corpora or corpora from other domains. Two assumptions are made: that domain relevant terms are more frequent in their domain and are rarely mentioned in other domains and that they have a uniform occurrence distribution in the documents of the domain. Several measures are proposed based on these assumptions including [PBB02], domain consensus [VMB01] and word impurity [LWY$^+$05]. The advantage of this approach is that it can be successfully applied to multi-word terms as

well as to single-word terms, unlike statistical approaches that deal mostly with longer phrases.

The expertise topic extraction approach proposed in this work is a hybrid approach as well, using syntactic patterns to identify term candidates and a combination of statistical measures to rank the candidates. The main novelty of our approach is that we make use of an automatically build domain model to measure the coherence of a term within a domain, as we will see in Chapter 3.

Benchmarking approaches

In our experiments for expertise topic extraction we employ as baseline two state of the art methods for term extraction, the NC-value approach [FAM00] and TermExtractor[1] [SV07]. The former is a hybrid method that ranks terms using only corpus statistics, while the latter exploits contrastive corpora. Our method for expertise topics extraction presented in Chapter 3 is an adaptation of the NC-value approach. TermExtractor was considered as well because it is a more sophisticated approach for term extraction that makes use of contrastive corpora.

NC-value is a well known term extraction method, initially developed for the biomedical domain. The C-value part of the method, defined in Equation 2.1, is primarily based on raw frequency counts and it deals with nested multi-word terms by penalising frequency counts of shorter embedded terms:

$$CV(a) = log_2|a| \cdot f(a) - \sum_{b \in T_a} f(b) \tag{2.1}$$

where a is the candidate string, $CV(a)$ is the C-value score for the candidate a, $|a|$ is used to denote the length of the term in number of words, $f(a)$ is the frequency of the candidate, T_a is the set of candidate terms that embed a, b is one of these terms, and $f(b)$ is the frequency of term b. This measure can be applied for extracting general terms with some modifications. The frequency score should perform well for general terms as they are used more often than more specific terms, but embedded terms are more general than longer terms that include them. The NC-value part incorporates context information in the C-value measure in a re-ranking step, by first selecting a set of top

[1] TermExtractor demo: `http://lcl.uniroma1.it/sso/index.jsp?returnURL=%2Ftermextractor%2F`

ranked terms using the C-value score. Context words (nouns, verbs and adjectives) are identified based on their occurrence with top candidates, extracted using the C-value measure, with the following weight:

$$w_{NCV}(b) = \frac{t(b)}{n} \tag{2.2}$$

where b is a context word, $t(b)$ is the number of top ranked terms that appear with b, and n is the number of top ranked terms used to identify context words. The NC-value measure is defined as follows:

$$NCV(a) = 0.8 \cdot CV(a) + 0.2 \sum_{b \in C_a} f_a(b) \cdot w_{NCV}(b) \tag{2.3}$$

where a is the candidate string, $NCV(a)$ is the NC-value score for candidate a, C_a is the set of context words, b is a context word from C_a, $f_a(b)$ is the frequency of b as a context word for a, and $w_{NCV}(b)$ is the weight defined in Equation 2.2.

TermExtractor is a popular approach that combines different term extraction techniques including domain relevance, domain consensus and lexical cohesion. Domain Relevance (DR) compares the probability of a term t in a given domain D_i with the maximum probability of the term in other domains used for contrast D_j and is measured as:

$$DR_{D_i}(t) = \frac{P(t/D_i)}{max_j\left(P(t/D_j)\right)}, j \neq i \tag{2.4}$$

Domain Consensus (DC) identifies terms that have an even probability distribution across the corpus that represents a domain of interest, and is estimated through entropy as follows:

$$DC_{D_i}(t) = -\sum_{d \in D_i} P(t/d) \cdot log\left(P(t/d)\right) \tag{2.5}$$

where d is a document in the domain D_i. Finally, the degree of cohesion among the words w_j that compose the term t is computed through a measure called Lexical Cohesion (LC). This measure is used to evaluate the association between each of the words that form a term . Let $|t|$ be the length of t in number of words, and $f(t, D_i)$ the frequency of t in the domain D_i, then Lexical Cohesion is defined as:

$$LC_{D_i}(t) = \frac{|t| \cdot f(t, D_i) \cdot log\left(f(t, D_i)\right)}{\sum_{w_j} f(w_j, D_i)} \tag{2.6}$$

The weight TE used for ranking terms by TermExtractor is a linear combination of the three methods described above:

$$TE(t, D_i) = \alpha \cdot DR + \beta \cdot DC + \gamma \cdot LC \tag{2.7}$$

While general terms typically have a high domain consensus, the domain relevance measure boosts narrow terms that have limited usage outside of the domain. For example the term *system* is not identified as relevant for Computer Science because it is frequently used in general language and in other specific domains as biology. In Chapter 3 we propose a different approach to compute domain specificity that can be applied for general terms by using a domain coherence measure that does not use external corpora.

In our experiments we used our own implementation of TermExtractor that is slightly different from the original system. The main difference is that our implementation does not use any of the structural relevance heuristics or the miscellaneous heuristics. Another difference is related to the criteria that was used to construct the background corpus. The contrastive corpus used in the original implementation of TermExtractor includes several domain specific corpora from domains such as medicine and computer networks. We did not have access to large-scale domain-specific corpora and we relied on general corpora instead. The Open American National Corpus[1] and a corpus of books from Project Gutenberg[2], are used as contrastive corpora in our implementation. The books selected from Project Gutenberg include the bible, the complete works of William Shakespeare, James Joyce's *Ulysses* and Tolstoy's *War and Peace*. Our choice of contrastive corpora is likely to reduce the overall performance of TermExtractor, but considering that our approach does not make use of any external corpora we considered that this does not affect the fairness of the comparison.

We consider only the default setting of TermExtractor assigning equal weights to each measure in Equation 2.7.

[1]Open American National Corpus: `http://www.americannationalcorpus.org/OANC/`
[2]Project Gutenberg: `http://www.gutenberg.org/`

2.2.2 Keyphrase extraction

Similar to term extraction, keyphrase extraction aims at extracting relevant phrases from text, but keyphrases are meant to describe the contents of a single document, while terms are typically extracted from large domain corpora [Ana94, BJ99]. Nevertheless, many of the techniques introduced for term extraction are successfully applied to keyphrase extraction [LR10]. The distinction between keyphrases and expertise topics is that keyphrases are not necessarily domain specific. Any phrase that describes the main contents of a document can be considered as a keyphrase. In the case of expertise topics, these are usually terms that show expertise in a specific domain.

In this section, we discuss general developments in the area of keyphrase extraction, but we also have a closer look at incipient work on features derived from external sources of knowledge in Section 2.2.2.3.

Typically keyphrase extraction systems work in two stages, similar to hybrid approaches for term extraction. In the first stage a general set of candidates is selected, then in the second stage candidates are further filtered to obtain the final result set. A basic approach for candidate selection is to generate all n-grams and then to filter them based on a stopword list, but a more sophisticated candidate selection method that produces better results [Hul04] is to use additional linguistic information in the form of part-of-speech tags or nominal group chunks. The second stage can be addressed by using an unsupervised approach or a supervised approach. The unsupervised approaches combine a set of features in a rank to select the most important keyphrases, while supervised approaches require a training corpus to learn a keyphrase extraction model.

2.2.2.1 Unsupervised approaches

[MT04] proposes an unsupervised graph based approach that considers single tokens as vertices in the graph and uses co-occurrence relations between tokens as edges. Candidates are ranked using PageRank and adjacent keywords are merged into keyphrases in a post-processing step. In [BC00] the frequency of noun phrase heads is exploited, using noun phrases as candidates and ranking them based on term frequency and term length. They select the most frequent noun phrase heads and they consider all the noun phrases that embed them as candidates. This candidate selection strategy has similarities to our proposed approach, but we only consider noun heads that are part

of an automatically constructed domain model. An unsupervised method that ranks filtered n-grams using TF-IDF and the position of the candidate in a document is proposed in [EbR09]. Single word phrases are favoured by measures based on frequency such as TF-IDF as they have a higher probability of occurrence, therefore a compound word boosting factor has to be used to increase performance.

2.2.2.2 Supervised approaches

Kea [FPW+99] and Extractor [Tur00] are two well known approaches that address the task of automatic keyphrase extraction. Kea is a supervised system that considers as candidates all n-grams of a certain length, and applies a Naive Bayes classifier using TF-IDF and position features [FPW+99] to identify keyphrases. A Naive Bayes classifier is also used in the systems proposed by [BF10, NL10] Extractor is a supervised system that selects stems and stemmed n-grams as candidates and tunes its parameters (mainly related to frequency, position, length) with a genetic algorithm. Decision trees can be applied as well for keyphrase extraction [MFW09, LR10, TTHG10]. Additionally, classifiers constructed by rule induction are proposed by [Hul04], using features such as term frequency, collection frequency, relative position and PoS tags.

2.2.2.3 Keyphrase extraction features

By far the most productive and widely applied property for keyphrase extraction is frequency, but a large number of other features are analysed in the literature. A first group of features deals with characteristics of the phrase taken in isolation. Lexical cohesion [PBB02] measures if a sequence of words appears together more often than they would appear by chance. This measure does not typically require knowledge from a background corpus and it is based on the domain corpus alone. The number of tokens in a keyphrase is also a useful clue as longer phrases tend to be mentioned less often and they are preferred as keyphrases as they are more specific and less ambiguous.

Another property of keyphrases is their ability to form new terms using other words by composition [EbR09, PT10], a property that was previously analysed in the area of term recognition [FAM00]. This feature is used to refine the final results by decrementing the frequency of a keyphrase if it is embedded in longer top ranked candidates. We refer to this property as embeddedness and we measure it by counting the candidates that include a keyphrase. In [BF10] the keyphrase suffix is restricted to a predefined

list of suffixes, considering at the same time the PoS sequence of tags of the candidate as a feature.

A second category of characteristics deal with keyphrase properties in relation to the document where it appears. Most existing keyphrase extraction systems opt for the standard TF-IDF metric to assign the relevance of a keyphrase to a document. The document structure plays also an important role and current systems take it into consideration either by using the more simple method of extracting the offset of the first or last position [FPW+99, EbR09, TTHG10] or by analysing the PDF document to extract sections such as title, abstract, introduction, conclusion [LR10, NL10]. These sections are more likely to introduce main concepts in the text than the full body of the document as well as highlighted words and phrases that are referred by acronyms [BF10, TTHG10].

Taking inspiration from automatic term extraction, a third category of features analyses the usage of the keyphrase throughout the corpus (e.g., domain consensus [VMB01]). Domain relevance is a measure that analyses keyphrase usage in comparison with a reference corpus [PBB02, PNMH08]. A general purpose corpus can be replaced by considering the whole web as a corpus and analysing web occurrences [EH09].

Keyphrase coherence

The content of a document is semantically connected, hence a desired property of keyphrases is their coherence with other terms mentioned in that document. Turney discusses a method for keyphrase extraction that computes the semantic relatedness between the top ranked keyphrases by using a statistical measure based on Pointwise Mutual Information (PMI) and a web search engine [Tur03]. He applies a two pass method that selects the best keyphrases in the first step and then re-ranks the answer set using the semantic relatedness between candidates and the top keyphrases. The main limitation of this approach is that sending a large number of queries to a web search engine is time consuming.

A simplification of this method is analysed in [EN10], removing the need of a two step ranking by computing the semantic relatedness of the candidates with author-provided general terms. The authors report only a minor importance for this feature and they suggest this is due to the high generality of the terms. Automatically constructing a domain model as done in Chapter 3, is a step further in this direction as the

words from a domain model are also general terms of a domain. We consider a much longer set of domain words compared to the one analysed by Eichler and Neumann and we analyse the semantic relatedness between keyphrases and the domain model in the whole corpus, not just in a document.

Similarly, a post-ranking step is performed on the final candidate list in [LR10] where the authors re-rank the keyphrases based on their probability of being selected as a keyphrase if another keyphrase is selected. This probability is computed on a repository of research publications that contains metadata about author-assigned keyphrases. Medelyan attempts to measure semantic relatedness using a measure called node degree that is computed by building a graph that has as edges the links between Wikipedia articles [Med09]. In comparison, we do not rely on external knowledge sources to measure coherence, and we do not require any predefined keyphrases.

External knowledge resources

The last group of properties introduced for keyphrase extraction make use of external knowledge resources. Controlled vocabularies are traditionally used for term assignment in libraries, but their main disadvantage is that they suffer from low coverage and are static in nature. This does not easily allow the integration of emergent terminology. Additionally, such a vocabulary is not available in every domain. Wikipedia mitigates some of these issues as it is a large scale resource that is constantly updated and that covers a large number of areas. This motivates the work reported in [Med09], where the author proposes Wikipedia as a source of evidence for keyphraseness. She assumes that, if a keyphrase appears in a document as an anchor, it is probably an important concept.

Another heuristic based on Wikipedia is the one proposed in [EH09], where candidates that correspond to an entry in Wikipedia are considered to be a good keyphrase as well. One of the disadvantages of using Wikipedia is that it contains articles for generic concepts such as “Calculation” or “Result”, which make this feature noisy [BF10]. This problem can be avoided by applying the feature for the most salient parts of a document alone (i.e., title and abstract). Checking if a keyphrase appears in terminological resources might serve a similar purpose [LR10]. The authors make use of a large terminological database that combines MeSH, a controlled vocabulary thesaurus for life science publications with the Gene Ontology, and linguistic resources from WordNet.

In this thesis we propose a method to extract domain specific terms that does not rely on external corpora or knowledge sources. In Chapter 3 we show that domain relevant terms can be extracted by constructing a domain model and by measuring the coherence of a term within a domain. The approach is more appropriate for technical domains that have a low coverage in Wikipedia and it avoids the need to merge a large number of terminological resources as done in [LR10].

2.2.3 Relation extraction

In this work we mainly focus on applying taxonomical relations in expert finding and expert profiling and we leave for future work the application of other types of relations. We briefly discuss methods for measuring the semantic relatedness of terms in section 2.2.3.1. Then, we present in more detail existing methods for taxonomy construction 2.2.3.2.

2.2.3.1 Semantic relatedness

Terms have to be further interpreted to identify the underlying concepts by dealing with problems such as homonymy and synonymy. Different measures of semantic distance between terms can be defined using the distributional hypothesis [Har68]. This hypothesis states that context can be a reliable source of information, as similar words tend to appear in similar contexts. Context can be defined in different ways, either syntactically by looking at predicate-argument structures [BNC00] or without syntactic pre-processing by analysing words that surround a term in a window of predefined size [BBQ04]. An alternative to using distributional similarity is to exploit semantic relations explicitly stated in text through extraction patterns, the so called Hearst patterns [Hea92].

In this work we make use of the distributional hypothesis, defining context as a window of words. Further semantics are added to this type of relations by proposing a global generality measure that is used to construct directed relations between expertise topics, as we will see in Chapter 4.

2.2.3.2 Taxonomy learning from text

Historically, concept hierarchies are considered as adequate solutions for organization, summarisation, and access to information, and are central to classification schemes,

ontologies, controlled vocabularies, thesauri, and indexing languages. Hierarchically organised knowledge areas facilitate the exploration and discovery of domain knowledge, and are a necessary instrument for self-assessment of expertise as well as for expert profiling and expert search. Previous work on expert finding applies topical structure in expert finding, observing that the performance of an expert finding system depends on the level of specificity of expertise topics [BBA+07]. Therefore, a domain taxonomy is an important design consideration in an expert search system, allowing us to determine expertise at different levels of granularity and to consider inexact matches of expertise.

Currently available taxonomies such as WordNet have a low coverage of domain concepts and are unavailable for many domains. Additionally, their manual construction and extension is time consuming and expensive as it requires expert knowledge. A possible solution to these problems, that received increased interest in the context of the Semantic Web, is to automatically derive concept hierarchies from text [BCM05]. Several types of relations between concepts are typically considered in concept hierarchies including *is-a*, *part-of*, and *type-of* relations. Taxonomy learning is concerned with the acquisition of hierarchical relations between pairs of concepts in a first stage, and with their organisation in a tree-like structure in the second stage.

The first stage of taxonomy learning received more attention and is relatively well understood. Hyponym relations can be acquired from text using lexico-syntactic patterns [Hea92], co-occurrence data [SC99], substring inclusion [NMWP99], or through hierarchical clustering [BCM05] and probabilistic topic modelling [WBB10, BMWM12, ZPVP07]. Each of these approaches has advantages and disadvantages. Although using patterns is a widely used approach that generally has high precision, it suffers from low recall and it requires manual definition of patterns. The first disadvantage can be addressed by relying on the large amount of textual data available on the web [KRH08], while the requirement for hand-crafted patterns can be avoided using pattern learning algorithms [SJN04]. Methods based on co-occurrence data use a simple definition of subsumption, with much weaker semantics than taxonomical relations, but they show good performance for navigational purposes. This method is extended to exploit high order occurrences in [DB07]. Substring inclusion is used to construct lexical hierarchies but the approach is too restrictive as it can be used only for a limited number of concepts. Clustering approaches are fully automatic and can be applied to any document

collection but the resulted groups are difficult to label and counter-intuitive. Labelling topics is also a challenge for probabilistic topic modelling methods.

Topical hierarchies proposed in this thesis are constructed using co-occurrence information, similar to the subsumption approach described in [SC99]. In this work, the authors state that a term x subsumes another term y if $P(x|y) = 1$ and $P(y|x) < 1$. Note that this measure only takes into account the occurrence of two terms of interest, without considering their relation with other terms in the graph. In our work we propose a global generality measure that considers the relation of a node with all the other nodes in the graph.

The second stage of taxonomy construction deals with the overall structure of the taxonomy, analysing entire paths not just local relations between pairs of concepts. Assuming that a concept hierarchy is already available, the second stage can be reduced to the more simple task of positioning missing concepts [SJN06, YC09] in an existing hierarchy. Recent approaches [KH10, NVF11] make no such simplifying assumption, analysing the overall structure of a graph where nodes are the terms and where edges represent the relations between them. The approach described in [KH10] gathers hierarchical relations from the web by making use of doubly-anchored patterns, for example *"<root> such as <seed> and *"*, or its reverse, *"* such as <term1> and <term2>"*. In the second stage, the noisy relations harvested using patterns are filtered by discarding nodes that have a low out-degree and in-degree, by eliminating cycles, and by selecting the longest path between node pairs when multiple paths are available.

Both approaches require some prior knowledge, the former algorithm starts with a root concept and basic level concepts or instances, while the latter, requires a small set of upper terms that are used as a stopping condition in an iterative algorithm. In contrast, we propose an approach that is completely automated and that does not require external corpora or knowledge sources. The approach for constructing topical hierarchies proposed in this thesis is an adaptation of the algorithm proposed in [NVF11]. The main difference is that we make use of a distributional approach for extracting relations between terms, instead of definitions. Our algorithm for constructing topical hierarchies is appropriate for technical domains and it can be applied when a limited amount of domain specific documents are available. In Chapter 4 we present a more detailed comparison of topical hierarchies with other methods for taxonomy construction and we discuss why this type of structure is more appropriate for Expertise Mining.

2.3 Summary

In this chapter we reviewed and briefly discussed limitations of current approaches for expert search, including knowledge-driven approaches and data-driven approaches. Then, we presented several key developments in the areas of term extraction, keyphrase extraction and taxonomy construction that motivate the Expertise Mining approach proposed in this thesis. The next two chapters address open issues in expert search by adapting and extending existing text analysis techniques in the context of our application. In Chapter 3 we discuss requirements specific to expertise topic extraction and we investigate term extraction methods that can compromise between the specificity of a term in a domain and its ability to summarise an expertise area. Our analysis of current taxonomy learning approaches motivates Chapter 4, where we show that hierarchical relations extracted from text are a useful source of information for Expertise Mining.

3

Domain adaptive expertise topic extraction through domain modelling

An expertise topic can be defined as a term that summarises the main subfields and concepts of a domain. Expertise topics, such as *Project Management*, *Software Development* and *Java Programming*, can be used to succinctly describe the expertise areas of a person. In this context, techniques developed for term extraction from text are directly applicable, but several challenges arise including the portability of the system across different domains and the identification of terms based on their level of specificity in a domain. Therefore, term extraction tools can not directly be used for expertise topic extraction and have to be adapted for the task at hand.

This chapter is partially based on [BPB13] and is organised as follows. First we present several considerations on the level of specificity of a term in Section 3.1. Then we introduce and define the notion of domain models in Section 3.2. In Section 3.3 we propose a method for constructing a domain model from a domain-specific corpus, then we discuss how domain models can be used to adapt term extraction techniques for expertise topic extraction in Section 3.4. After discussing several methods for extracting expertise topics, we investigate the problem of semantically grounding expertise topics in Section 3.5. Our experimental setup is presented in Section 3.6, followed by an experimental evaluation of the constructed domain models in Section 3.7. We conclude this chapter with a discussion and analysis of our findings in Section 3.8.

3.1 Specificity levels of terms

The vocabulary used in a domain specific corpus contains a combination of domain specific terms, words from general language, and terms from other related domains, as can be seen in Figure 3.1. The examples provided in this figure were selected from a corpus of scientific publications in computational linguistics. Term extraction aims to distinguish between domain specific terms presented on the right side of the picture and terms outside the domain. Expertise topics are usually a subset of domain terms of an intermediate level of specificity. General terms are not appropriate for describing expertise because a person is not regarded as an expert if they have commonsense knowledge or knowledge in areas that are part of general education. Also, when constructing expertise profiles, highly specialised terms can be summarised with broader terms, resulting in more concise expertise profiles. In our example, *algorithm* is a term that is too generic when describing computer science experts, while *generalized augmented transition network grammars* can be summarised by using the broader term *transition network grammars*.

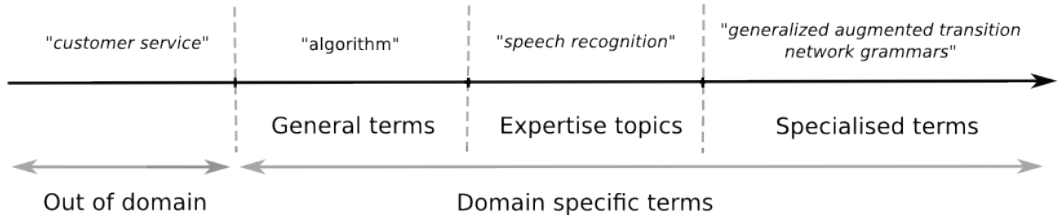

Figure 3.1: Classification of terms based on their level of specificity

Where exactly we draw the line between these levels of specificity depends on the target domain but also on the users of the system. While highly specialised topics could be of interest to expert users, a higher abstraction level is appropriate for the general public. However, identifying the appropriate level of specificity depending on the user is beyond the scope of this work. We focus instead on adapting existing term extraction techniques for the task of identifying intermediate level terms.

Existing approaches for term extraction rely on contrastive corpora to distinguish between domain-specific terms and out-of-domain terms. This makes them better suited for extracting specialised terms, and less appropriate for extracting general terms. Extracting terms of different levels of specificity is a relevant problem in expertise mining,

as well as information retrieval and text summarisation, that received less attention in the literature. To address these challenges, we propose an alternative measure of domain specificity that is appropriate for extracting general terms. Our approach is based on the notion of term coherence with an automatically constructed domain model.

Previous work on term extraction focused on unithood and termhood as two distinctive properties of terms [KU96]. While unithood is based on the assumption that terms are syntagmatic linguistic units, termhood is related to the representativeness of a term in a domain. In this work we analyse existing termhood measures and we adapt them in the Expertise Mining context.

3.2 Domain models

Domain knowledge can be explicitly represented in the form of vocabularies of terms, thesauri, or ontologies. In this work, we hypothesise that a more schematic representation of domain knowledge, in the form of a relatively short list of words, can be used to guide the selection of domain-specific terms. Such a domain model can be manually identified using existing background knowledge, but this might not always be available. For this reason, an automatic approach for constructing a domain model from domain corpora is needed.

A domain model is defined as a representative subset of a domain vocabulary, that has minimal size and high distribution in domain corpora. General terms are preferred when constructing a domain model because they are broad enough to be widely distributed in domain corpora, allowing us to select a relatively small subset of words from a domain lexicon. The size of the domain model has a direct impact on computation time because the domain model is used to measure the coherence of candidate terms as we will see in the following section.

A domain model is represented as a vector of words that contribute to determine the domain of the whole corpus. As such, domain models are similar with category profiles constructed for text categorisation [Mos03], but in this case no predefined list of categories is required or any training data. Domain modelling was shown to be useful for text categorisation and word sense disambiguation in [NFS+11], where the authors construct a semantic representation for a domain, called semantic model vector. In this previous work, a domain model is implemented as a weighted list of synsets. Let

Δ be the domain model, and w_1 to w_n a set of words specific to the domain, then the domain model is formally defined as a vector of words:

$$\Delta = \{w_1, ..., w_n\} \tag{3.1}$$

The number of words n can be empirically set according to a cutoff associated weight. Previous work on using domain information for word sense disambiguation [MPG02] has shown that only about 21% of the words in a text actually carry information about the prevalent domain of the whole text, and that nouns have the most significant contribution (79.4%). This finding motivates our selection of a relatively small number of words for a domain model, as well as the decision to favour nouns. When compared to latent topics as constructed through topic modelling [BNJ03], a domain model provides less structure, by identifying a single vector of representative words for the whole domain corpus.

The domain model proposed in this work is derived from the corpus itself, without the need for external corpora, by identifying terms that are general enough to have broad coverage of a particular domain. An automatic method for identifying the upper level concepts of a domain has applications beyond the task of term extraction. Although not named as such, a domain model is effectively used for text summarisation [TM02], where the authors manually identify a set of 37 nouns including *theory*, *method*, *prototype* and *algorithm*. This manually identified lexicon is used in the construction of lexical patterns for sentence selection. Another area that could benefit from the methods proposed here is the construction of lexical taxonomies. For instance, upper nodes are manually selected from the topmost synsets in WordNet in recent work [NVF11]. This method can be further automated by automatically identifying the most general terms in a domain. This direction is further investigated in Chapter 4, where we explore the construction of topical hierarchies.

3.3 Constructing domain models using domain coherence

Given a domain corpus, representative words of the domain can be selected using a single-word term extraction technique. Several assumptions are made to identify words that are used to construct a domain model from a domain corpus. The first three

assumptions are used for candidate word selection, while the fourth assumption is used to filter the candidate words:

1. **Length:** It is only single-word terms that are considered as longer terms tend to be more specific;
2. **Distribution:** Candidate words should have a high distribution in a domain corpus (the word should appear in at least one quarter of the documents in the corpus);
3. **Saliency:** Candidate words should be content bearing (i.e., nouns, verbs, adjectives);
4. **Semantic Relatedness:** A term is more general if it is semantically related to a large number of domain-specific terms.

The distribution assumption implies that rare terms are more specific, similarly with the frequency-based measure of tag generality used in [BKH+11]. This might not always be the case, for example a simple search with a search engine shows that *artefact* or *silverware* are more rarely used than the term *spoon*, although the first two concepts are more generic. However, in this work we are interested in extracting basic-level categories as theorised in psychology [Haj13]. A basic-level category is the preferred level of naming, that is the taxonomical level at which categories are most cognitively efficient. For example *dime* is always called a *dime* and not *metal object*, *1952 dime* or *10 cents*. A counter example can be found for the length assumption as well, as the longer term *inorganic matter* is more general than the single word *knife*, but in this case we would simply consider as a candidate the single word *matter* which is more generic than the compound term. Length and frequency of occurrence are proposed as general criteria for identifying basic-level categories in [Gre05].

A possible solution for building a domain model is to use a standard termhood measure for single-word terms, and select the top ranked candidate words. But most of the approaches for extracting single-word terms make use of contrastive corpora, favouring specific words that are rarely used outside of the domain. An alternative solution is to use coherence, interpreted as semantic relatedness, which is shown to play an important role in the task of keyphrase extraction [Tur03]. We generalise Turney's measure to quantify the coherence of a term in a domain, instead of the coherence of a

keyphrase in a document. This is done by computing the coherence of a term with the domain model, and not the coherence of pairs of terms. Because the words from the domain model are specifically selected to have high frequency, we can rely on statistics from the domain corpus alone and we do not require any external corpora.

Similar to this previous work, we choose Pointwise Mutual Information (PMI) as a measure of semantic relatedness. This measure was shown to outperform other coherence measures when applied to the task of measuring topic coherence [NLGB10]. First, we extract multi-word terms using a standard term extraction technique such as the one presented in Section 3.4.2, then we use the top ranked terms to rank the candidate words for the domain model using the following scoring function that measures domain coherence:

$$s(\theta) = \sum_{\sigma \in \Omega} PMI(\theta, \sigma) \tag{3.2}$$

Which can be rewritten by replacing the formula for PMI as:

$$s(\theta) = \sum_{\sigma \in \Omega} log\left(\frac{P(\theta, \sigma)}{P(\theta) \cdot P(\sigma)}\right) \tag{3.3}$$

where θ is a word considered as a candidate for the domain model, σ is a multi-word term, Ω is the set of extracted terms, $P(\theta, \sigma)$ is the probability that the word θ appears in the context of the term σ, $P(\theta)$ is the probability of appearance of θ, and $P(\sigma)$ is the probability of appearance of σ. In this work context is defined as a window of words. The set Ω contains the best terms extracted by our basic term extraction method described in Section 3.4.2, but any other term extraction method can be applied in this step. At this stage it is only a relatively small number of automatically extracted terms that are used because span searches are relatively expensive to compute. In our experiments we considered the top 200 ranked terms.

In Algorithm 1 we show how domain coherence can be used to construct a domain model, using a short list of automatically extracted terms. Context is defined as a window of words of predefined size. First, we consider as candidates those words (i.e., nouns, verbs, adjectives) that have high distribution, and which are mentioned in a considerable proportion of the corpus. The filtering step discards words that are mentioned in a small number of documents. Then, each word is scored based on their domain coherence which is measured based on the provided list of terms. The domain

model is finally constructed by selecting the top ranked words based on their domain coherence.

Algorithm 1: The algorithm that constructs a domain model using a list of automatically extracted terms.

input : Window size w
List of words words of size n
List of top ranked terms terms of size t
Size of the domain model d
output: Domain model $domainModel$

```
1 words ← filterWords(words);
2 for i ← 0 to n do
3 |   s ← 0;
4 |   for j ← 0 to t do
5 |   |   s ← s + PMI(words[i], terms[j], w);
6 |   end
7 |   scores[i] ← s;
8 end
9 domainModel ← selectTop(words, scores, d);
```

For example, lets assume that an automatic method for term extraction selects the following best ranked terms from a corpus in Computer Science: *linear programming*, *programming language*, and *data mining*. Also, lets assume that the nouns *algorithm*, *formula*, *software*, *system*, and *smoothness* are the content bearing words extracted from a domain corpus. In the filtering step (line 1 in the algorithm), the word *smoothness* is discarded as less than 10% of the documents in the corpus mention it. This threshold can be adjusted depending on the corpus size and on the size of the resulting domain model. Domain coherence is computed by calculating the PMI for each of the three considered terms (lines 3-7). In this way the normalised values for domain coherence will be 0.95 for *algorithm*, 0.35 for *formula*, 0.68 for *software*, and 0.91 for *system*. If we choose to construct a domain model of 3 words the domain model will be a vector with 3 elements: *algorithm*, *software*, and *system*. While if we choose to construct a domain model with 2 words the domain model will be composed of the words *algorithm* and *system*.

A small sample of words from domain models built for Computer Science, Agriculture and the Biomedical domain, using our domain coherence method, is presented in

Table 3.1. Additionally, a sample list of top ranked words for a domain model in the Computer Science domain is provided in Appendix 2.

Computer Science	Biomed	Food and Agriculture
development	mechanism	control
software	evidence	farm
framework	antibody	supply
information	molecule	food
system	system	forest

Table 3.1: Domain models extracted for different knowledge areas

We observe that a domain model contains several words that are unlikely to be used in other domains, for example *software* from the Computer Science domain model, or *farm* from the Agriculture domain model. These domain models also contain more generic words, for instance *control* or *evidence*, that are likely to appear often in many different domains. Some of the words appear in more than one model, for example *system* that is used in the Computer Science model and in the Biomedicine model, although it refers to slightly different concepts. All of these words are likely to be used in general language. Therefore, measures based on contrastive corpora are unsuitable to extract them, as they favour words that are often used in a domain, but that are rarely used outside the domain.

3.4 Applying domain models for term extraction

A domain model can be used to measure the coherence of a candidate term within a domain. Context words are considered as a valuable source of information for term extraction in previous work [FAM00]. In their work, all the modifiers (i.e., nouns, verbs, adjectives) used in the immediate context of top ranked candidate terms are considered as termhood clues. We propose a more principled selection of context words, assuming that the main concepts of a domain are a suitable approximation of domain semantics. While approaches that make use of contrastive corpora favour the leaves of the hierarchy, our approach based on lexical coherence with general terms is more suitable for identifying intermediate level terms.

The method proposed in the previous section tackles the question about how to build a domain model, but the question about how to use this domain model to ex-

tract terms remains unanswered. Take for example the following structure from the AGROVOC vocabulary[1]: *resources* → *natural resources* → *mineral resources* → *lignite*, where *resources* is an upper level term, *natural resources* and *mineral resources* are intermediate level terms, and *lignite* is a leaf. The underlying assumption of our approach is that, top level terms (e.g., *resource*) can be used to extract intermediate level terms, in our example *natural resources* and *mineral resources*. Intermediate terms can be extracted using a domain model by relying on the notion of compositionality but also by analysing words that appear in the context of a term. Context is needed as well because terms are not necessarily compositional.

In previous work, context is mostly used for conceptual analysis, that is the extraction of definitions and relations between terms. We consider its application for term identification, similar to a previous work in term extraction [FAM00]. A first solution is to follow a well established term extraction technique and extract candidate terms based on noun phrase analysis as described in Section 3.4.1, and to use the domain model in a post-ranking step, as we will see in Section 3.4.3. An alternative, discussed in Section 3.4.4, is to use the domain model for candidate term selection by constructing extraction patterns. Both approaches extend the basic term extraction method described in Section 3.4.2.

3.4.1 Generation and filtering of candidate terms

Candidate terms are discovered relying on a syntactic description of a term, a widely used method previously proposed in [BJ99] for terminology acquisition. The syntactic pattern for a term is defined through a shallow grammar for noun phrases as a sequence of part-of-speech tags. We consider that a noun phrase is a head noun accompanied by a set of modifiers, i.e., nouns, adjectives, that include proper nouns, cardinal numbers ,e.g., *P2P systems*, and gerunds (e.g., *ontology mapping*, *data mining*). Terms that contain the preposition *of* (e.g., *quality of service*) or the conjunction *and* (e.g., *search and rescue*) are also allowed.

We discard candidate terms longer than five words and we ignore those that include one letter words or non-alphanumerical characters. We chose to focus on multi-word terms because single-word terms are more difficult to extract and they are also more ambiguous. Several approaches [PNMH08, TH03] use a reference corpus to model

[1]AGROVOC: `http://aims.fao.org/standards/agrovoc/about`

general language usage, instead we choose to use documents available on the Web by sending queries to an external web search engine. We experimentally choose as an upper threshold for the number of hits 10^9, and as a lower threshold a minimum of 5 hits. Thereby, candidate terms that are too general or too specific and possibly misspelled can be rejected from the final result set based on the number of hits returned by the search engine. In this way we ignore general combinations of words that could appear in any document and that are a major source of noise.

3.4.2 Basic term extraction approach

The basic term extraction approach described in this section is a variation of the C-value method [FAM00], a well known approach for multi-word term extraction, which is mainly based on plain frequency counts. This method was described in more detail in Section 2.2.1. The reason why we selected the C-value method is that although more sophisticated statistical and information-theoretic measures are proposed for term extraction, it has been shown that these measures show little improvement when compared with plain frequency counts [WH06].

The main difference between our basic approach (*Basic*) and C-value is the way we take into consideration embedded terms, that is terms that are contained in longer terms as substrings (e.g., *natural language* is embedded in *natural language processing*). Previously, this information has been used to decrease frequency counts, as shorter terms are counted both when they appear by themselves and when they are embedded in a longer term. We argue that the number of longer terms that embed a term can be used as a termhood measure. For example the term *information retrieval* is used to create several more specific terms such as *information retrieval system*, *information retrieval metric*, and *visual information retrieval*. In our experiments, this measure can only be applied for embedded multi-word terms, as single-word terms are too ambiguous. Hence, the *Basic* scoring method b is defined as:

$$b(\tau) = |\tau| \log f(\tau) + \alpha e_\tau \tag{3.4}$$

where τ is the candidate term, $|\tau|$ is the length of τ, f is its frequency in the corpus, and e_τ is the number of terms that embed the candidate term τ. The parameter α is used to linearly combine the embeddedness weight and is empirically set to 0.72 in our

experiments. This was done by choosing the best setting through an exhaustive search on the interval 0 to 10, with a step of 0.5.

3.4.3 Using domain coherence as a termhood measure

A first solution for applying a domain model to term extraction is to rely on the notion of domain coherence. Here, domain coherence is defined as the semantic relatedness between a candidate term and the domain model described above. The assumption is that a true term should have a high semantic relatedness with representative words of domain semantics. The same measure of semantic relatedness is used as for the domain model, the PMI measure. The domain coherence DC of a candidate term τ is defined as follows:

$$DC(\tau) = \sum_{\theta \in \Delta} PMI(\tau, \theta) \tag{3.5}$$

where θ is a word from the domain model, and Δ is the domain model constructed using Equation 3.2. Note that a similar measure was used in Equation 3.2 to select generic words for the domain model. This method favours more generic terms than contrastive corpora approaches, therefore it is better suited for extracting intermediate level terms. Constructing the domain model with high distribution terms is crucial for ensuring a high recall.

Domain coherence is used in a post-raking step to filter candidate terms extracted using the *Basic* term extraction approach from Section 3.4.2. This method is referred to as *PostRankDC* in the experimental section, standing for Post-Ranking using Domain Coherence (DC).

3.4.4 Building extraction patterns for selecting candidate terms

An alternative approach for using the domain model during the ranking step as presented in the previous section, is to use the domain model to filter and reduce the set of candidate terms considered in Section 3.3. The assumption that the members of a class appear often in the vicinity of other members in conjunctions, enumerations, appositions and compounded terms is the basis of similar work on semantic lexicon construction [RS97]. The goal is to use the domain model to construct context patterns that can be used to extract candidate terms, a widely used approach in information

extraction. In this way, we consider only candidate terms that are mentioned in the immediate vicinity of at least one word from the domain model, instead of analysing all the noun phrases used in a corpus.

The following extraction patterns can be considered for selecting candidate terms. A first category of extraction patterns selects candidate terms that include a domain model word as a substring. If we take for example the word *information* from a domain model, these patterns will select noun phrases such as ***information*** *retrieval*, ***information*** *theory*, or *quantum* ***information***. The second category of extraction patterns selects noun phrases that are introduced by a domain model word followed by one of the prepositions: *for*, *of* and *in*:

*...**approach for** data assimilation...*
*...**development in** Information Retrieval...*

The extraction patterns allow us to select as candidate terms for further analysis the terms *data assimilation* and *Information Retrieval*. In both cases the same syntactic description of a term and filtering methods are used as in Section 3.4.1. Extraction patterns are used to reduce the set of candidate terms extracted by the method presented in Section 3.4.1. This reduced list of candidate terms can be further ranked with the *Basic* term extraction approach described in Section 3.4.2 to obtain the final result list.

3.5 Grounding expertise topics on the LOD cloud

Existing work on using knowledge bases in combination with information retrieval techniques for semantic query expansion shows that background knowledge is a valuable resource for expert search [Dem07, TMS08]. Additional background knowledge, as found on the Linked Open Data (LOD) cloud [1], can provide information at different stages of Expertise Mining. The LOD cloud is a rich and continuously growing resource. On one hand, manually curated concepts are available from a large number of domain-specific ontologies and thesauri. Also, the LOD cloud contains a large number of datasets about scientific publications and patent descriptions that can be used as additional evidence of expertise.

[1] Linked Data: `http://linkeddata.org`

A first step in the direction of exploiting this potential is to provide an entry point in the LOD cloud through DBpedia [1], one of the most widely connected knowledge sources, that is often used as an entry point in the LOD cloud. Two naive but promising approaches for semantic term grounding on DBpedia are described and evaluated in section 3.7.3. Our goal is to associate as many terms as possible with a concept from the LOD cloud through DBpedia URIs, because these URIs can be further used as entry points to other domain specific knowledge sources by using search services such as [TDO07]. Where available, concept descriptions are collected as well and used in our system. Initially we find all candidate URIs using the following DBpedia URI pattern.

http://dbpedia.org/resource/{*DBpedia_concept_label*}

Where *DBpedia_concept_label* is the expertise topic as extracted from our corpus. A large number of candidate URIs are generated starting from a multi-word term as each word from the concept label can start with a letter in lower case or upper case in the DBpedia URI. Take for instance the expertise topic *"Natural Language Processing"*, all possible case variations are generated to obtain the following URI:

http://dbpedia.org/page/Natural_language_processing

To ensure that only DBpedia articles that describe an entity are associated with an expertise topic, we discard category articles and we consider only articles that match the *dbpedia-owl:title* or the final part of the candidate URI with the topic. Multiple morphological variations are extracted and stored from our corpus for each expertise topic. Each of these variations is used to search for a URI, increasing in this way the number of matches.

3.6 Experimental setup

In the previous sections we introduced an approach for domain modelling and two different methods for applying a domain model for expertise topic extraction. Next, an empirical evaluation and comparison is performed, providing first an introduction to the evaluation framework used in our experiments. We begin this section with a discussion of traditional evaluation methods for term extraction and their limitations when applied

[1] DBpedia:`http://dbpedia.org/`

to expertise topic extraction in Section 3.6.1. Then, we describe the datasets used in our experiments in section 3.6.2. To answer research question RQ 1.3, addressing the independence of our approach to a domain, we consider corpora from three different technical domains: Computer Science, Biomedicine, Food and Agriculture which are described in Section 3.6.2.

3.6.1 Evaluation metrics

In this section we discuss the evaluation metrics used in our experiments. First we describe the evaluation metrics used to directly evaluate domain models in Section 3.6.1.1 and then we present the metrics applied for evaluating the results for expertise topic extraction in Section 3.6.1.2.

3.6.1.1 Evaluation metrics for domain modelling

Domain models are evaluated by comparing against a list of manually identified words. Standard information retrieval measures such as precision, recall and F-score are used to measure the performance of our approach. Precision is defined as the proportion of retrieved documents that are relevant, while recall is the proportion of relevant documents that are retrieved. F-score is a measure that gives an estimate of the trade-off between precision and recall by computing their harmonic mean. Both precision and recall are computed on unordered sets of documents and they have to be extended when evaluating ranked lists of results. Many other effectiveness measures are discussed in [BV00], but we limit ourselves to these three well-known measures.

The results of the domain modelling evaluation are reported in terms of the F-score calculated at different cut-off points which is defined as follows:

F-score (F@N) The F-score computed when N results are retrieved, which is used to report F-score after analysing the top 100, 200, or 300 words.

3.6.1.2 Evaluation metrics for expertise topic extraction

The expertise topic extraction task is evaluated based on the quality of the ranked list of terms extracted by the system. The standard measures for information retrieval effectiveness, such as precision, recall and F-score, can therefore be applied. The main measure used for evaluating term extraction is Precision at top K%, which is defined

as follows:

Precision (P@K%) The Precision computed when K% of the results are retrieved, which is used to report precision after analysing the top 20%, 40%, 60%, 80%, and 100% results of the ranked list.

This measure is similar with precision at top N (P@N), but we report performance for larger portions of the ranked list, aggregating precision for a larger number of results.

Previous term extraction studies rely on direct human judgement to evaluate the top terms ranked by a system. Evaluating the portability of a system across different domains is a challenge in this setting, because it is difficult to involve domain experts even for one domain. On the other hand, datasets annotated for applications where term extraction plays an important role, such as keyphrase extraction or index term assignment, are more and more widely available. These datasets can be used for an application-based evaluation of term extraction systems.

Another limitation of previous evaluation of term extraction is that performance is often estimated by manually analysing a small subset of extracted terms. Recall cannot be analysed in such a setup, with most studies reporting the results in terms of precision alone. Instead, we propose an application-based evaluation of term extraction, making use of datasets annotated for keyphrase extraction and information retrieval. These datasets allow a more fine-grained evaluation, at the corpus level and at the document level, of both precision and recall. In this way, instead of manually analysing a limited number of terms for the whole corpus, we can automatically evaluate a larger list of candidate terms against gold standard terms.

3.6.2 Datasets

In this section we describe the datasets used in our experiments. First we describe the dataset used in the intrinsic evaluation of domain models in Section 3.6.2.1 and then we discuss the datasets used for evaluating expertise topic extraction in Section 3.6.2.2.

3.6.2.1 Dataset for domain modelling

The proposed method for building a domain model is analysed through an intrinsic evaluation in the Computer Science domain. A dataset was gathered for evaluating

the extraction of general terms from domain-specific corpora. Our choice of domains was conditioned by the availability of domain experts. The annotation of words for a domain model was done in two steps to reduce the complexity of the task. In the first step one expert analysed an extended list of terms with the purpose of selecting words for a domain model. In the second step several experts analysed a subset of these terms to measure agreement.

In the first step the expert is asked to analyse the nouns used in a taxonomy of Computer Science subjects, the ACM Computing Classification System[1]. The expert is provided with the list of nouns and their frequency in the taxonomy and is required to identify nouns that refer to generic concepts, that are not specific to a particular subfield of the domain. A set of 80 nouns are selected in this manner including *system*, *information*, and *software*. The complete list of words from this manually constructed domain model of Computer Science is presented in Table 3.2.

algorithm	distributed	interaction	optimisation	software
analysis	engineering	interface	pattern	solution
approach	environment	interpretation	prediction	specification
approximation	estimation	language	probability	standard
architecture	evaluation	machine	problem	standardisation
automation	execution	management	procedure	statistic
classification	feature	measurement	process	strategy
computation	framework	mechanism	processing	structure
computer	generation	method	processor	study
control	graphic	methodology	program	synthesis
data	hardware	metric	programming	system
database	hierarchy	modelling	protocol	technique
definition	implementation	model	reasoning	technology
design	information	module	representation	theory
development	integrated	multiparadigm	science	tool
device	integration	network	service	workbench

Table 3.2: A manually constructed domain model for Computer Science

This gold standard is built by one annotator because the task requires analysing and filtering several hundred words. Inter-annotator agreement is computed by analysing a subset of the selected words through a survey that involved 27 participants. More details about the instructions provided to the participants and their anonymised answers

[1]ACM Computing Classification System: `http://www.acm.org/about/class/1998/`

can be found in Appendix 1. A quarter of the words identified by the domain expert are randomly combined with the same number of randomly selected words from the rejected list, which are used as negative examples. The participants are given an alphabetically sorted list of words that contain the positive and negative examples presented in Table 3.3. We used the Fleiss kappa statistic to calculate the interrater agreement. Kappa is 0.34, lying in the fair agreement range.

A qualitative analysis of the answers shows that 80% of the words from our gold standard domain model are selected by at least half of the participants. The positive examples that received less than half of the votes are *approximation*, *execution*, *generation*, and *probability*. Even so, these words received a minimum of 9 votes each. A possible reason why these words are rejected by the majority of the participants could be that they are arguably too general for the given domain. Another conclusion is that although the gold standard list is overall considered to be correct by a large number of participants, the list is not exhaustive as two words that were initially rejected by the first annotator (i.e., *concept* and *user*) are considered to be correct by a majority of experts.

Positives		**Negatives**	
algorithm	machine	agent	key
analysis	optimisation	circuit-switching	mathematics
approximation	probability	concept	moment
data	processing	connection	multimedia
definition	protocol	debuggers	sorting
execution	service	depth	speech
feature	solution	directory	supplier
generation	standard	framebuffer	text
integration	technology	frames	user
language	theory	grayscale	word

Table 3.3: Positive and negative examples used in the domain modelling survey

3.6.2.2 Datasets for expertise topic extraction

In this section we describe the datasets used to evaluate our approach for expertise topic extraction. We make use of a dataset provided by the organisers of the Semantic Evaluation (SemEval) workshop for the task of Automatic Keyphrase Extraction from Scientific Articles. We also rely on three domain-specific datasets to evaluate how well our approach performs across domains.

SemEval 2010 collection

The SemEval 2010 competition included a task targeting the Automatic Keyphrase Extraction from Scientific Articles [KMKB10]. Given a set of scientific articles participants are required to assign keyphrases extracted from text to each document. We participated in this task with an unsupervised approach for keyphrase extraction that does not only consider a general description of a term to select candidate keyphrases but also takes into consideration context information [BB10a].

The SemEval task organizers provided two sets of scientific articles, a set of 144 documents for training and a set of 100 documents for testing. The collection consists of ACM publications from four subdomains (i.e., C.2.4 Distributed Systems, H.3.3 Information Search and Retrieval, I.2.6 Learning and J.4 Social and Behavioral Sciences). The average length of the articles is between 6 and 8 pages including tables and pictures. Three sets of answers are provided: author-assigned keyphrases, reader-assigned keyphrases and combined keyphrases (combination of the first two sets). The participants are asked to assign a number of exactly 15 keyphrases per document. All reader-assigned keyphrases are extracted from the papers, whereas some of the author-assigned keyphrases do not occur explicitly in the text. Several alternations of genitive keyphrases are accepted, for example *policy of school*, *school policy*, and *school's policy* are all considered to be correct. In case that the semantics changes due to the alternation, the alternation is not included in the answer set.

The SemEval dataset has several drawbacks that make it of limited use for our evaluation of domain modelling. The dataset contains a relatively small number of documents from a small number of domains. While it allows us to evaluate the performance of our system for keyphrase extraction, it does not allow us to analyse if we are able to model multiple domains. For this reason we also consider several other domain-specific datasets described in the next section.

Domain-specific collections

Domain independence is an important requirement for expertise topic extraction, as no assumption can be made about the application area of the expertise mining system. The system should achieve acceptable results on academic content from various scientific areas as well as enterprise documents. For this purpose we consider three corpora

from a wide range of domains including Computer Science, Biomedicine and Food and Agriculture. *Krapivin* [KAM09] is a corpus of Computer Science scientific publications that provides author and reviewer assigned keyphrases. *GENIA* [OTK+01] is a corpus of biomed abstracts that are exhaustively annotated with terms, with about 35% of all noun phrases being annotated as correct terms. The Food and Agriculture corpus (*FAO*) contains several reports collected from the site of the Food and Agriculture Organization of the United Nations[1]. This dataset provides index terms assigned to each document by professional indexers [MW08].

It is not only the document size that varies considerably across these three corpora, but also the number of annotations assigned to each document as can be seen in Table 3.4. This table presents the number of documents (Documents), the number of tokens (Tokens), and the average number of annotations per document (Avg. Ann.). *FAO* is the smallest corpus in terms of the amount of documents but it has the largest size in tokens number. This implies there is a much larger number of candidate terms and a smaller number of true positives, making it the most challenging use case of the three. At the opposite end of the scale the GENIA corpus is composed of shorter documents, with a smaller number of candidate terms, that are annotated with a relatively high number of correct terms. This indicates that term extraction systems should yield better results in the biomedical use case.

Corpus	Documents	Tokens	Avg.Ann.
Krapivin	2304	$22 \cdot 10^6$	5
GENIA	1999	$0.5 \cdot 10^6$	37
FAO	780	$28 \cdot 10^6$	8

Table 3.4: Statistics about the corpora employed in our experiments

3.7 Experimental evaluation

We have detailed our approach for constructing a domain model in Section 3.3, then we have discussed how a domain model can be applied to improve term extraction in Section 3.4, followed by a discussion of our experimental setup in Section 3.6. In this section, we present an experimental evaluation of our method, answering the research

[1]Food and Agriculture Organization of the United States: `http://www.fao.org`

questions RQ1.1 and RQ1.2 in Sections 3.7.1 and 3.7.2, respectively. These research questions are concerned with the intrinsic and extrinsic evaluation of a domain model and its application for improving term extraction results.

3.7.1 Evaluating the domain model

In this section, we address the research question RQ 1.1, related to the construction of a domain model. The task of constructing a domain model is cast as the task of ranking candidate words selected from a domain specific corpus using various scoring functions. A set of experiments that deal with the intrinsic evaluation of a domain model is presented, using a manually constructed gold standard dataset. As a benchmark we make use of two methods proposed for term extraction and a method used for constructing concept hierarchies. Additionally, several other benchmarks based on probabilistic topic modelling are discussed and compared with our approach. Results are evaluated in the Computer Science domain using the gold standard dataset constructed in Section 3.6.2.1. The *Krapivin* [KAM09] corpus described in Section 3.6.2.2 was used to extract a domain model for Computer Science.

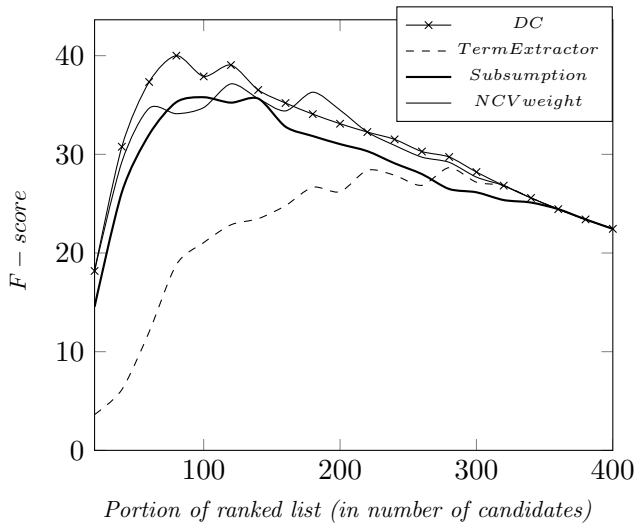

Figure 3.2: Methods for extracting a domain model

The first two considered benchmarks are the contrastive approach used in TermExtractor and the more simple frequency-based method used by NC-value to select context words. For more details about these two approaches check Section 2.2.1. Furthermore,

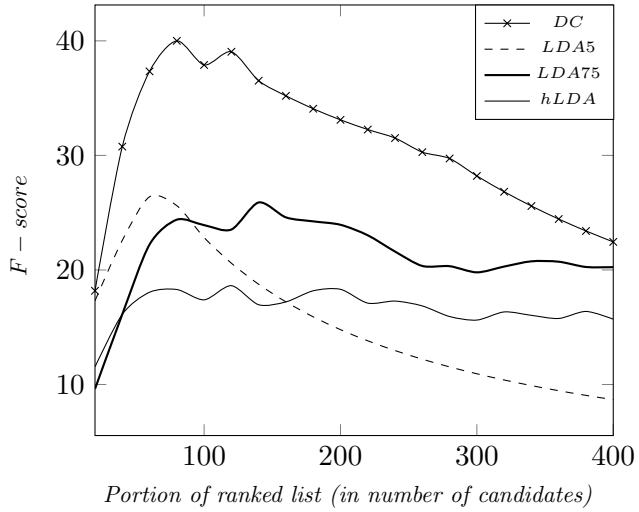

Figure 3.3: Comparison of domain modelling with topic modelling approaches

we consider a statistical method used for the construction of subsumption hierarchies in document browsing [SC99]. In our implementation, context is defined as a window of 5 words. All nouns that are mentioned in at least one quarter of the documents are considered as candidate words for the domain model, excluding the ones that appear in a stopword list. In our implementation we used a widely used stopword list that contains 429 words [1].

The results of this experiment are shown in Figure 3.2, where performance is measured in terms of F-score at top N results (F@N), which is described in Section 3.6.1.1. Several conclusions can be drawn from this experiment. First, the methods that analyse the context of top ranked terms (i.e., our domain coherence measure, DC, the weight used for context words in the NC-value, w_{NCV}, and the *Subsumption* approach) perform better than the contrastive measure used in TermExtractor, with statistically significant gains. Also, our domain coherence method outperforms the more simple frequency-based weight used in NC-value and the subsumption score, although the results are not statistically significant in this case. As expected, the words ranked high by TermExtractor are too specific for a generic domain model.

[1] Stopword list: `http://www.lextek.com/manuals/onix/stopwords1.html`

Comparison with latent semantics

Because a domain model is a semantic group of words, it has many similarities with topics extracted using topic modelling. Therefore a natural question is whether techniques developed to discover latent semantics in a domain corpus are suitable to identify a domain model. When compared to topic modelling, a domain model provides less structure, as it identifies only a single topic for the whole domain corpus. A popular probabilistic approach to topic modelling, Latent Dirichlet Allocation (LDA) [BNJ03], is used in our experiments.

Topics are modelled as probability distributions over words, therefore topics can be seen as a list of words ranked based on their probability of occurrence. To make the approaches comparable, for each topic we select the top 20 words with the highest probability of occurrence. In topic modelling, the generality of topics depends on the number of extracted topics, the larger the number of topics, the more specific the thematic information represented by them. As we are interested in identifying general words, we experiment with different settings for the number of extracted topics (i.e., 5, 10, 25, 50, 75, 100, 150, and 200 topics).

Figure 3.3 compares the results obtained by our domain modelling approach DC and the best performing LDA settings. These were found when extracting 5 ($LDA5$) and 75 topics ($LDA75$). At most 20 words are analysed for each topic and a total of 400 words were considered in our evaluation. The $LDA5$ benchmark is an exception because when limiting the number of topics to 5 with maximum 20 words per topic, only 100 words are evaluated. The $LDA5$ benchmark does not match any of the gold standard words after this point because no data was available and the F-score decreases steadily for larger cut-off points.

Another solution for identifying topics based on their generality is to make use of a method that learns topic hierarchies, such as the hierarchical LDA approach ($hLDA$) described in [BGJT04]. Default settings are used for $hLDA$, extracting a three level hierarchy of 20 supertopics and 30 subtopics, which are all considered for our evaluation. In this case, only the best ten words are considered for each topic. In our experiments, the implementation available in Mallet [McC02] is used for both approaches.

A manual analysis of the results shows that the probability of occurrence of a word in a latent topic is not related with the generality of the word. General words

are combined with more specific words to form a topic. On the other hand domain modelling is more successful in bringing general words at the top of the list as can be seen in Figure 3.3. The LDA approach outperforms the *hLDA* approach, but both approaches under-perform when compared to domain modelling.

The experiments presented in this section answer the research question RQ 1.1, related to effective ways to automatically construct a domain model. Our approach based on domain coherence is more effective for constructing a domain model than existing term extraction approaches, subsumption and topic modelling approaches.

3.7.2 Term extraction evaluation

Domain models are constructed to improve the extraction of intermediate level terms, therefore in a second set of experiments we describe an extrinsic evaluation on tasks such as keyphrase extraction and index term extraction. We present our participation in the SemEval competition, in a task that deals with the automatic extraction of keyphrases from scientific articles [BB10a], in Section 3.7.2.1.

Domain independence is evaluated across multiple domains in a traditional evaluation for term extraction in Section 3.7.2.2, as well as indirectly in an application-based evaluation in Section 3.7.2.3. All three corpora described in Section 3.6.2.2 are annotated for tasks where termhood plays an important role (i.e., keyphrase extraction, information extraction, and information retrieval) but the annotations are assigned at document level. Therefore, we envision two sets of experiments: a standard term extraction evaluation where the top ranked terms are evaluated against the list of unique annotations from our datasets in Section 3.7.2.2 and a second set of experiments where each term extraction approach is used to assign candidate terms to documents in combination with a document relevance measure in Section 3.7.2.3.

3.7.2.1 SemEval 2010 participation

This section is partially based on [BB10a]. A shared task offers the opportunity to compare results with a large number of state-of-the-art systems. We participated in the SemEval workshop with a keyphrase extraction system (referred to as *DERIUNLP*) that implements a variation of the context patterns method described in section 3.4.4. The main difference is that the ranking function does not take into consideration term

embeddedness. The extraction patterns for selecting candidate keyphrases are constructed using the manually extracted domain model that served as a gold standard in Section 3.6.2.1. The organisers provided participants with three different baselines, an unsupervised method and two supervised methods. The supervised methods use the Naive Bayes (*NB*) and the maximum entropy (*ME*) algorithms, as implemented in WEKA[1]. These baseline systems select candidate keyphrases from text by considering all unigrams, bigrams and trigrams. The unsupervised baseline ranks candidate keyphrases using TF-IDF as a scoring function, while the supervised baselines use TF-IDF as a feature.

The results are presented in Tables 3.5 and 3.6, where column labels specify the cut-off points in the output list and the evaluation metric. The first character of the column name indicates if the evaluation is done for top 5, 10, or 15 candidate keyphrases, while the second character (i.e., *P*, *R*, *F*) identifies the evaluation metric (i.e., micro-averaged Precision, Recall and F-score, respectively). Our *DERIUNLP* run considerably outperformes all baselines on all the metrics, as can be seen in Table 3.5. We include for comparison the performance of our system for a run that considers substring inclusion. This run, called *SaffronDM* in Table 3.5, obtains even higher performance on this dataset.

Method	**5P**	**5R**	**5F**	**10P**	**10R**	**10F**	**15P**	**15R**	**15F**
TF-IDF	22	7.5	11.19	17.7	12.07	14.35	14.93	15.28	15.1
NB	21.4	7.3	10.89	17.3	11.8	14.03	14.53	14.87	14.7
ME	21.4	7.3	10.89	17.3	11.8	14.03	14.53	14.87	14.7
DERIUNLP	27.4	9.35	13.94	23	15.69	18.65	22	22.51	22.25
SaffronDM	**35.2**	**12.01**	**17.9**	**27.0**	**18.42**	**21.9**	**23**	**23.5**	**23.26**
SaffronFreq	15.83	5.13	7.75	13.40	8.68	10.54	13.33	12.96	13.14

Table 3.5: Baseline and DERIUNLP performance over combined keywords

Our system does not use plain frequency counts as a feature, but instead relies on the number of cooccurrences with domain model words. To show the contribution of the domain model, Table 3.5 reports the results of a version of our system (called *SaffronFreq*) that uses the overall frequency of a candidate keyphrase instead. The results of the *SaffronFreq* run drop considerably, suggesting that co-occurrence with words

[1]WEKA:`http://www.cs.waikato.ac.nz/ml/weka/`

System	5P	5R	5F	10P	10R	10F	15P	15R	15F
Best	39.0	13.3	19.8	32.0	21.8	26.0	27.2	27.8	27.5
Average	29.6	10.1	15	26.1	17.8	21.2	21.9	22.4	22.2
Worst	9.4	3.2	4.8	5.9	4.0	4.8	5.3	5.4	5.3
DERIUNLP	27.4	9.4	13.9	23.0	15.7	18.7	22.0	22.5	22.3

Table 3.6: Participant performance over combined keywords

from the domain model is more informative for keyphrase extraction than frequency alone.

In Table 3.6 we take a critical look at our results in comparison with the results of other participants. Even though our system considers in the first stage a significantly limited set of candidate keyphrases selected using context patterns, the results are only slightly below the average results of other participants. Our system was ranked on the 8th position out of 19 systems when considering performance in terms of F-score at top 15 in the result list, on the 10th position at top 10 and on the 13th position at top 5.

While the results are comparable with other state of the art systems, it is worth noting that our system was the third best unsupervised system in this competition. The best performing unsupervised systems in this competition were KP-Miner [EBR10] and KX [PT10]. The results achieved by our approach (i.e., DERIUNLP) are compared against the results of these two systems in Table 3.7. KP-Miner outperforms our system by a wider margin at top 5 (32% improvement) than at top 15 (13% improvement), which indicates that our system is more competitive when extracting a larger number of keyphrases.

System	Rank	5P	5R	5F	10P	10R	10F	15P	15R	15F
KP-Miner	3	36.0	12.3	18.3	28.6	19.5	23.2	24.9	25.5	25.2
KX_FBK	7	34.2	11.7	17.4	27.0	18.4	21.9	23.6	24.2	23.9
DERIUNLP	8	27.4	9.4	13.9	23.0	15.7	18.7	22.0	22.5	22.3

Table 3.7: Performance of unsupervised systems over combined keywords

Also of interest is that the best performing system, [LR10], has access to a large terminological database that combines many existing terminological resources. It can be assumed that the performance of this system on a different domain depends on the coverage of their terminological database. Furthermore, several participants exploited

the structure of a scientific article by analysing metadata available only in the PDF format. While this approach provides valuable features for the extraction of keyphrases from scientific articles, it cannot be easily applied to other types of documents. Each type of document and of document structure would require a different set of heuristics. A requirement of our application scenario, Expertise Mining, is that the proposed solution should be domain independent. Therefore, we conclude that features based on PDF metadata are not appropriate in our context.

3.7.2.2 Standard term extraction evaluation

In this section, we answer research question RQ 1.3 and we partially answer research question RQ 1.4. We compare the *Basic* term extraction method presented in Section 3.4.2 and the method based on domain coherence described in Section 3.4.3 against the NC-value and TermExtractor methods which are used as benchmarks. The method for selecting candidate terms described in Section 3.4.1 is used for all the evaluated approaches. Also, to assure the results are comparable, the same number of context words is used in NC-value as the size of the domain model used in our approach.

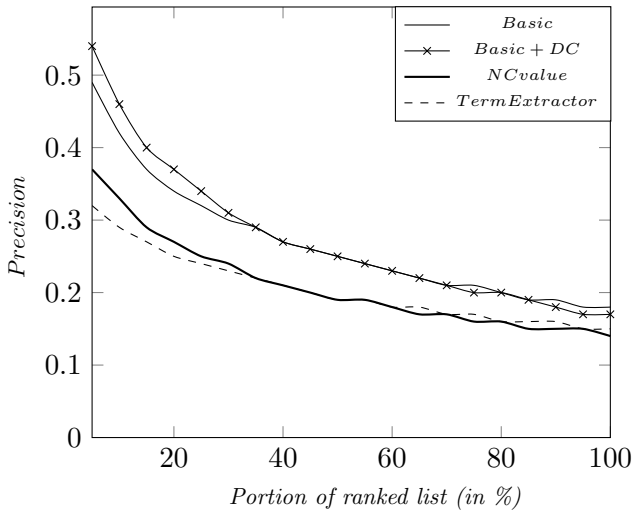

Figure 3.4: Term extraction precision for top 10k terms in Computer Science

While keyphrases and index terms suit our purposes well as they are terms of an intermediate level of specificity meant to summarise or classify documents, many of the terms annotated in GENIA are specific to a given document and not necessarily representative of the Biomedical domain. A solution to this problem is to filter gold

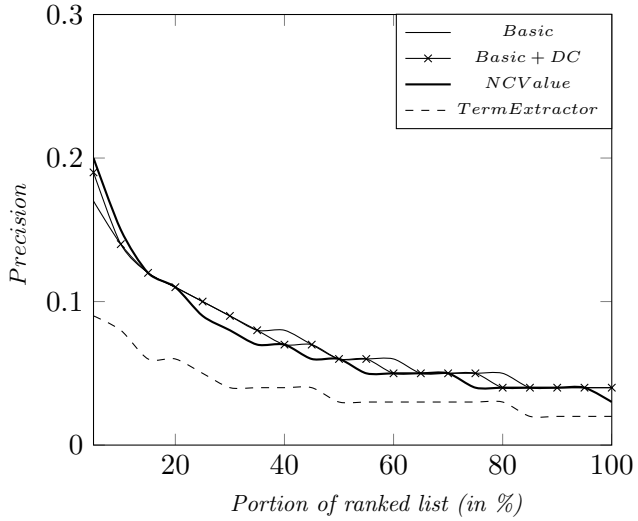

Figure 3.5: Term extraction precision for top 10k index terms in Food and Agriculture

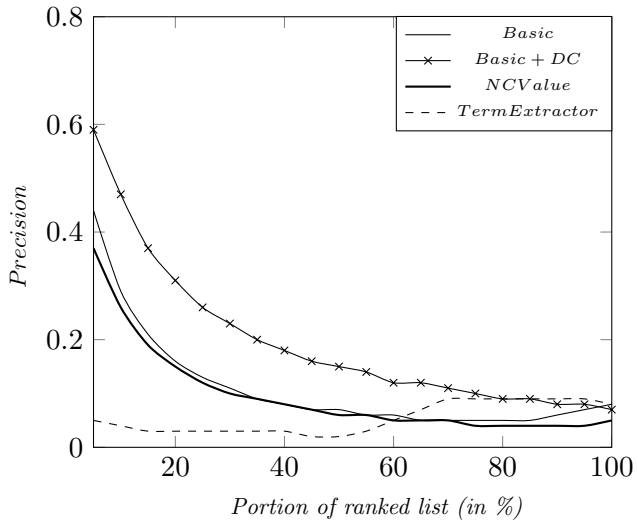

Figure 3.6: Term extraction precision for top 10k terms from the Biomedical domain

standard terms based on their distribution in the corpus, assuming that general terms are mentioned in several documents throughout the corpus. The frequency of a term has long been associated with its level of specificity [JS07], therefore we can assume that this filtering step will reduce the annotated terms to more general ones. To this end, all the gold standard terms that appear in less than 1% of the documents from the GENIA corpus are discarded in our term extraction experiment. This solution is

not without flaws, as it might disregard a large number of general terms that are rarely used in our corpus. In this way, a reduced list of about 5% of the unique gold standard terms provided in the GENIA dataset are considered. Also, it is worth mentioning that TermExtractor achieves the best results on the complete set of annotated terms, which seems to indicate that this system performs better for specialised terms than for general ones.

In our experiments, the top ten thousand ranked terms are evaluated for each of the three datasets. We analyse portions of the ranked lists computed using the *Basic* approach, the *Basic* approach linearly combined with the domain coherence measure (*Basic+DC*), and the two benchmarks, *NC-value* and *TermExtractor*. The measure used is P@K%, where portion is defined as a percentage of top ranked results from the entire ranked list. The precision for a portion of the list is averaged using the overall number of candidate terms considered.

First, we observe that all methods achieve better results on the GENIA (Figure 3.6) and the Krapivin corpus (Figure 3.4) than on the Food and Agriculture corpus. The best methods achieve a maximum precision close to 60% at the top of the ranked list on the GENIA dataset and of more than 50% on the Krapivin dataset. The Food and Agriculture use case is more challenging, as the best method achieves a precision of less than 20%, as can be seen in Figure 3.5. Also, the contrastive corpora measure employed in TermExtractor yields considerably worse results on all three domains. A qualitative analysis of the results showed that TermExtractor is more sensitive to misspelled words, general modifiers (e.g., *large* knowledge base), and out-of-domain terms that are rarely mentioned in contrastive corpora. This indicates that the results of TermExtractor could be improved through a more careful selection of external corpora.

The *Basic* method, which rewards embedded terms, outperforms the NC-value method on the Computer Science domain, and in the biomedical domain, but it performs slightly worse on the Agriculture domain. The combination of the *Basic* method with the domain coherence measure (referred to as *Basic + DC* in the legend) yields the most stable behaviour, outperforming all other measures across the three domains, considerably so in the biomedical domain (Figure 3.6) and at the top of the ranked list in Computer Science (Figure 3.4). In Computer Science, domain coherence significantly outperforms the best performing state-of-the-art method, NC-value (Figure 3.4). For example at 20% of the ranked list the improvement is 37%. In Biomedicine,

the improvement is statistically significant, with a gain of 106% at top 20% of the list (Figure 3.6). These results lead to the conclusion that using a domain model is more appropriate for extracting intermediate level terms than using statistical approaches based on contrastive corpora.

3.7.2.3 Application-based evaluation

An important reason for developing term extraction techniques is for their contribution in specific applications, for example keyphrase extraction or index term assignment. Thus, a reasonable evaluation of term extraction techniques is to evaluate their performance in a given application. This section complements the standard evaluation approach presented in the previous section with an application-based evaluation, addressing the research question RQ 1.4, about the portability of the expertise topic extraction approach across domains.

In this set of experiments we considered again the three domain-specific datasets described in Section 3.6.2.2, but we evaluated the extracted terms at the document level. Instead of evaluating a single list of ranked terms for the whole corpus, we evaluated a list of ranked terms for each of the documents separately and then we aggregated the results. This is in line with the usual evaluation of tasks such as keyphrase extraction or index term assignment, for which the datasets were initially constructed. For this purpose, each term extraction approach has to be adapted for the task at hand. To keep the results comparable, we adapted all the considered term extraction approaches in the same way.

Typically, a termhood measure is combined with various measures of document relevance to perform keyphrase extraction or index term assignment, because candidate terms have to be assigned at the document level. In our experiments we considered the standard information retrieval measure *TF-IDF*, to measure the relevance of a term to a document. The same method for selecting terms proposed in section 3.4.1 is considered, and then candidate terms are ranked using different methods for term extraction, generically called *termhood* in the following equation. To assign terms to documents we combined the *termhood* score which is measured at the corpus level with the *TF-IDF* relevance measure as follows:

$$t(\tau, \delta) = termhood(\tau) \cdot tfidf(\tau, \delta) \tag{3.6}$$

where t is the scoring function used to assign candidate terms to documents, τ is the candidate term, δ is the given document, *termhood* is the scoring function used for term extraction (e.g., NC-value, TermExtractor) and *tfidf* measures the relevance of a term τ for a document δ. In this way, the top ranked candidate terms using the scoring function t are assigned to the document δ.

In this set of experiments the best results are obtained by using domain coherence as a post-processing step, method which is called *PostRankDC* and which was described in Section 3.4.3. The top 30 candidate terms, selected using our *Basic* approach described in Equation 3.4, are re-ranked based on their domain coherence with the domain model using the scoring function t.

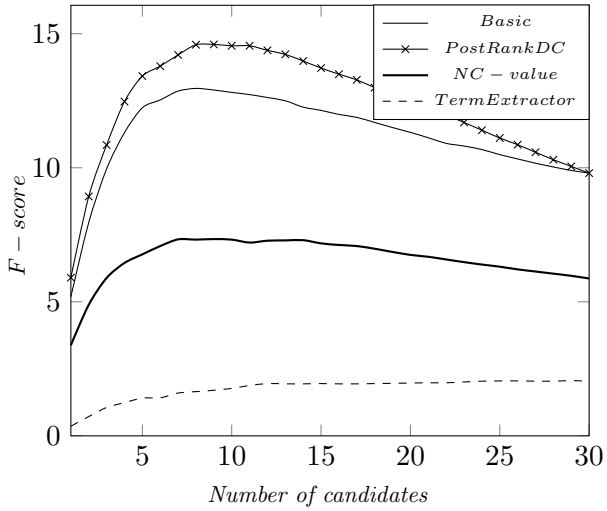

Figure 3.7: Keyphrase extraction evaluation on the Krapivin corpus

The application-based evaluation proposed in this work allows us to evaluate both precision and recall, therefore F-score can be employed as an evaluation metric. The results for keyphrase extraction in Computer Science are presented in Figure 3.7, while the results for index term extraction in the Agriculture domain are shown in Figure 3.8, and the results for term extraction at the document level in the Biomedical domain appear in Figure 3.9. On the x-axis we display different cut-off points of the ranked output list, and on the y-axis we plot the F-score in percentages.

All three methods yield a higher performance on the GENIA corpus, because a considerably higher proportion of all the noun phrases in the text are annotated as correct terms compared to the two other datasets. Although the GENIA documents

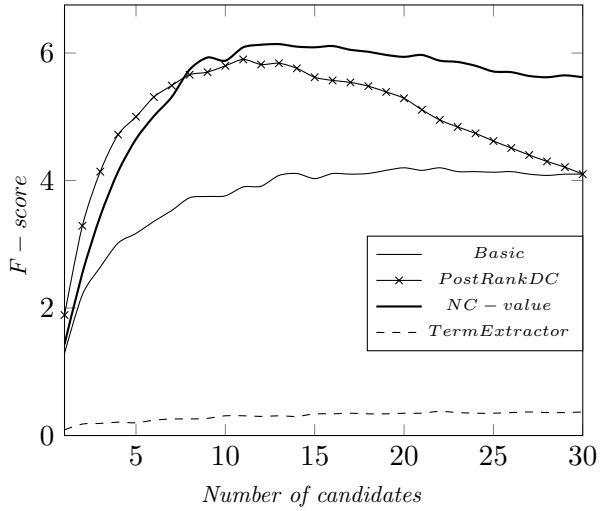

Figure 3.8: Index term evaluation on the FAO corpus

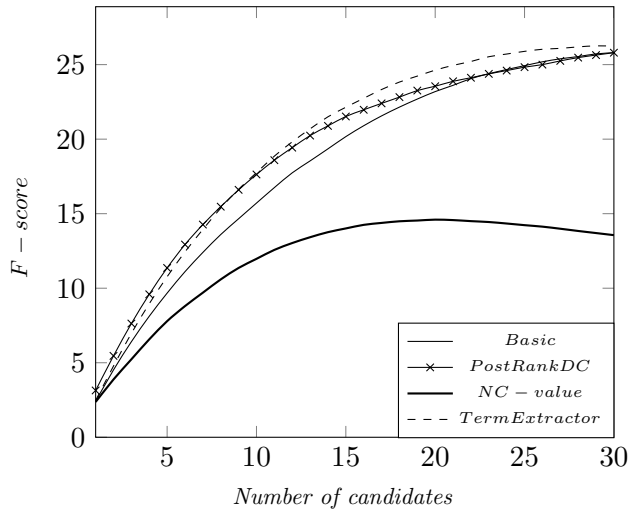

Figure 3.9: Term extraction at the document level on the GENIA corpus

are in average much shorter than the other documents, there are more than four times correct terms than in documents from the other domains. A random baseline would also achieve higher results on this dataset. The results on the Agriculture corpus are again the lowest, because a larger number of candidate terms has to be analysed, compared to the other two domains. The contrastive measure employed by TermExtractor is not suitable for extracting generic terms, such as keyphrases or index terms, as can be seen in Figure 3.7 and Figure 3.8, but it outperforms the other methods when extracting

more specific biomedical terms.

The *Basic* method outperforms the NC-value approach on the Krapivin corpus and on the GENIA corpus, but not on the FAO corpus. This leads to the conclusion that embedded terms have a different behaviour across domains. We can observe that the domain coherence approach (*PostRankDC*) considerably improves over our *Basic* approach on all three domains. After the post-ranking step, the improvement is statistically significant compared to the best performing state-of-the-art method on the Computer Science dataset, NC-value. In this domain the improvement is 99% better than NC-value when reporting the top 10 ranked terms per document. NC-value outperforms TermExtractor in Computer Science and Agriculture, but TermExtractor performs better in Biomedicine, where the output terms should be more specific. These results confirm our assumption that, although both NC-value and TermExtractor make use of domain-independent features for ranking, their performance varies across domains and applications. At the same time, combining our domain coherence approach (*PostRankDC*) with our *Basic* method in a post-ranking step, in the method *PostRankDC*, displays a more stable behaviour, achieving the best performance on the Computer Science domain (Krapivin) and results similar to those of the best method in Biomedicine (GENIA) and Agriculture (FAO).

3.7.3 Evaluating the semantic grounding of expertise topics

In this section we compare two approaches for grounding expertise topics on DBpedia. The first approach (A1) matches a candidate DBpedia URI with an expertise topic, using the string as it appears in the corpus. The second approach (A2) makes use of the lemmatised form of the expertise topic. Stemming was also considered but this approach resulted in a decrease in performance, as stems are more ambiguous [1].

In order to evaluate our URI discovery approach, we build a small gold standard dataset by manually annotating 186 expertise topics with DBpedia URIs. These expertise topics are extracted from a corpus of scientific publications from the Semantic Web research field. This corpus is described in more detail in the next chapter, in Section 4.3.1.2. First of all, we note that we were only able to find a corresponding DBpedia

[1] An approach based on a semantic web search engine that uses keyphrase search to find structured data was also considered, restricting the search to the DBPedia domain. The results were disappointing because only a limited number of retrieved results can be analysed. Often, the relevant DBpedia concept does not appear in the top results.

concept for about half of the analysed expertise topics. This is because we are dealing with a general knowledge datasource that has a limited coverage of specialised technical domains.

Approach	True Positives	False Positives	True Negatives	False Negatives
A1	93	4	82	7
A2	90	1	85	10

Table 3.8: DBpedia URI extraction results

The number of positive and negative matches is shown in table 3.8. Although both approaches achieve similar results in terms of F-score, the approach that makes use of lemmatisation (A2) achieves better precision, as can be seen in table 3.9. To extract descriptions or definitions of concepts we rely on the *dbpedia-owl:abstract* property, or the *rdfs:comment* property in the absence of the former. For now we are interested in English definitions, therefore we consider triples tagged with the property *lang='en'* alone. Even though English descriptions are available for a larger number of topics, this tag is not always present. Therefore, we can only retrieve descriptions for a smaller number of topics. A manual analysis of matching errors showed that expertise topics that include an acronym (e.g. "NLG system" instead of "Natural Language Generation system") are more difficult to associate with a DBpedia concept, as often acronyms are ambiguous.

Other general purpose data sources, such as Freebase [1], or domain-specific data sources can be linked in a similar manner. A complex problem that we do not address in this work is the disambiguation of an expertise topic when multiple concepts from different domains can be matched. Usually, DBpedia provides a disambiguation page for such cases. In our implementation we did not analyse concepts that redirect to a disambiguation page, grounding only those expertise topics that are specific enough to be used in a single domain.

[1] http://www.freebase.com/

Approach	Precision	Recall	F-score
A1	0.96	0.93	0.94
A2	0.99	0.90	0.94

Table 3.9: Precision and recall for DBpedia URI extraction

3.8 Summary

In this chapter, we proposed an approach to identify intermediate level terms through domain modelling. We had a closer look at two important requirements for a term extraction system, the extraction of domain-specific terms and the need for domain-independent approaches. We argued that approaches that make use of contrastive corpora are only suitable for updating existing terminology resources with more specific terms and not for summarising expertise areas as required in Expertise Mining. The main contributions described in this chapter are the following:

1. A method for extracting top level terms from a domain corpus
2. A novel domain coherence metric based on semantic relatedness with a domain model
3. A method to construct context patterns for selecting candidate terms
4. A novel application-based evaluation for term extraction systems

Also, as a first step in the direction of making use of background knowledge from the Linked Open Data cloud for Expertise Mining, we associated expertise topics with corresponding concepts from DBpedia. Experiments presented in Section 3.7 show that the performance of current term extraction approaches depends on the domain although these systems make use of domain independent features. Instead, our domain coherence approach based on a domain model performs well across domains, while the performance of the two benchmark systems varies across domains. The application-based evaluation introduced here allows us to investigate more subtle differences between term extraction methods on a larger number of candidate terms. At the same time, this method of evaluation allows us to analyse both precision and recall, making possible a more thorough evaluation than previously done.

A major insight from this chapter is that our method for selecting general terms from a corpus can be used as a global generality measure. In the following chapter, we investigate its application for inducing generalisation hierarchies from text.

4

Constructing topical hierarchies for Expertise Mining

In this chapter we investigate the role of knowledge organisation in Expertise Mining. In particular, we analyse how knowledge structures, in the form of topical hierarchies, can be derived from domain corpora and applied to improve expert finding and expert profiling. First, a qualitative analysis of several automatically constructed taxonomies is provided. This analysis is done in the context of their application to Expertise Mining for tasks such as expert profiling and expert finding. This part of the thesis is organised as follows. We introduce topical hierarchies and we propose an approach to construct a topical hierarchy using a novel generality measure in Section 4.1. Then, we turn to the tasks of expert profiling and expert finding in Section 4.2, discussing several ways to leverage a topical hierarchy for Expertise Mining. The experimental setup for this chapter is presented in Section 4.3, followed by an experimental evaluation of the proposed methods in Section 4.3. This chapter is concluded with a summary of our findings in Section 4.5.

4.1 Constructing topical hierarchies from text using a global generality measure

A large number of hierarchical relations are readily available in various lexical resources and on the Linked Open Data cloud, but not all domains are covered and existing taxonomies are often too small to fully describe a domain. Automatically constructed taxonomies can provide a higher level of granularity that could make possible the identification of different levels of competency within a domain. Previous work on building

concept hierarchies from text was either concerned with constructing application specific hierarchies, or with building exhaustive hierarchies that can inform a wide number of applications. The former results in relatively small-sized and focused hierarchies. For example, a document browsing hierarchy has to take into account the usability and navigability of the structure, as well as other interface-related restrictions [SHR07]. Similarly, a large-sized taxonomy is a considerable obstacle for document classification as a large amount of training data is required, leading to increased misclassification rates due to the large number of classes [WUH+08]. Application-driven development of taxonomies allows a more objective evaluation, using performance metrics specific to each considered application.

The latter approach focuses on collecting a large number of pairs of concepts that are related through a hierarchical relation. This approach results in structures that have a wide coverage, but that are not necessarily organised in a tree like structure, but rather in a noisy graph. Currently, there is an increased interest in going beyond the collection of relation pairs, by analysing the global structure of the graph and filtering inconsistencies between local edges [KH10, NVF11]. Because exhaustive taxonomies are not developed in the context of a specific application, they are generally evaluated through comparison to existing manually constructed taxonomies or through expert assessment. Although most studies choose one of these methods to evaluate lexical taxonomies, both approaches have their limitations. Gold-standard evaluation is limited to a predefined list of concepts, and is less appropriate for highly specialised technical domains. Manual evaluation performed by domain experts is limited to the evaluation of pairs of concepts, because longer paths are considerably more complex and are deemed infeasible for a user survey.

Indirectly evaluating taxonomies in the context of a given application is a possible solution to these fallacies. Unlike other applications such as document browsing and document classification, Expertise Mining is not restricted to small-sized hierarchies, and can theoretically benefit from using large-scale taxonomies. Taxonomies are applied in Expertise Mining to inform expertise measures, therefore there are no restrictions related to the size of the taxonomy. Expertise Mining is an area that was not previously considered for an application-driven evaluation of taxonomies.

A **topical hierarchy** is a hierarchy of nodes, where nodes correspond to expertise topics and links between nodes correspond to *broader* relations between expertise

topics. This type of relation is defined in the SKOS[1] vocabulary as a hierarchical link between two concepts that indicates that one concept is in some way more general ("broader") than the other ("narrower"). Nodes at a higher level represent more general information, while the nodes at a lower level contain more detailed information. The root of the hierarchy should be the topic that best describes the domain, that is the name of the field. Ideally, each term should have a well defined position in the hierarchy of a domain, and there should be at least a path that connects it to the root. Automatically identifying the name of a field is not trivial, because even between experts there is not always an agreement on the name of a field. For example, one of the research areas that we consider in this chapter is known as Computational Linguistics, Natural Language Processing as well as Human Language Technology.

More formally, a topical hierarchy is a directed acyclic graph $G = (V, E)$, where $V = \{v_1, v_2, .., v_n\}$ is the set of vertices, each corresponding to an expertise topic, and E is the set of directed edges between terms in which $E_{a,b}$ indicates that there is a directed edge between v_a and v_b.

4.1.1 Analysis of existing hierarchies

Depending on their purpose, taxonomies are constructed using different types of relations between terms including generic-specific relations, instance relations, or meronymy relations. This is an important design consideration in our application scenario as well, therefore we performed a qualitative analysis of existing taxonomies in the context of Expertise Mining. With this objective in mind, we selected a classification hierarchy for Computer Science and several manually constructed taxonomies about tourism, finance, and plants. Each of these hierarchies are contrasted with our definition of topical hierarchies and are analysed based on their applicability to Expertise Mining.

For our analysis of the ACM Computing Classification System [2] we chose as an example the Machine Learning subtree[3], which is displayed in Figure 4.1 using a graph visualisation tool. The size and the colour of the nodes, as well as the labels are proportional with the degree of the node. The nodes with the highest degree are presented in red and the nodes that have the smallest number of neighbours in blue.

[1] SKOS vocabulary:`http://www.w3.org/2004/02/skos/core`

[2] ACM Classification System: `http://www.acm.org/about/class/2012`

[3] The visualisations presented in this chapter are constructed using Gephi [4]

We used the Force-Atlas and the Yifan Hu [Hu05] algorithms to generate the layout of the graph. The same settings are used to construct the visualisations of all the other graphs discussed in this chapter.

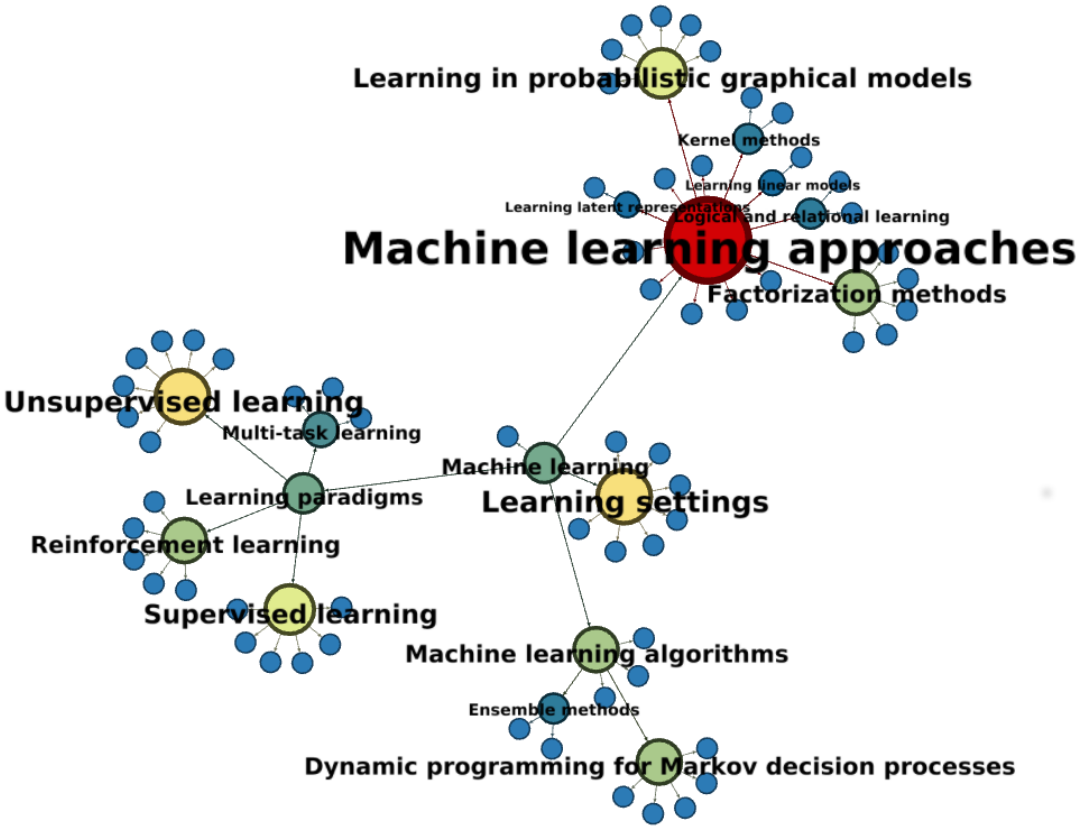

Figure 4.1: The Machine Learning subtree from the ACM Computing Classification System

At the center of the graph you can see the name of the field, *Machine learning*, which is also the root of the hierarchy. Directly connected to it are a small number of broad categories including *Machine learning approaches* and *Machine learning algorithms*. The majority of the nodes are multi-word terms, due to the highly technical and specialised nature of the field. In the original structure, relations between nodes are not named. A manual inspection shows that only a subset of the relations present in this hierarchy can be interpreted as *is-a* relations. This relation is defined in the RDF schema of WordNet[1] using the relations *hyponymOf* and *hypernymOf*. Although

[1]WordNet RDF schema available at: `http://www.w3.org/2006/03/wn/wn20/schemas/wnfull.rdfs`

some of the relations that can be seen in Figure 4.1 are *is-a* relations (e.g., the relation between *Supervised learning* and *Learning paradigms*), a more loosely defined relation such as the *skos:broader* relation describes better these edges.

The ACM Classification System respects our definition of a topical hierarchy, but it has the limitation that it contains only a small number of carefully selected terms. This is to ensure that the resulting structure that should describe the Computer Science field in its entirety remains manageable for a human indexer. For instance, the subtree for the Natural Language Processing research area is rather superficial, having only eight subordinate nodes (i.e., *Information extraction*, *Machine translation*, *Discourse, dialogue and pragmatics*, *Natural language generation*, *Speech recognition*, *Lexical semantics*, *Phonology / morphology*, and *Language resources*) and a depth of one level. This hierarchy is too coarse grained to be effectively used for measuring expertise and for constructing informative expert profiles in more specialised areas.

Most work on constructing lexical taxonomies from text is concerned with the construction of strict *is-a* hierarchies, but these structures are less informative for Expertise Mining as we will see from the following examples. We analysed a couple of manually constructed *is-a* taxonomies from the tourism and finance domains, described in [CHS05]. These hierarchies are shown in Figure 4.2 and Figure 4.3, respectively. Both taxonomies are a directed acyclic graph and not a tree, with several nodes connected to more than one parent. The relatively small-sized tourism hierarchy is connected in one component through the addition of an artificial *root* node, while the larger finance taxonomy has a large number of disconnected components that are not connected to this root. In both cases the root node can not be simply renamed with the name of the domain, because this would require the introduction of a different type of relation than *is-a* with the subordinate nodes.

The root node of the tourism hierarchy connects abstract concepts such as *thing*, *non-material thing*, *situation*, *spatial concept* and *qualitative time concept* that are not domain specific, with a domain-specific concept, *accommodation equipment*, that should intuitively be at a lower level in the hierarchy compared to the previous concepts. The same can be observed in the finance taxonomy, where the root node connects general domain nodes such as *mathematical concept*, *situation*, *thing*, and *spatial thing*, with a domain-specific node, *intangible*. Broad concepts that are part of a top ontology are

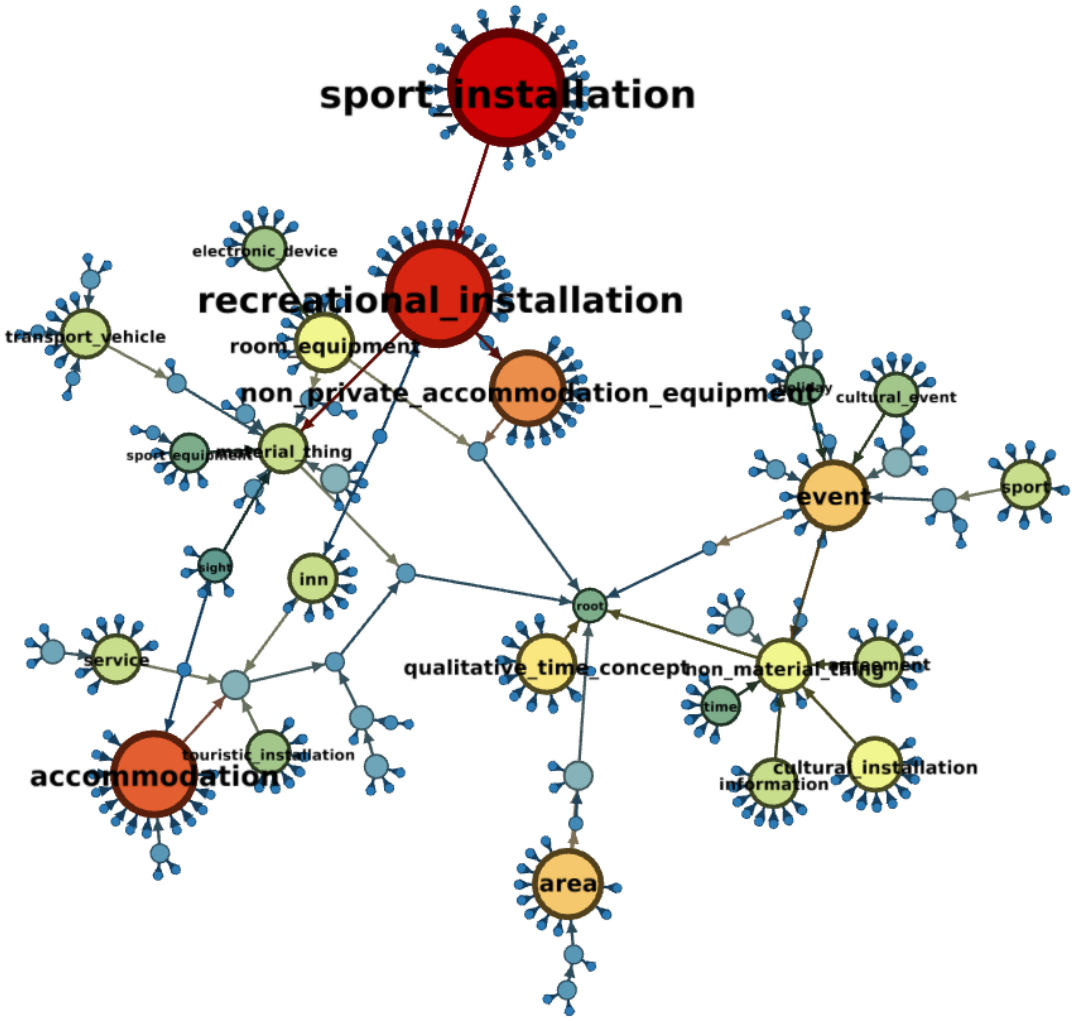

Figure 4.2: Hand-crafted *is-a* tourism taxonomy

central in hypernym-hyponim taxonomies but these concepts are not relevant to Expertise Mining. This is because commonsense knowledge or knowledge that is acquired as part of a general education does not form the object of expertise.

Most studies about extracting taxonomies from text attempt to reconstruct the hyponymy structure available in WordNet. Figure 4.4 shows a subset of this hierarchy for the plants domain, which was previously used to evaluate the taxonomy construction algorithm presented in [KH10]. Compared to the tourism and the finance taxonomies discussed above, the WordNet nodes considered in this evaluation are more specific. In this case, the links with nodes from a top level ontology are not considered relevant and

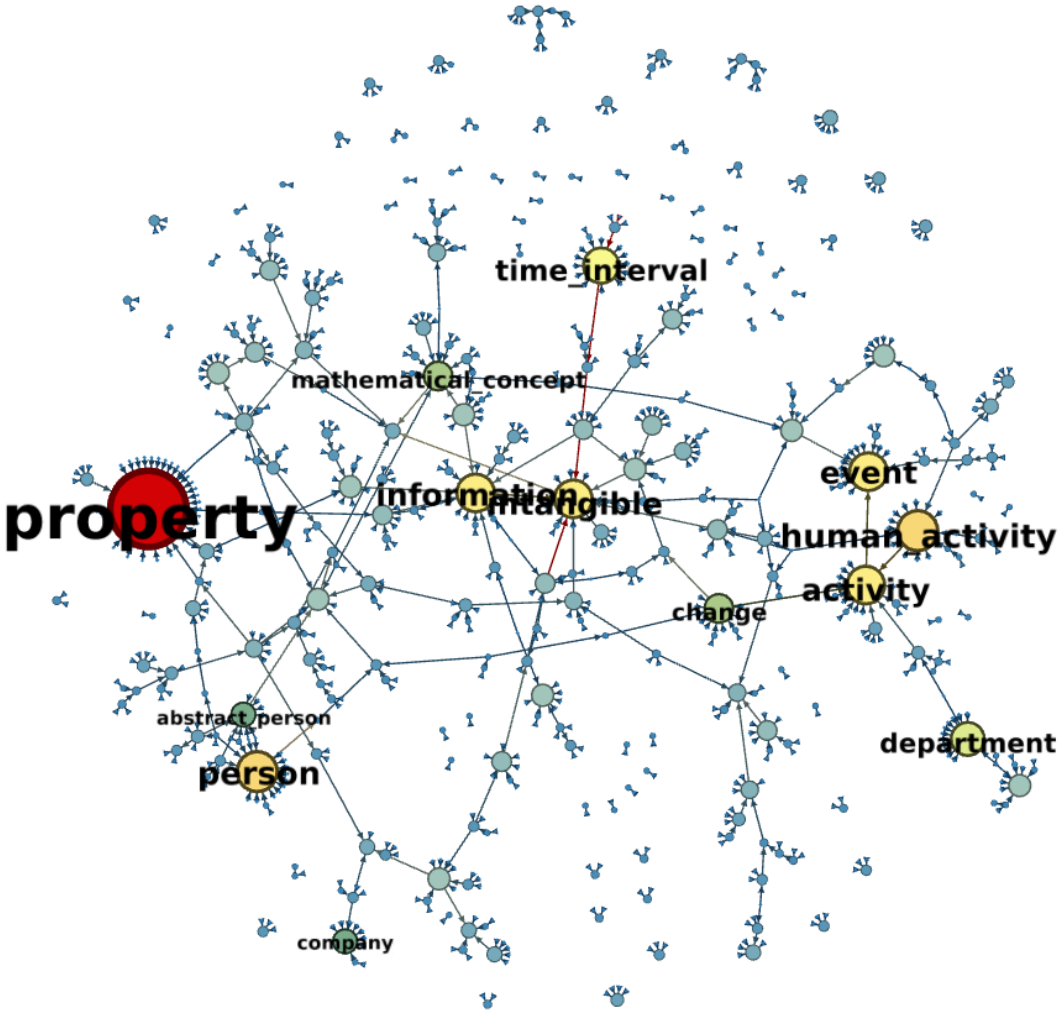

Figure 4.3: Hand-crafted *is-a* finance taxonomy

the broadest term is the root of the taxonomy, the node *plants* which is domain specific. For this reason, WordNet hierarchies selected in this way seem to be more appropriate for Expertise Mining but this resource has low coverage of technical domains.

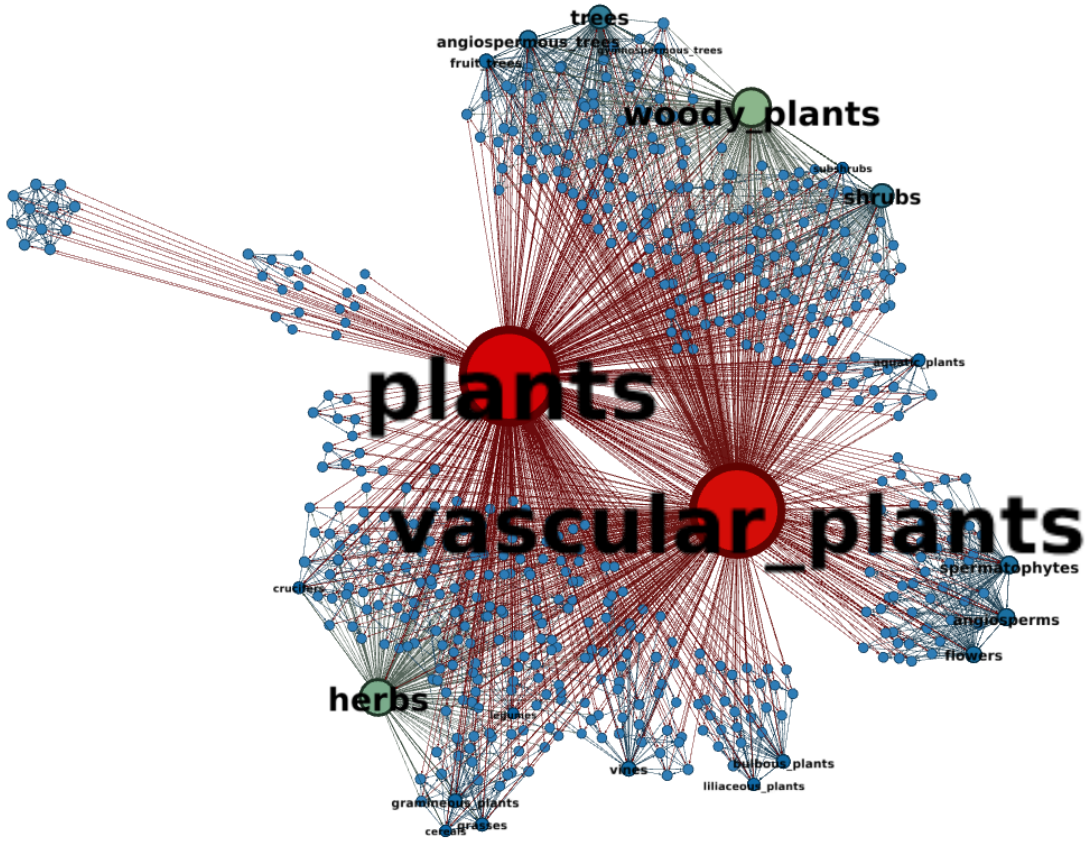

Figure 4.4: WordNet *is-a* taxonomy for the plants domain using doubly-anchored patterns

A large proportion of the total number of edges from this taxonomy of plants are links between the two most highly connected nodes in the graph, the *plants* and the *vascular plants* nodes. About 47% of the edges are connected to these two main nodes. A similar observation can be made about the two other WordNet hierarchies used as gold standard in [KH10], about animals and vehicles. In the animals taxonomy 40% of the edges are connected to 3 main nodes (i.e., *animals*, *vertebrates*, *chordates*), while in the vehicles taxonomy 42% of the edges are connected to the *vehicles* and *wheeled vehicles* classes. This would be equivalent with connecting a large number of subtopics directly with the name of the field.

An algorithm that attempts to reproduce the WordNet hierarchy, is deemed successful if it is capable to collect a large number of instances of a class. For example, in the case of the Wordnet plants taxonomy, an algorithm that is able to find all the instances of the two main classes, *plants* and *vascular plants*, would achieve a recall of 47%. While this type of edges are interesting for a general purpose resource such as WordNet, these edges are less informative when used for distinguishing different levels of competency in a given field. This is because the resulting hierarchies have a shallow depth with the majority of nodes connected to a few main classes. An important focus in a system that attempts to reproduce WordNet hyponymy relations is instance gathering, and less the construction of comprehensive classification structures as needed for expert finding and for exploratory search of expertise.

4.1.2 Measuring global generality within a domain

In the previous chapter we proposed a measure for selecting generic words that are representative for a domain, and that can be used to construct a domain model. Naturally, such a generality measure can have multiple applications. In this chapter we apply this method to measure the term generality instead of selecting general words for a domain model, and we provide a different interpretation of this measure in the context of constructing topical hierarchies. Similar to the measure for domain coherence introduced in the previous chapter, we define the global generality g of a term t as:

$$g(t) = \sum_{\tau \in T} Sim(t, \tau) \tag{4.1}$$

where τ is a term from the set of extracted terms T, and $Sim(t, \tau)$ is the similarity between the two terms. In this definition, the generality of a term is computed based on how close the term is related to a large number of terms from the domain. The intuition is that a generic term is often used with a large number of different terms, while a specific term is mostly used with a small number of closely related terms. This is a natural characteristic of a topical hierarchy. If we consider the path length between two terms in a topical hierarchy as a measure of similarity, the terms placed on the top levels are connected through a short path with a large number of nodes. On the contrary, the leaves of the hierarchy are connected through short paths only with a smaller number of nodes from the same subtree and their ancestors, but have relatively

longer paths to other nodes. This measure is a global measure of generality as it does not account only for the relation between two terms as done in previous work, but it takes into consideration the relation with all the other terms from the domain as well.

Several measures of similarity can be applied in Equation 4.1, but Pointwise Mutual Information (PMI) achieved the best results on the dataset introduced in Section 3.6.2.1 when compared to other popular measures of similarity such as log-likelihood. Therefore, we define the global generality measure as:

$$g(t) = \sum_{\tau \in T} PMI(t, \tau) \tag{4.2}$$

Which can be rewritten by replacing the formula for PMI as:

$$g(t) = \sum_{\tau \in T} log\left(\frac{P(t, \tau)}{P(t) \cdot P(\tau)}\right) \tag{4.3}$$

where $P(t, \tau)$ is the probability that the word t appears in the context of the term τ, $P(t)$ is the probability of appearance of t, and $P(\tau)$ is the probability of appearance of τ.

4.1.3 A graph-based algorithm for constructing topical hierarchies

The global generality measure described in the previous section can be used to assess if a term is more generic than another term, but it is not sufficient by itself to identify edges for a tree-like topical hierarchy. A possible solution is to use this information in combination with a graph-pruning algorithm, such as the one proposed in [NVF11], which takes as input a highly connected, noisy graph and filters inconsistencies between edges. The main difference compared to this work is that we do not consider strictly defined *is-a* relations but we enrich more loosely defined relatedness relations with generality information. Because we do not make use of extraction patterns but of co-occurrence information, our approach has the advantage that it can identify hierarchical relations which are not explicitly mentioned in text. Also, this method can be applied to smaller domain-specific corpora because it does not require large amounts of text to ensure an acceptable recall.

Relations constructed based on semantic relatedness have lower precision than relations extracted using Hearst patterns [Hea92]. Therefore we filter relations for terms that are mentioned together in less than a minimum number of documents, to make

sure that the relations are not limited to a small number of documents. This threshold affects the general connectivity of the graph and has to be adjusted based on the size of the corpus. For a small corpus, the threshold should be small to allow relatively rare terms to be connected to at least one other term in the graph. When enough documents are available, a larger threshold ensures that the initial graph contains relations between strongly connected terms.

The workflow for learning a topical hierarchy using a global generality measure is summarised in Figure 4.5.

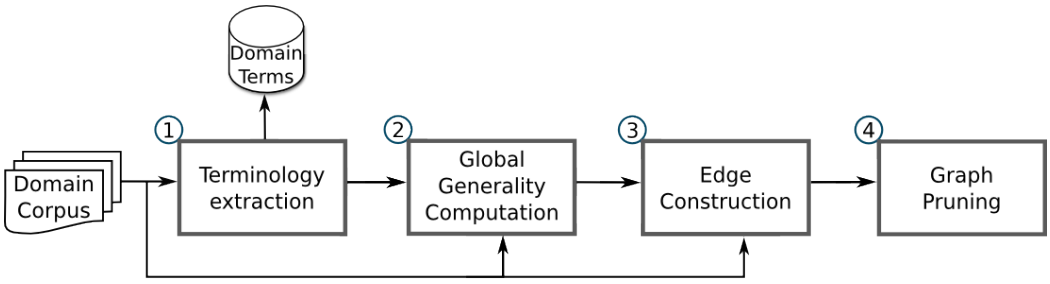

Figure 4.5: Learning workflow for constructing a topical hierarchy using a global generality measure

In the first step, expertise topics are extracted from a domain corpus using the method presented in Section 3.4.4, and added as nodes in the graph (1). This results in a set of vertices V, where each vertice corresponds to an expertise topic. Next, we compute the global generality measure for each term, incorporating the semantic relatedness with all the other terms from the hierarchy (2). In our implementation, an index for co-word analysis is used to measure relationship strength between two research terms. This index is defined as follows:

$$w_{ij} = \frac{|D_{ij}|}{|D_i| \cdot |D_j|} \tag{4.4}$$

where $|D_i|$ is the number of documents that mention the term t_i in our corpus, $|D_j|$ is the number of documents that mention the term t_j, and $|D_{ij}|$ is the number of documents in which both terms appear. Our implementation relies on domain corpora alone and does not require additional data from the web, avoiding the need for a domain filtering step as done in [NVF11].

The global generality of a term is used in the edge construction step to construct a directed graph (3). At the end of this step, edges are added to the graph for every pair

of terms t_i and t_j if the weight w_{ij} is larger than an empirically set threshold. Edges are added for all the pairs that appear together in a minimum number of documents. Edge direction is inferred based on the global generality of each node. The edge $E_{i,j}$ is added to the graph if $g(t_i) > g(t_j)$, otherwise the edge $E_{j,i}$ is added. In other words, each edge is directed from a more generic term to a more specific one.

The edge construction step results in a large number of edges and a highly connected graph, with several cycles and multiple parents for each node. At this stage, any type of semantic relations can connect two nodes, including associative and ad-hoc relations. Additionally, we add edges based on string inclusion, directing the edges from shorter terms, that are more generic, to longer, more specific terms. For example, and edge will be added between the nodes *machine translation* and *statistical machine translation* because the former nodes is embedded in the later node. The weight of the edges based on string inclusion is set to a maximum value as the edges inferred in this way are more reliable than edges constructed based on co-occurrence. We only consider the case when the shorter term appears as head in the longer phrase. This is an inexpensive and reliable method that was previously used to construct lexical hierarchies [NMWP99].

The noisy graph obtained in this way is used as input for the fourth step, that eliminates the cycles, allowing at most one parent for each node (4). The Chu-Liu/Edmonds algorithm [CL65, Edm67] for optimal branching is used in the fourth step for pruning the graph. An optimal branching is a rooted tree where every node but the root has in-degree 1, and that has a maximum overall weight. In case there are multiple disconnected components in the graph, the following steps are performed separately for each component. This algorithm requires as input a manually selected root node. In our implementation, the root node is considered to be the node with the maximum global generality value, that is the most generic node in the graph. We disconnect all the other nodes that have no incoming edges, to make sure there is just one possible root in the graph. This is a requirement of the graph pruning which is implemented to construct a tree from a graph that has only one root node that has no incoming edges.

Our edge weighting strategy, that weights the edges based on the relation strength between two vertices, favours closely related terms during the pruning step. Each resulting component is connected to the main component by connecting local roots to the main root node. The graph pruning algorithm constructs a tree structure where the root is the most generic expertise topic, and the leaves are specific terms.

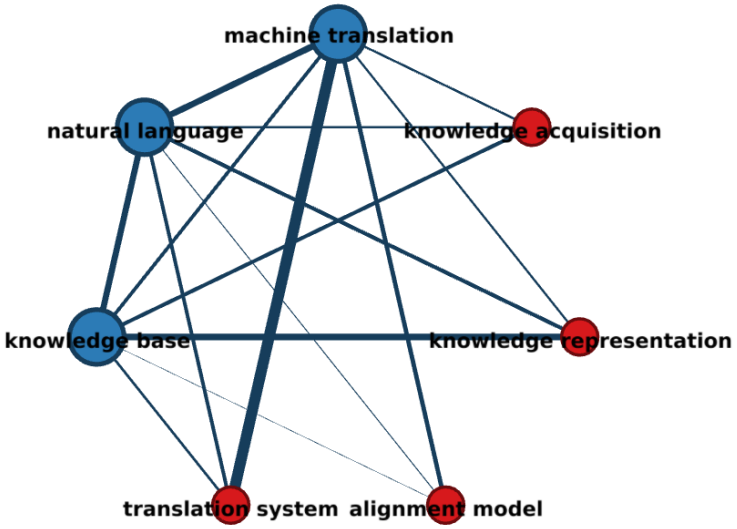

Figure 4.6: Undirected edges between seven terms in Computational Linguistics

Let's consider for example that the following expertise topics are extracted from a corpus in the Computational Linguistics domain: *natural language*, *machine translation*, *translation system*, *alignment model*, *knowledge base*, *knowledge representation*, and *knowledge acquisition*. Weighted undirected edges can easily be added between these terms based on their semantic relatedness, as can be seen in Figure 4.6.

Example terms	Global generality
natural language	1
machine translation	0.5
knowledge base	0.26
translation system	0.2
knowledge representation	0.15
knowledge acquisition	0.08
alignment model	0

Table 4.1: Global generality values for selected terms in Computational Linguistics

The global generality values for each of the terms considered in our example are presented in Table 4.1. In this list, the most general node based on its global generality is the *natural language* node and the most specific node is the *alignment model* node. The *natural language* term is more generic because it forms the object of study of the whole field, while *alignment model* is a term which is used specifically in the *machine*

translation area. Global generality values are used to find the direction of each edge, transforming the initial undirected graph to a directed graph as can be seen in Figure 4.7. For example, the initially undirected edge between *natural language* and *machine translation* is directed from the broader term *natural language* with a global generality of 1 to the narrower term *machine translation* that has a global generality of 0.5.

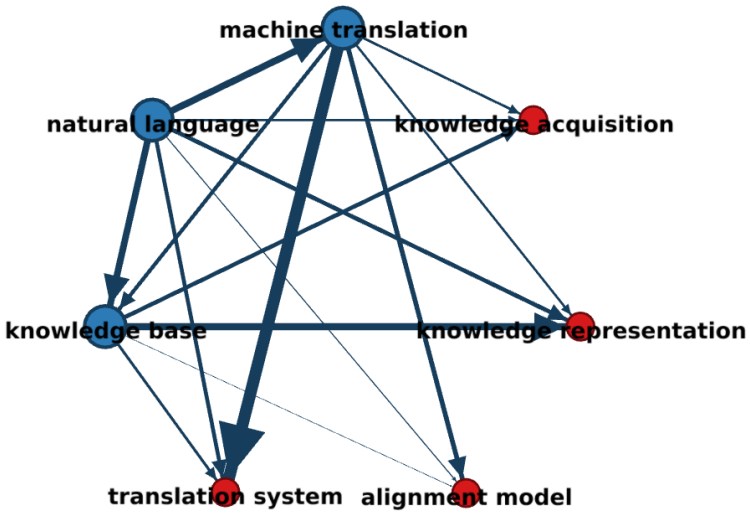

Figure 4.7: Directed edges between seven terms in Computational Linguistics

In our example, the most generic term is the node *natural language*, which is selected as a root. The graph pruning algorithm constructs a tree with the root node *natural language*, that has broader terms at the top of the hierarchy and specialised terms as leaves. The resulting hierarchy for our example is shown in Figure 4.8. Closely related terms such as *knowledge base*, *knowledge acquisition*, and *knowledge representation* are clustered in the same subtree of the hierarchy, placing the broader term, *knowledge base*, as a parent of the other two more specific terms. Note that the relations between the nodes are not *is-a* relations but the more general SKOS *broader* relation. For example we can not say that *machine translation is-a natural language*, but rather that *natural language* is *broader* than *machine translation*.

Compared to the taxonomies analysed in Section 4.1.1, this topical hierarchy is more appropriate for Expertise Mining because it contains only technical terms that are suitable expertise topics in the field of Computational Linguistics. Additionally, such a structure can be used to derive content based measures of expertise based on the relations between terms. For example, using the small hierarchy provided as example

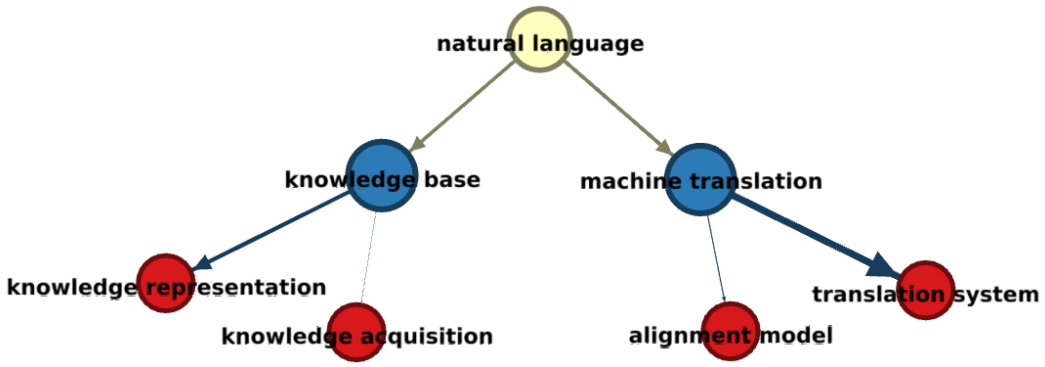

Figure 4.8: Topical hierarchy for seven terms in Computational Linguistics

in this section, we can infer that an expert in *machine translation* is familiar with translation systems and alignment models, as we will see in the following section.

4.2 Applying topical hierarchies to Expertise Mining

In this section, we discuss how automatically extracted expertise topics and relations between them, structured as a topical hierarchy, can be used to inform Expertise Mining tasks such as expert profiling, discussed in Section 4.2.1 and expert finding, presented in Section 4.2.2. This section is partially based on [BB10b, Bor10].

4.2.1 Expert profiling

An expert profile is a description of a person's expertise and interests, that can inform the selection of an expert for a given task. In this work, whenever we refer to an expert profile, we mean a topical profile and not a person's contact information. A motivation for constructing expert profiles is that although a person frequently writes about a subject area, a person is rarely an expert on every aspect of a topic [MM07b]. Therefore, a structured expert profile provides context to an expertise topic, showing which aspects of a topic are covered by an expert. We start by providing a formal definition of an expert profile, than we briefly discuss a method for selecting expertise topics for expert profiles and we propose a relevance-based method for constructing expert profiles.

Following [BdR07], we define the topical profile of a person as a vector of expertise topics along with scores that measure the expertise of that individual. The expert

profile p of an individual i is defined as:

$$p(i) = \{S(i, t_1), S(i, t_2), ..., S(i, t_n)\} \tag{4.5}$$

where t_1, t_2,...,t_n are the expertise topics extracted from a domain-specific corpus.

A first step in constructing expert profiles is to identify terms that are appropriate descriptors of expertise. A recent study [BBB+13] identified several requirements for an expert profile including coherence, completeness, conciseness and diversity. Another requirement mentioned in this study is that expertise topics have to be at the right level of specificity. The methods proposed in Chapter 3 are specifically designed for this purpose, because they extract terms that are specific to a domain, but that are general enough to refer to a knowledge area instead of a specific term. Using more generic terms allows us to construct a complete but concise profile for an expert. A large number of expertise topics is extracted for each document, but only the top ranked keyphrases are considered for expert profiling. Keyphrases are assigned to documents by combining the overall termhood rank of a candidate term with a measure of relevance for each document, as described in Equation 3.6. In our experiments, only the top 20 keyphrases are used for expert profiling.

Once a list of expertise topics is identified, we can proceed to the second step of expert profiling, the assignment of scores to each expertise topic for a given expert. We rely on the notion of relevance, effectively used for document retrieval, to associated expertise topics with individuals. A person's interests and expertise are inferred based on their authored documents. Each expertise topic mentioned in one of these documents is assigned to their expert profile using an adaptation of the standard information retrieval measure tf-idf [BYRN99]. The measure is applied on a corpus of aggregated documents.

Each individual i is represented by a virtual document d_i that is constructed by aggregating all the documents authored by that person. This allows us to estimate the relevance of a term for an individual by computing the relevance of a term for the virtual document d_i. Documents are aggregated for each individual i, resulting in a set of aggregated documents D defined as follows:

$$D = \{d_{i1}, d_{i2}, ..., d_{in}\} \tag{4.6}$$

Where n is the total number of individuals.

The tf score is defined in Equation 4.7 as the frequency $f(t, d_i)$ of a term t in the aggregated document d_i authored by an individual i, normalised by the frequency $f(t, D)$ of the term in the whole set of aggregated documents D.

$$tf(t, i) = \frac{f(t, d_i)}{f(t, D)} \tag{4.7}$$

The idf score for a term t and a set of aggregated documents D is defined in Equation 4.8 as the total number of individuals n divided by the number of individuals that mention the term.

$$idf(t, D) = \frac{n}{|\{d_i \in D : t \in d_i\}|} \tag{4.8}$$

In this way, the $tfidf$ scoring function combines tf and idf as follows:

$$tfidf(t, i) = tf(t, i) \cdot idf(t, I) \tag{4.9}$$

Finally, an expertise topic is added in the expert profile of a person using the following scoring function:

$$S(i, t) = termhood(t) \cdot tfidf(t, i) \tag{4.10}$$

Where $S(i, t)$ represents the score for an expertise topic t and an individual i, $termhood(t)$ represents the rank computed in Section 3.4.2 for the topic t and $tfidf(t, i)$ stands for the tf-idf measure for the topic t on the aggregated document of an individual i. In this way, we construct profiles with terms that are representative for the domain as well as highly relevant for a given individual.

A topical hierarchy can further provide useful information for constructing diverse profiles, that have the right level of specificity. For example, when several closely related expertise topics are relevant to a person, only one of them should be reported. At the same time, the leaves of the hierarchy are often too specific to efficiently summarise expertise and should not be included in the profile. We leave this direction of research for future work.

4.2.2 Expert finding

Expert finding is the task of identifying a list of people who are knowledgeable about a given expertise topic, according to [BBA+07]. The expert finding task does not necessarily require the extraction of expertise topics because the user searches for experts by providing a query that describes the topic of interest. Even so, in this section we will show that expert finding results can be improved when expertise topics and the relations between them are automatically extracted from text.

Within an organisation or community, several competent people have to be ranked based on their relative expertise. Documents written by a person are used as an indirect evidence of expertise, assuming that an expert often mentions his areas of interest. Data-driven approaches such as the ones developed by the information retrieval community [PC06, BdRA06, SH08], do not attempt to map domain knowledge to identify knowledgeable individuals. Instead, associations between people and expertise topics are uncovered based on co-occurrences between topics and people in the same context. Information retrieval approaches do not make use of performance indicators to measure expertise, as done in competency management, but rely instead on the broader notion of relevance [BdRA06].

Relevance-based approaches rely on exact matches of expertise and are less reliable for general expertise topics that are mentioned less often in text. If a person does not mention the name of a domain, they will not be considered an expert although they often discuss about related subfields. For example, although a person does not explicitly mention *Semantic Web* in their documents, if they frequently discuss about *Linked Data*, *RDF data*, *ontologies* and *SPARQL* they are likely to be Semantic Web experts. Although latent semantics methods such as LSA [DDL+90], pLSI [Hof99], LDA [WC06] address this problem, they are more appropriate for single-word terms and have to be extended to n-gram models to cover longer terms, which are computationally more expensive.

In this chapter we extend existing relevance-based measures of expertise, using topical hierarchies to introduce a semantically motivated measure of expertise. Again, we consider the tf-idf measure described in the previous section in Equation 4.9 to measure the relevance of a given expertise topic for a person. People are represented by an aggregated document that is constructed by concatenating all the documents

authored by a person. Therefore, the relevance score $R(i,t)$ that measures the interest of an individual i for a given topic t is defined as:

$$R(i,t) = tfidf(t,i) \tag{4.11}$$

Expertise is closely related to the notion of experience. The assumption is that the more a person works on a topic, the more knowledgeable they are. This performance indicator is similar to the frequency indicator mentioned in [Paq07]. We estimate the experience of a person based on the number of documents that they wrote about an expertise topics. It is only those documents for which the expertise topic is extracted as a top ranked keyphrase that are considered. Let $D_{i,t}$ be the set of documents authored by the individual i, that have the expertise topic t as a keyphrase. Then, the experience score $E(i,t)$ is defined as:

$$E(i,t) = |D_{i,t}| \tag{4.12}$$

where $|D_{i,t}|$ is the cardinality, or the total number of documents, in the set of documents $D_{i,t}$. It can be argued that it is not only the number of documents that indicates expertise, but the quality of those documents as well. For example, in a peer-review setting, the impact of a publication measured using citation counts is often used as an indicator of publication quality. Similarly, page rank can be used as a quality indicator for web pages, the number of comments for blogs, the number of retweets for tweets, the number of followers for users. But each of these indicators is specific to the content type and have to be investigated separately depending on the domain, therefore we leave the integration of document quality measures for future work.

Relevance and expertise measure different aspects of expertise and can be combined to take advantage of both features as follows:

$$RE(i,t) = R(i,t) \cdot E(i,t) \tag{4.13}$$

Both the relevance score and the experience score rely on query occurrences alone. A topical hierarchy, such as the one constructed in Section 4.1.3, can provide valuable information for improving expert finding results. When the subtopics of an expertise topic are known, we can evaluate the expertise of a person based on their knowledge of specialised fields. A previous study showed that experts have increased knowledge at more specific category levels than novices [TT91]. We introduce a novel measure for

expertise called Area Coverage that measures whether an expert has in depth knowledge of an expertise topic. Let $Desc(t)$ be the set of descendants of a node t, then the Area Coverage score $C(i,t)$ is defined as:

$$C(i,t) = \frac{|\{t' \in Desc(t) : t \in p(i)\}|}{|Desc(t)|} \tag{4.14}$$

where $p(i)$ is the profile of an individual i constructed using the method presented in the previous section. In other words, Area Coverage is defined as the proportion of descendants of a query that appear in the profile of a person. Area coverage is larger than zero only for topics that have more than one descendant, therefore this measure does not contribute to finding experts for specialised topics that appear as leaves in a topical hierarchy.

Finally, the score $REC(i,t)$ used to rank people for expert finding is defined as follows:

$$REC(i,t) = RE(i,t) \cdot C(i,t) \tag{4.15}$$

This score combines several performance indicators, measuring the expertise of a person based on the relevance of an expertise topic, the number of documents about the given topic, as well as his depth of knowledge of the field, also called Area Coverage.

4.3 Experimental setup

In this section we present the experimental setup used for an empirical investigation of research question RQ2. This research question is concerned with extracting and applying topical hierarchies to Expertise Mining. We discuss our choice of evaluation metrics and we present three datasets annotated for Expertise Mining in Section 4.3.1. The benchmarks used in our experiments are described in Section 4.3.2.

4.3.1 Evaluation metrics and datasets

This section describes several evaluation metrics that are commonly used to evaluate expert profiling and expert finding systems, in Section 4.3.1.1. Additionally, we describe three domain-specific datasets that were gathered based on data about organisers and program committees of a large number of workshops in Computer Science in Section 4.3.1.2.

4.3.1.1 Evaluation metrics

Given the tasks at hand, several evaluation measures for document retrieval can be used. The expert profiling and the expert finding tasks are evaluated based on the quality of ranked lists of expertise topics and of experts, respectively. From an evaluation point of view, this is not different from evaluating a ranked list of documents. The most basic evaluation measures used in information retrieval are precision and recall. These measure the proportion of retrieved documents that are relevant and the proportion of relevant documents that are retrieved, respectively. Other frequently used effectiveness measures include:

Precision at N (P@N) This is the precision computed when N results are retrieved, which is usually used to report early precision at top 5, 10, or 20 results.

Average Precision (AP) Precision is calculated for every retrieved relevant result and then averaged across all the results.

Reciprocal Rank (RR) This is the reciprocal of the first retrieved relevant document, which is defined as 0 when the output does not contain any relevant documents.

To get a more stable measurement of performance, these measures are commonly averaged over the number of queries. In our experiments, we report the values for the Mean Average Precision (MAP), and Mean Reciprocal Rank (MRR). In this setting, recall is less important than achieving a high precision for the top ranked results. It is more important to recommend true experts than to find all experts in a field.

4.3.1.2 Domain-specific datasets based on workshop program committees

Evaluating expert search systems remains a challenge, despite a number of data sets that have been made publicly available in recent years [BCdVS07, BBA+07, SdVC07]. Traditionally, relevance assessments for expert finding were gathered either through self-assessment or based on opinions of co-workers. On one hand, self-assessed expert profiles are subjective and incomplete, while on the other hand opinions of colleagues are biased towards their social and geographical network. A recent study showed that,

in an organisational setting, people are more likely to recommend as experts peers from their collaboration network or people that are geographically close [SB11].

We address these limitations by exploiting expertise data generated over the course of several decades in a peer-review setting, more specifically data about 590 conference workshops in Computer Science. Conference workshops are focused events, organised around a narrow set of (interrelated) topics. The process of composing program committees for workshops is particularly interesting, as organisers have a broad knowledge of the domain and of domain experts. Workshop organisers tend to be topical experts, as well as the program committee (PC) members that are invited to review submitted contributions to the workshop. A peer-review setting alleviates the problem of subjective assessments of expertise, as several community members have to reach an agreement. Typically, workshop proposals are subject to careful scrutiny before being accepted as part of a conference. Whether the organisers and the PC members are well-recognised experts in their field is an important criteria for the acceptance of a workshop proposal. We limit ourselves to data about workshops, as conferences are more broad in scope than a workshop, making the assignment of expertise topics to specific committee members less reliable.

Another limitation of existing datasets for expert search is the sparsity of topic-person associations. This is partially due to the fact that relevance assessments are gathered through heavily involved and time-consuming interviews with experts. Instead, calls for workshop papers (CfPs) are more easy to collect and readily provide rich descriptions of expertise topics, that can be associated with organisers and PC members. Workshop CfPs are a valuable source of topic-expert associations, as they contain extensive lists of domain experts. This information can be used as a gold standard for expert finding, but only as a silver standard for expert profiling as the profiles inferred from workshop topics are incomplete. It is unlikely that a researcher will be invited as a workshop PC member in each of the areas that they are an expert in.

The three new test collections presented in this section, are focused around entire research communities in a specific research domain as opposed to one organisation. Previous test collections for expert search focus on searching a single organisation for experts and expertise. Different types of organisations are considered, such as research institutes [BCdVS07, SdVC07] and universities [BBA+07, SB11]. Instead, our test collection is focused around entire research communities in specific research domains,

with members that are geographically dispersed and that have similar interests to some extent. In particular, we gathered data produced by two communities with a long tradition in Computer Science, Information Retrieval (IR) and Computational Linguistics (CL), as well as the more recent Semantic Web (SW) community. Because the members of a research community are more similar than the members of an organisation, these datasets require methods that distinguish expertise at more fine-grained levels. Take for example the university use case, where expertise topics as broad as mathematics, philosophy, or physics are informative enough to distinguish between experts from different departments. When analysing expertise in a research community such as the IR or the CL community, more fine-grained expertise topics are required.

We first provide some general considerations about the datasets, and then we describe some of the particularities of each dataset. For each test collection, we describe the process of collecting the documents, creating the topics and producing the relevance assessments. Table 4.2 contains an overview of the main characteristics of our three test collections, including information about the total number of documents, workshops, and authors from each research area. Each dataset consists of a corpus of documents and information about their authors along with a collection of workshops from the same research areas that can be used as a basis for gold standard evaluation. The CL and the SW datasets are collected and maintained by the research communities that initially published the works. Therefore, relatively clean metadata about academic events and scientific publications, including full-text content, are directly available from the same location. The same cannot be said about the IR collection that has to be gathered from different sources including DBLP, Google Scholar, and ArnetMiner [TZY+08]. This resulted in a smaller coverage of full-text publications as can be seen in Table 4.2.

Organisers and PC members correspond naturally to (a subset of) the relevant experts on the topic of the workshop. To generate a topic description for each workshop, we extracted the title of the workshop as well as a short and a long description of the purpose of the workshop. The long description of the workshop was typically taken from the starting page and covered the complete description of the goals and focus of the workshop (except for the areas of interest). The short description typically corresponded to the first paragraph of the long description: one or two sentences containing

a concise teaser description of the workshop[1]. In addition, we extracted the areas of interest from each workshop website, which are often presented as a bullet-point list of research areas. Extracting this workshop information was done by a group of annotators, researchers in Computer Science, which have previous annotation experience. Annotators were provided with the description of the workshop and were briefly instructed by email to identify expertise topics from call for papers.

It is often the case that a workshop is organised with the intent to create a platform for communication between research areas with overlapping interests. For example the workshop on the emergent topic of "Computational Neurolinguistics" organised by the CL community in 2010 was meant to bring together researchers from the areas of computational linguistics and cognitive neuroscience that have an interest in machine learning methods. Such workshops have PC members with expertise backgrounds that match one of these areas, or a combination of both. To allow a more fine-grained identification of expertise topics, beyond workshop titles, we manually annotated each workshop from the IR, CL and SW areas with expertise topics. The list of organisers and PC members for each workshop served as our relevance assessments. Members of organising committees and program committees are considered to be relevant experts for the topic of the workshop. Five different annotators were involved in constructing the topic set for the IR collection; for the CL and SW collections, 2 annotators were involved.

The datasets proposed in this work are used to investigate different applications of Expertise Mining. The CL and the SW datasets provide a large number of documents and a list of associated terms extracted from workshop descriptions that can be used for term extraction or expertise topic extraction evaluation. The main difference between these two tasks is that generally expertise topics have to be broad enough to summarize a knowledge area, while terms can be more specific. Another application that can be addressed is the assignment of experts to program committees using workshop descriptions. In this way, workshop organisers can identify PC members based on their interests mentioned in previous publications. Expert finding is a similar task that takes as input more focused keyphrase-based descriptions of expertise topics instead of

[1]Distinguishing between what constitutes a short and long version of the description was left up to the individual annotator; we did not check for inter-annotator agreement, although incidental inspection suggested a consistent extraction process.

a workshop description. Finally, extracted terms can be assigned to topical profiles, i.e., expert profiling. Profiling a person requires the identification of areas of skills and knowledge that best describe their interests and expertise.

	IR	CL	SW
#documents	24,690	10,921	2,311
% of full-text documents	54.1%	100%	55%
#workshops	60	340	190
#unique authors	26,098	9,983	4,480
#authors/document	2.7	2.2	3.3
#experts/workshop	14.9	25.8	24.9
#expertise topics	488	4,660	6,751

Table 4.2: Overview of workshop based test collections (IR = Information Retrieval, CL = Computational Linguistics, SW = Semantic Web)

We now present specific details about each dataset.

Semantic Web workshop dataset

The first dataset is a corpus of scientific publications from Semantic Web conferences[1] that were published in the proceedings of several conferences, including: ISWC, EKAW, ESWC, WWW, ASWC, and I-Semantics [2]. This dataset is available through a public SPARQL endpoint. Full content is available for 55% of the publications, with abstract only available for the rest of the publications. Each researcher and document is identified through a unique URI. Workshops along with information about title, year, description, website, organisers, and PC members, can be queried by selecting all the events of type *WorkshopEvent*. In some cases this information is missing or incomplete and we used the workshop website where it was still active to manually extract the data. We were not able to do this for a subset of workshops that did not have active websites any more. A large number of terms that appear in workshop descriptions, describing the main topics of interest of the workshop, were manually annotated. For example the workshop with the title "1st International Workshop on Stream Reasoning" from 2009 was annotated with terms including *stream reasoning* and *reasoning* that appear in the

[1] Available at `http://data.semanticweb.org`, last accessed July 15, 2013.
[2] The complete list of conferences can be found here: `http://data.semanticweb.org/conference`

title of the workshop, as well as *network monitoring*, *data streams*, *traffic engineering*, and *sensor networks* that are mentioned only in the CfP. These terms are associated with each PC member and workshop organiser to be used as relevance assessments for expert finding and expert profiling.

Computational Linguistics workshop dataset

The ACL Anthology Reference Corpus[1] is a dataset made available by the Computational Linguistics community. Documents are collected from events such as ACL, EACL, NAACL, SemEval, ANLP, EMNLP, Coling, HLT, IJCNLP, and LREC. Metadata about researchers and documents is provided in an XML format and authors are identified using their names as they appear in publications. Several name variations are used to identify the same person. Workshop descriptions are easily identified in the directory structure of the dataset,as they are grouped together under the same folder, and organised based on the year when the event was held. Each workshop is associated with a document that describes that event. One annotator manually extracted information about the organisers, PC members, year, and title from these documents. At the same time, each workshop was annotated by two annotators with expertise topics that describe the main areas of interest of the workshop. For example, the workshop on "Biomedical Information Extraction" from 2009 was annotated with terms such as: *biomedical information extraction*, *biomedicine*, *health care*, *healthcare delivery*, *personalized medicine*, and *clinical narrative*. All of these expertise topics were explicitly mentioned in the workshop description.

Information Retrieval workshop dataset

The third dataset covers the related fields of information retrieval (IR), digital libraries (DL), and recommender systems (RS). To construct a test collection covering all of these research fields, we used the DBLP Computer Science Bibliography[2], a computer science bibliography website that tracks the most important journals and conference proceedings in computer science. Our initial motivations for constructing

[1] Available at `http://acl-arc.comp.nus.edu.sg/`, last accessed July 15, 2013.
[2] Available at `http://dblp.uni-trier.de/`, last accessed July 9, 2013.

a test collection around DBLP were two-fold: (1) the fields of IR, DL, and RS are well-covered in DBLP, and (2) a special version of the DBLP data set, augmented with citation information, is available from the team behind ArnetMiner, which allows for investigations into the use of citation information for expert search.

To make the augmented DBLP collection suited to expert search evaluation, we need realistic topic descriptions as relevance judgments at the expert level. Workshops organized at major conferences covering the fields of IR, DL, and RS are used to collect relevance judgments. To identify relevant workshops, we visited the websites of the CIKM, ECDL, ECIR, IIiX, JCDL, RecSys, SIGIR, TPDL, WSDM, and WWW conferences, which have substantial portions of their program dedicated to IR, DL, and RS. We collect links to workshop websites for all workshops organized at those conferences between 2001 and 2012. This resulted in a list of 60 different workshops with websites that were still online at the time of writing[1].

As a starting point, a test collection covering the aforementioned fields was constructed by using the augmented DBLP data set released by the team behind ArnetMiner[2]. This data set is a October 2010 crawl of of the DBLP data set containing 1,632,442 different papers with 2,327,450 citation relationships between papers in the data set[3]. As this augmented data set contains publications from all fields of computer science, we filtered out all publications not belonging to IR, DL, and RS by restricting the collection to publications in relevant journals, conferences, and workshops. This additional filtering step resulted in a final list of 78 *curated venues* (core plus additional)[4] covering a total of 24,690 publications.

In addition to citation information, the augmented DBLP data set is also extended with abstracts wherever available. However, the team behind ArnetMiner was only able to add abstracts for 33.7% of the 1.6 million publications (and 43.5% of the 24,690 publications in our test collection). We therefore attempted to download the full-text versions of al 24,690 publications using Google Scholar. We constructed a search query consisting of the last name of the first author and the full title without surrounding quotes[5]. We then extracted the download link from the top results returned by Google

[1]The list of 60 active workshops can be viewed at `http://itlab.dbit.dk/~toine/?page_id=631`.
[2]ArnetMiner:`http://arnetminer.org/`
[3]Available at `http://arnetminer.org/DBLP_Citation`, last accessed July 9, 2013.
[4]The list of curated active workshops is available at `http://itlab.dbit.dk/~toine/?page_id=631`.
[5]A preliminary test on just the publications from the core venues showed that adding quotes around the publication title decreased recall from 80.3% to 70.86%.

Scholar (if available). We were able to find download URLs for 14,823 of the 24,690 publications in our filtered DBLP data set for a recall of 60.04%, where recall is defined as the percentage of papers in our filtered DBLP data set that we could find download URLs for.

While this is not as high as we would like, it does represent a substantial improvement over the percentage of abstracts present in the augmented DBLP data set. Moreover, a recall rate of 100% is impossible to achieve as tutorials, keynote abstracts, and even entire proceedings are typically not available online in full-text, but they are present in the DBLP data set. Around 90.15% of download URLs obtained in this manner were functional, which means we were able to download full-text publication files for 13,363 publications (or 54.12% of our entire curated data set). We performed a check of 100 randomly selected full-text files to see if these were indeed the publications we were looking for and achieved a precision of 97% on this sample. We therefore assume that the false positive rate of our approach is acceptably low.

4.3.1.3 The UvT Expert dataset

The domain-specific datasets introduced in the previous section contain a large amount of scientific publications that are focused on a given field of research. These datasets allow us to investigate expertise in a given research community. Previous studies on Expert Search put more effort into analysing expertise inside knowledge-intensive organisations. The UvT dataset, introduced in [BBA+07], contains information about the employees of Tilburg University, that is collected from a publicly accessible database. The UvT dataset is more heterogeneous than the workshop datasets, as it gathers information from manually provided summaries of research and courses, personal homepages, as well as publications. Table 4.3 gives an overview of the size of the UvT dataset. The UvT dataset is topically more diverse than the datasets presented in the previous section, covering broad areas of study such as economics, law, information technology, public administration or criminology. Although expertise topics are available in Dutch and English, in our experiments we considered only 981 expertise topics available in English.

About 7% of the publications are available as full content, with most publications being available as citations only. The large and diverse number of expertise topics combined with the limited availability of textual descriptions leads to challenges related

	RD	CD	PUB	HP
#documents	316	840	27,682	6.724
#people	316	318	734	318

Table 4.3: Overview of the UvT Expert Dataset, including Research Descriptions (RD), Course Descriptions (CD), Publications (PUB), and Personal Homepages (HP)

to data sparseness. Nevertheless, the expert finding and expert profiling tasks are easier on the UvT dataset, as we will see in Section Sect:evalExTE. This is due to the fact that most documents are high quality summaries of expertise and that there are a relatively smaller number of people in the dataset. Additionally, there is a small number of overlapping expert profiles, because in a university less people have similar interests than in a research community.

4.3.2 Baseline approaches for Expertise Mining

The approaches proposed in this section are evaluated against two information retrieval methods for expert finding and expert profiling. Both methods model documents and expertise topics as bags of words and take a generative probabilistic approach, ranking expertise topics t by the probability $P(t|i)$ that they are generated by the individual i [BAdR09]. The same probability is used for ranking expertise topics in a person's profile, as well as for finding knowledgeable people for expert finding. The first model ($LM1$) constructs a multinomial language model θ_i for each individual, over the vocabulary of documents authored by them. This is similar to our approach that computes the relevance of a topic for an individual on a document that aggregates all the documents authored by that person.

The assumption is that expertise topics are sampled independently from this multinomial distribution. Therefore, the probability $P(t|i)$ can be computed as:

$$P(t|i) = P(t|\theta_i) = \prod_{w \in t} P(w|\theta_i)^{n(w,t)} \tag{4.16}$$

where $n(w,t)$ is the number of times the word w appears in the expertise topic t. Smoothing using collection word probabilities is applied to estimate $P(w|\theta_i)$. The smoothing parameters are estimated with an unsupervised method, using Dirichlet

smoothing and the average number of words associated with people as the smoothing parameter.

The second model ($LM2$) considered as baseline estimates a language model θ_d for each document from the set D_i of documents authored by the individual i. Words from an expertise topic t are sampled independently, summing the probabilities to generate an expertise topic for each of these documents. In this case, the probability $P(t|i)$ is calculated using the following equation:

$$P(t|i) = \sum_{d \in D_i} P(t|\theta_d) = \sum_{d \in D_i} \prod_{w \in t} P(w|\theta_d)^{n(w,t)} \tag{4.17}$$

Again, the probability $P(w|\theta_d)$ is estimated by using the same unsupervised smoothing method. In this case, the smoothing parameter for Dirichlet smoothing is the average document length in the corpus.

4.4 Experimental evaluation

In the previous section, we presented the experimental setup for this chapter, including evaluation metrics, datasets and baseline approaches that are used in our experiments. We continue this section by presenting an experimental evaluation of the methods proposed in Section 4.2. First, we evaluate our method for extracting terms vs. expertise topics proposed in Chapter 3, on the workshop datasets introduced in this chapter. Then, we compare our approach for constructing topical hierarchies with existing methods for extracting taxonomies from text in Section 4.4.2. Finally, we describe the outcomes of our system for expert profiling (Section 4.2.1) and expert finding (Section 4.2.2) in comparison with two language modelling benchmarks. We summarise our findings in Section 4.5.

4.4.1 Evaluation of expertise topic extraction

The previous chapter presented an extensive evaluation of the proposed method for term extraction using domain modelling. Generic tasks such as keyphrase extraction and index term assignment were considered for this purpose. In this section we evaluate our approach on the more specific task of extracting expertise topics. We hypothesise that our domain modelling approach introduced in Section 3.4.4 is suitable for extracting expertise topics.

To this end, we make use of the relatively large number of expertise topics annotated in the CL and SW datasets, which were described in Section 4.3.1.2. Our goal is to answer the question whether our domain modelling approach is more appropriate for extracting expertise topics than previously used relevance measures (RQ2.1). The IR dataset was not considered in this experiment, as a much smaller number of knowledge areas is available compared to the other two datasets. The standard information retrieval measure tf-idf was previously used to extract expertise topics for topic-centric expert search [JLsK+07, LYL08]. Therefore, we use this measure as a baseline approach in our experiments. We adopt the same methodology used to evaluate term extraction, which was discussed in section 3.6.1.2. Again, the evaluation measure applied is Precision at top K%, which is defined as follows:

Precision (P@K%) The Precision computed when K% of the results are retrieved, which is used to report precision after analysing the top 20%, 40%, 60%, 80%, and 100% results of the ranked list.

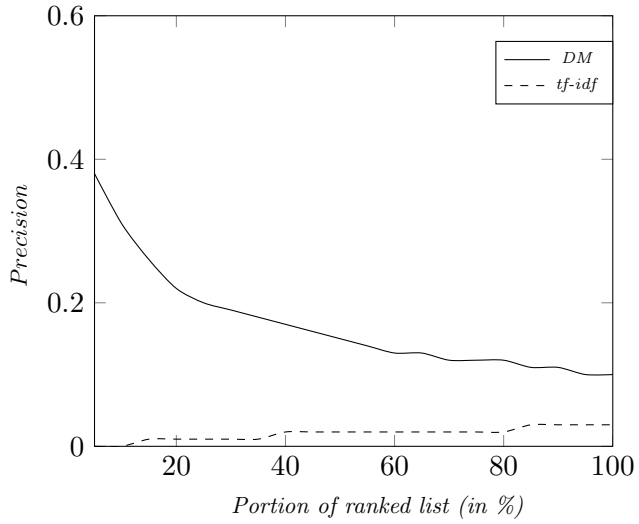

Figure 4.9: Precision for top 10k terms for the domain modelling approach and the tf-idf approach on the CL workshops dataset

Figure 4.9 shows the results for the CL corpus. Precision is calculated for the top ranked terms extracted using the domain modelling approach *DM* and the terms with the highest tf-idf value across the corpus. These results include partial matches,

considering correct any term that includes a gold standard term as a substring. For example the term *statistical machine translation* will be considered correct if the gold standard includes an expertise topic called *machine translation*. Our approach identifies a much larger number of expertise topics, with precision approaching 40% at the top of the list. This is only an estimate of precision, as our gold standard terms do not cover all the correct expertise topics, but only those that are mentioned in workshop CfPs.

Similar results can be observed for the SW dataset, presented in Figure 4.10, with the difference that the task is easier for both systems due to the fact that the SW corpus has a smaller size and is more focused. Also, there are a relatively higher number of annotated expertise topics for each expert. At the top of the list, almost half of the terms extracted using domain modelling are expertise topics. Again, a smaller number of expertise topics are identified using tf-idf. After the first 40% of the ranked lists, both approaches converge because the SW corpus is about ten times smaller than the CL corpus, in terms of full content publications available, therefore a smaller number of promising candidate terms is extracted.

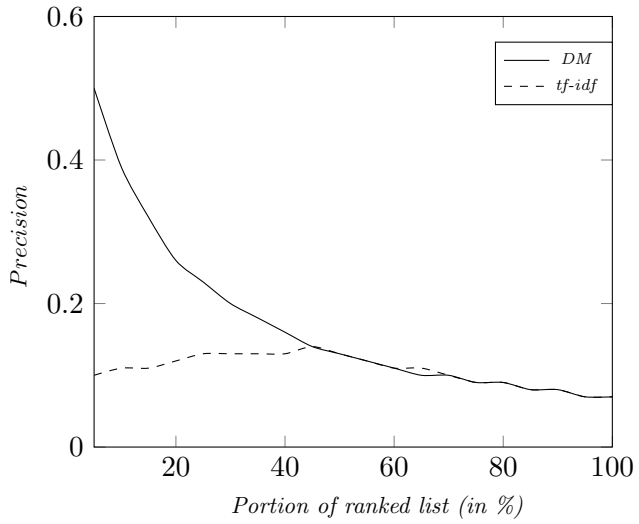

Figure 4.10: Precision for top 10k terms for the domain modelling approach and the tf-idf approach on the SW workshops dataset

A visual investigation of highly ranked terms based on tf-idf shows that these terms are more specific than expertise topics. For example, terms such as *chest radiographs*, *Lebanese hostage takers*, and *adhesive characters* are some of the terms that received high scores using tf-idf. These terms are highly relevant for a given document, but are

not representative for the domain. This explains the small number of matches at the top of the list. Instead, using our approach, terms such as *word sense disambiguation*, *speech recognition*, *statistical machine translation* are ranked higher. Terms extracted using domain modelling are more appropriate as expertise topics, because they have an intermediate level of specificity, that is they are specific to the domain at hand but they are broad enough to summarise different areas of expertise.

Overall, our approach achieves better results than previously used relevance based methods, as can be seen in figures 4.9 and 4.10. The experiments presented in this section confirm our hypothesis that a term extraction technique which relies on domain modelling is more suited for extracting expertise topics than previously used relevance measures such as tf-idf.

4.4.2 Constructing topical hierarchies vs. taxonomies

In this section we take a closer look at recent methods for taxonomy construction from text, discussing their applicability to Expertise Mining. In particular, we compare our approach for constructing a topical hierarchy with two approaches for taxonomy construction. The first approach is based on doubly anchored patterns [KH10], while the second one is the OntoLearn Reloaded approach [NVF11]. Both approaches rely on the massive amount of data available on the web to collect hyponym-hypernym relations between terms of interest. While the former approach uses doubly-anchored patterns and a root concept as seed to gather relationship pairs in a bootstrapping fashion, the latter approach exploits explicit *is-a* relations from term definitions.

Figure 4.11 shows the hierarchy generated in [KH10] when reconstructing the WordNet hierarchy for plants presented in Figure 4.4. Compared to occurrence-based methods for identifying relations, such as the subsumption method proposed in [SC99], pattern-based methods generally achieve high precision but lower recall. Additionally, patterns work better for single-word terms such as *plants* and *herbs*, that are frequently used on the web. Recall is considerably lower for multi-word terms, such as *vascular_plants* (the second largest node, at the right of the figure, in yellow) and *woody_plants* (the second topmost node, in blue). Although a similar number of edges are provided in the gold standard taxonomy for *plants* compared to *vascular_plants* and for *herbs* compared to *woody_plants*, only about half the number of edges were found for

the longer terms compared to single-word terms. This problem is even more pervasive in technical domains, where a large number of terms are multi-word expressions.

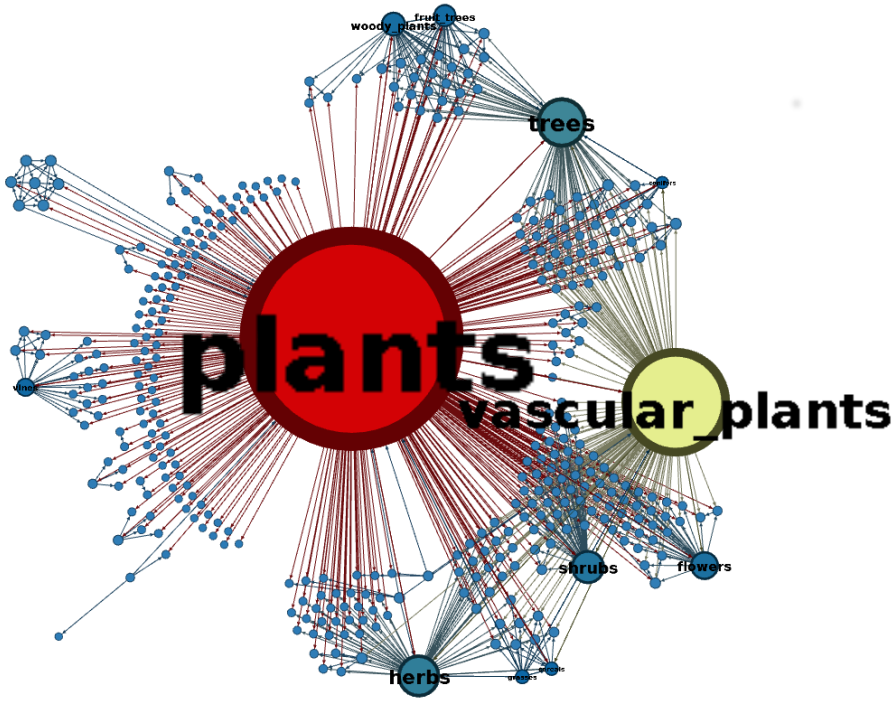

Figure 4.11: Automatically constructed *is-a* taxonomy for the plants domain

A more promising approach for constructing a hierarchy of technical terms is the OntoLearn Reloaded approach described in [NVF11]. This method is specifically designed to for constructing taxonomies for technical domains, Artificial Intelligence (AI). Additional information about the size of the graph can be seen in Table 4.4. The extracted taxonomy can be seen in Figure 4.12. Note that this is a different visualisation than the ones presented in [NVF11], but the underlying structure is the one made available online by the authors. The main difference is that in our visualisation the size and the colour of the nodes is used to represent the degree of a node. This was meant to highlight nodes that have a large number of descendants.

The root of the OntoLearn Reloaded taxonomy for Artificial Intelligence is the node *abstraction*. This node has as subordinate nodes other abstract terms such as *event*, *property*, *knowledge*, *communication*, *data/information*, *system*. These terms are domain relevant, but they are too broad to identify an expertise area. Although this approach has a higher coverage of technical terms than the pattern-based approach,

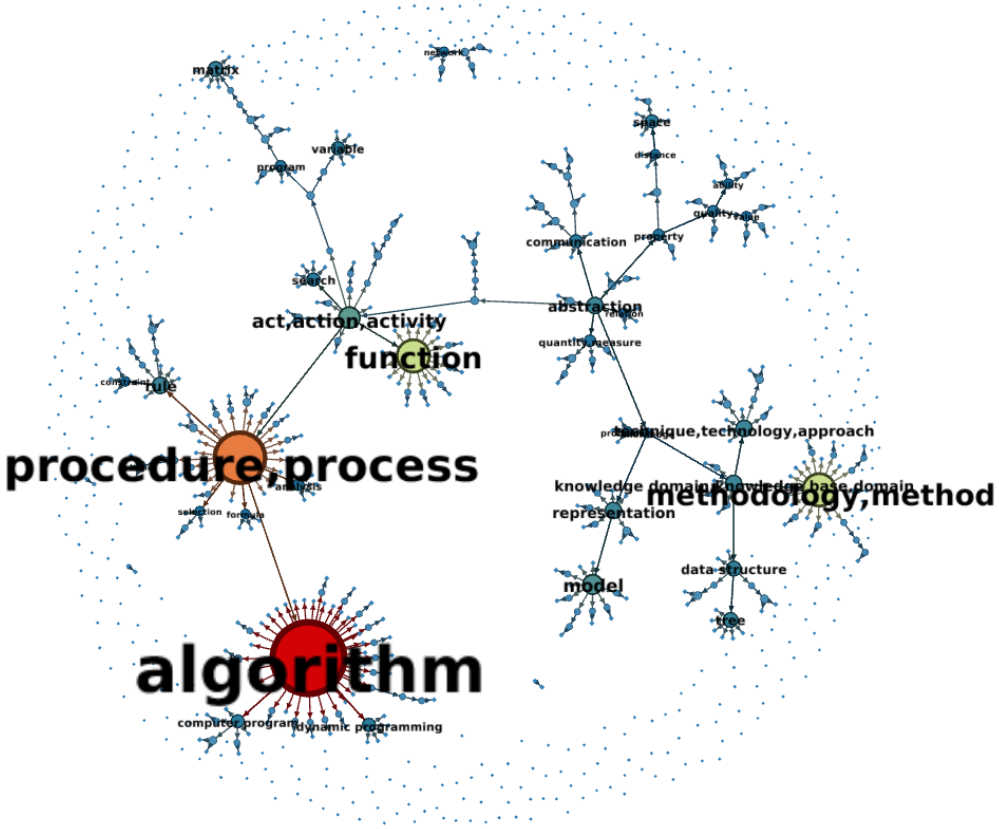

Figure 4.12: OntoLearn Reloaded *is-a* taxonomy for Artificial Intelligence

about half of the nodes are not connected to any other nodes. Furthermore, the six most connected nodes (i.e., *algorithm*, *function*, *methodology/method*, *procedure/process*, *act/action/activity*, *model*) are also abstract terms, not expertise topics. The edges between these six highly connected nodes and their direct descendants represent about 30% of the total number of edges in the graph. The taxonomy contains a large number of multiword terms as well, but this is not evident in our visualisation because these terms have a small number of child nodes and they appear closer to the leaves of the graph.

The same can be said about the OntoLearn Reloaded taxonomy constructed for a subfield of Artificial Intelligence, Computational Linguistics, that is shown in Figure 4.13. The same root node was identified, the *abstraction* node. This node is connected to a similar list of abstract concepts, including: *system*, *data/information*, *knowledge*, *communication*, *quantity/measure*, *event*, *property*. Again, about 40% of the nodes are

not connected by any edge. In this case, the six nodes with the highest degree are *procedure/process*, *model*, *rule*, *function*, *grammar* and *system*. With the exception of *grammar*, the other nodes are too general to be considered as acceptable descriptors of expertise. Altogether, these nodes are connected by about a quarter of the edges in the graph.

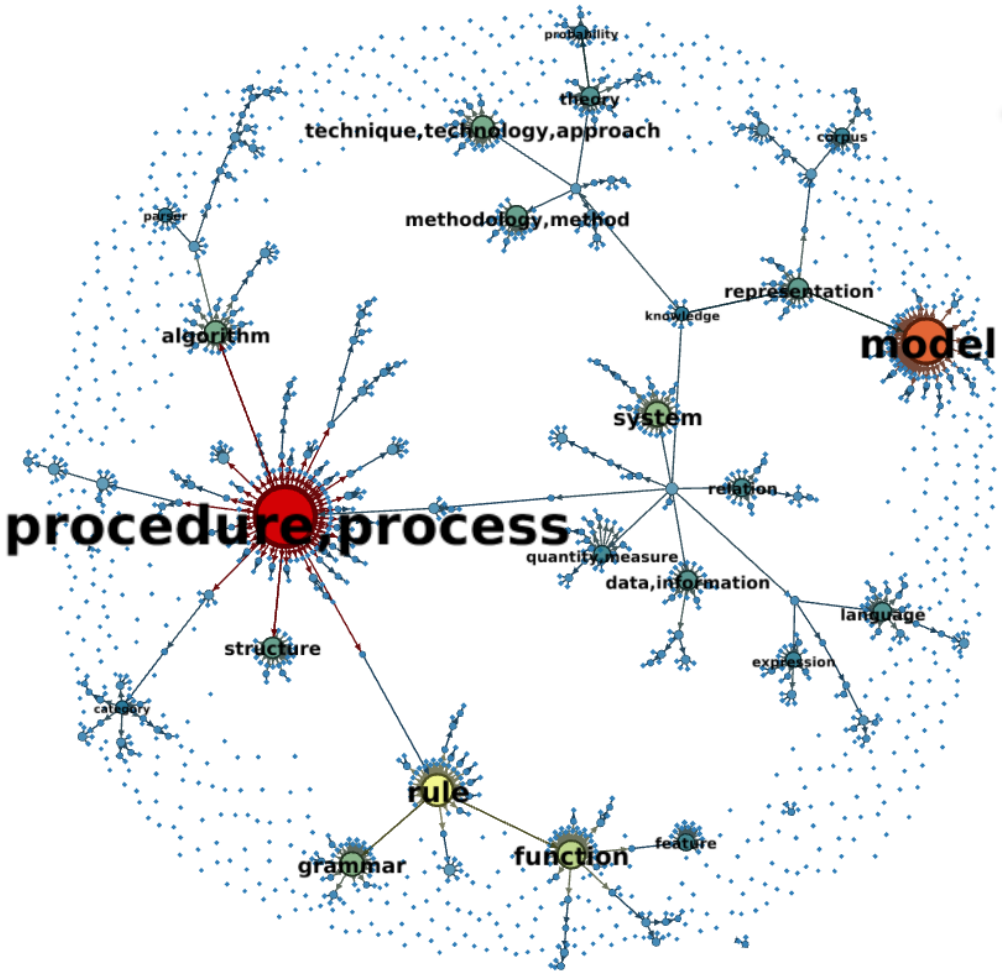

Figure 4.13: OntoLearn Reloaded *is-a* taxonomy for Computational Linguistics

To come back to our small example based on seven Computational Linguistics terms, used in Section 4.1.3, we analyse a subset of the OntoLearn Reloaded taxonomy, that is shown in Figure 4.14. The majority of the seven terms used in our example can be found in the OntoLearn Reloaded taxonomy as well, with the exception of *alignment model*, *natural language*, and *translation system*, which appear as part of longer terms

such as *statistical alignment model*, *machine translation system*, and *natural language processing*, respectively. The small taxonomy presented in Figure 4.14 was constructed by selecting the above mentioned nodes in the OntoLearn Reloaded taxonomy, as well as all ancestors up to their common root, which is the node *abstraction*. A first observation is that the OntoLearn Reloaded hierarchy is rich at the abstract levels, but more shallow for specialised terms. Most of the nodes that are closely related to the root in our hierarchy presented in Figure 4.8, appear as leaves in the OntoLearn Reloaded taxonomy, including the name of the field, *natural language processing*. This is a side-effect of iteratively searching for hypernyms in increasingly general definitions. The OntoLearn Reloaded approach enforces strict semantics on pairs of relations between terms, resulting in highly accurate structures, but that have a low coverage when identifying relations between specialised terms, such as expertise topics.

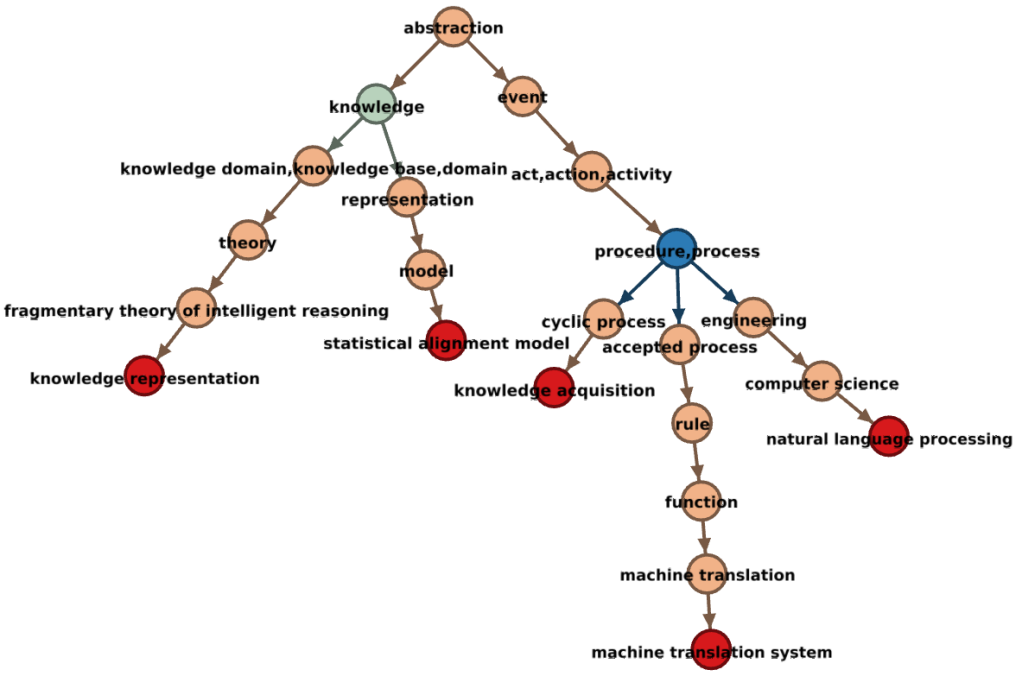

Figure 4.14: Ontolearn Reloaded taxonomy for seven Computational Linguistics terms

The application of automatically constructed is-a taxonomies to Expertise Mining is hindered by the low connectivity of the resulting graph, by the large number of abstract nodes in the top levels of the taxonomy, as well as by the considerable number of edges that are connected to a high level concept. In comparison, our algorithm for

Approach	Domain	#Nodes	#Edges
OntoLearn Reloaded	AI	868	415
OntoLearn Reloaded	CL	1626	926
Topical Hierarchy	CL	1626	1609

Table 4.4: Graph size of OntoLearn Reloaded taxonomies for Artificial Intelligence and Computational Linguistics compared to a topical hierarchy for Computational Linguistics

constructing a topical hierarchy, relies on co-occurrence information that is available for a much larger number of nodes, even when using a moderate-sized corpus. This allows us to avoid using out-of-domain corpora which is an additional source of noise that is likely to introduce out-of-domain nodes and edges.

Take for example the topical hierarchy presented in Figure 4.15, that is constructed using a Computational Linguistics corpus. The same number of nodes was considered as for the Ontolearn Reloaded taxonomy constructed for the same domain. The root of the tree is the node *natural language*, which is connected to two accepted names of the field, *natural language processing* and *computational linguistics*, through the node *language processing*. Several subfields, such as *speech recognition*, *information retrieval* and *statistical machine translation* can be identified as highly connected nodes. A much larger percentage of nodes can be connected using co-occurrence information than by relying on patterns or existing definitions from the web. It is only 1% of the nodes that are not connected in the topical hierarchy, compared to 40% in the case of the Ontolearn Reloaded taxonomy, although in our case all the terms used to construct a topical hierarchy are multi-word expressions and not generic single-word terms. Topical hierarchies, such as the one presented in Figure 4.15, have a more rich structure and a larger number of edges, than OntoLearn Reloaded taxonomies.

A disadvantage of relying on co-occurrence information is that this requires a sufficient amount of domain-specific documents. On the other hand, Ontolearn Reloaded can be applied to construct a taxonomy even for a single document because relations are extracted on the web.

We conclude that acquiring is-a relations is an easier task for single-word nodes and abstract nodes, than for expertise topics, that are longer, specific, technical terms. At a superficial investigation, topical hierarchies constructed based on co-occurrence information are more informative for the task at hand. In the following section, we

Figure 4.15: Topical hierarchy for Computational Linguistics

discuss an application-based evaluation of concept hierarchies, in the context of Expertise Mining. This application domain is well suited for this purpose, because relatively clean datasets can be gathered about expertise, as shown in Section 4.3.1. In this way, we can automatically evaluate longer paths between concepts, not just relation pairs as was previously done through user studies.

4.4.3 Expert profiling evaluation

In this section we investigate the RQ2.2 research question, which addresses the problem of constructing expert profiles directly from text, without making use of manually identified expertise topics. Our hypothesis is that expert profiles constructed using

automatically-extracted expertise topics will be at least equal or improve compared to expert profiles constructed using manually identified knowledge areas. The topic-centric approach (TC) for expert profiling proposed in Section 4.2.1 is evaluated on three domain-specific datasets, as well as a fourth dataset about the employees of Tilburg University, which are described in Section 4.3.1.2.

Our topic-centric approach can be applied for expert profiling without the need for controlled vocabularies, as expertise topics are directly extracted from text. Instead, the language modelling approaches used as a baseline in this section, can only be applied on datasets where such resources are readily available. Both baseline approaches described in section 4.3.2 are provided with gold standard expertise topics as input, while the topic-centric approach automatically extracts expertise topics from text.

We make use of the evaluation measures listed in Section 4.3.1.1, presenting the experimental outcomes in Table 4.7. First, we note that the results achieved for the IR dataset are higher in absolute terms than the results for the SW and the CL datasets. This is due to the smaller number of expertise topics that are available compared to the SW and the ACL datasets. The TC approach outperforms both the LM1 model and the LM2 model, on the ACL and the SW dataset. The most considerable improvement compared to the LM1 and LM2 models can be seen on the SW dataset. This is the most focused dataset out of the three domain-specific datasets, covering a relatively narrow and recent area of research.

Dataset	Measure	LM1	LM2	TC
	MAP	0.0256	0.0233	**0.0392**
CL	MRR	0.1857	0.2044	**0.2767**
	MAP	0.0082	0.0088	**0.0369**
SW	MRR	0.1271	0.1161	**0.3437**
	MAP	0.1052	**0.1679**	0.0879
IR	MRR	**0.3761**	0.3677	0.3364
	MAP	0.1299	**0.1380**	0.0459
UvT	MRR	0.3066	**0.3136**	0.1662

Table 4.5: Expert profiling results for the language modelling approach (LM) and the topic centric approach (TC)

The language modelling approaches achieve better results on the IR and the UvT

datasets, with the LM2 approach outperforming the LM1 approach on most measures. The gap between the language modelling approaches and the TC approach is more narrow on the IR dataset. Not surprisingly, our method for extracting expertise topics is under-performing when applied to a corpus that covers diverse expertise areas, such as the UvT dataset. Another possible explanation can be found in the number of documents that are available for each person. The LM1 and LM2 models achieve the worse results on the SW dataset, where only 8% of the people are associated with more than 3 documents.

4.4.4 Expert finding evaluation

We turn now to the RQ2.3 research question, related to the use of automatically constructed topical hierarchies for expert finding. Our assumption is that a topical hierarchy provides valuable information for expert finding, allowing us to rank experts based on the level of specificity of expertise topics mentioned in their documents. We test this hypothesis by comparing the Area Coverage measure of expertise, which is based on a topical hierarchy, with more simple relevance-based approaches. The experiments presented in this section make use of datasets and evaluation measures discussed in Section 4.3.

We compare several topic-centric methods for expert finding with two language-modelling baselines that were previously applied for this task, which were described in Section 4.3.2. The expert finding methods evaluated in this section include Experience (E), Relevance and Experience (RE) and Relevance, Experience and Area Coverage (REC). These methods are described in Equations 4.12, 4.13, and 4.15 respectively, in Section 4.2.2. To apply the Area Coverage measure of expertise, a topical hierarchy has to be provided for each domain. Therefore we construct a topical hierarchy using the algorithm proposed in Section 4.1.3 for each of the domains presented in Section 4.3.1.2, including Computational Linguistics (CL), Semantic Web(SW), Information Retrieval (IR), and Tilburg University (UvT).

A short summary about the constructed topical hierarchies for each domain is presented in Table 4.6. Depending on the size of each dataset, a different number of expertise topics is extracted and subsequently considered for constructing a topical hierarchy. The CL dataset is the largest dataset, allowing us to filter the edges in a pre-processing step based on the number of documents that provide evidence for the

Dataset	CL	SW	UvT	IR
#Nodes	15,000	5.000	5,000	4,000
#Edges	14,976	4,506	4.939	3,939
#Min Docs	3	1	1	3
Window size	5	50	50	5

Table 4.6: Graph size for topical hierarchies constructed for Computational Linguistics (CL), Semantic Web(SW), Information Retrieval (IR), and Tilburg University (UvT)

relation. An edge is added in the noisy graph only if at least three different documents provide evidence for the relation. This condition is not used for the other smaller datasets because it reduces the number of edges and the connectivity of the graph. For the same reason, the window size used to count co-occurrences of terms is larger for the smaller datasets than for the CL dataset.

The results for the expert finding task are presented in Table 4.7.

Dataset	Measure	LM1	LM2	E	RE	REC
CL	MAP	0.0071	0.0056	0.0335	0.0335	**0.0340**
	MRR	0.0631	0.0562	0.2734	0.2738	**0.2754**
	P@5	0.0202	0.0173	0.1340	0.1339	**0.1347**
SW	MAP	0.0070	0.0067	**0.0327**	0.0305	0.0314
	MRR	0.0528	0.0522	**0.2262**	0.2115	0.2095
	P@5	0.0182	0.0188	**0.1065**	0.0967	0.0994
IR	MAP	0.0599	0.0402	0.1592	**0.1669**	0.1657
	MRR	0.1454	0.1231	0.4056	**0.4141**	0.4120
	P@5	0.0614	0.0485	0.1771	0.1771	**0.1783**
UvT	MAP	**0.2009**	0.1994	0.1155	0.1151	0.1158
	MRR	0.3551	**0.3571**	0.2298	0.2266	0.2281
	P@5	**0.1357**	0.1347	0.0850	0.0846	0.0841

Table 4.7: Expert finding results for the language modelling approach (LM), Experience (E), Relevance and Experience (RE), and Relevance, Experience and Area Coverage (REC)

We note that topic-centric approaches (E, RE, REC) outperform language modelling approaches on domain-specific datasets such as the CL, SW, and IR datasets. Our experimental results lead us to the conclusion that the more specialised a dataset is, the less reliable relevance-based assessment of expertise is. In the case of the Semantic Web dataset, which is the most focused dataset, using the relevance-based measure (RE)

even decreases performance compared to the expertise score (E). Language modelling approaches outperform topic-centric approaches on the UvT dataset alone, which is the most broad dataset among the four considered datasets. This is because expertise profiles have a larger degree of overlap when dealing with focused datasets that describe a narrow domain. For example, it is easier to distinguish between experts in history and mathematics using relevance-based methods, but more difficult to distinguish between two experts in Semantic Web that address similar topics in their publications.

Using a topical hierarchy by computing Area Coverage improves the results across all datasets except the IR dataset, in terms of MAP. In terms of P@5, the results are improved on all datasets except on the UvT dataset. These results confirm our hypothesis that topical hierarchies can inform expert finding. Furthermore, our algorithm for constructing topical hierarchies builds structures of sufficient quality to bring improvements in the expert finding task. In the following section we further investigate the overall quality of topical hierarchies through a comparison with hierarchical clustering.

4.4.5 Comparing topical hierarchies with hierarchical clustering

Because topical hierarchies make use of similarities between terms, they rely on the same type of information as hierarchical clustering approaches. In this section, we compare topical hierarchies with hierarchical clustering based on the improvements that the resulting hierarchies bring to the task of expert finding. An agglomerative approach with complete linkage clustering is used in our experiments [DE84], because this approach was shown to outperform other clustering methods when applied to hierarchy construction [CHS04]. To make sure that the two approaches are comparable, we use the same nodes and the same similarity metric in both cases.

The same list of expertise topics are used for clustering as for the topical hierarchies described in Table 4.6. We use the measure presented in Equation 4.4 to compute the similarity between expertise topics. The constructed dendrogram is converted in a hierarchy of expertise topics by labelling the resulting clusters. Each intermediate cluster in the hierarchy is labelled with the most frequent expertise topic. The agglomerative clustering algorithm merges two clusters at each step, which results in a large number of self-referring edges when the same label is identified for the merged cluster. For the purpose of our experiments, all such edges are ignored. Table 4.8 presents the results for the REC score described in Equation 4.15. The Area Coverage measure is

Dataset	Measure	HC	TC
CL	MAP	0.0333	**0.0340**
	MRR	0.2745	**0.2754**
	P@5	0.1344	**0.1347**
SW	MAP	0.0300	**0.0314**
	MRR	**0.2144**	0.2095
	P@5	0.0985	**0.0994**
IR	MAP	0.1581	**0.1657**
	MRR	0.4052	**0.4120**
	P@5	0.1643	**0.1783**
UvT	MAP	0.1130	**0.1158**
	MRR	0.2200	**0.2281**
	P@5	0.0838	**0.0841**

Table 4.8: Expert finding results using Area Coverage computed based on Hierarchical Clustering (HC) and Topical Hierarchies (TC)

computed using hierarchies through Hierarchical Clustering (method HC in the table) and using our algorithm for constructing Topical Hierarchies (called TC).

Computing Area Coverage using a topical hierarchy achieves better results for expert finding than using a hierarchical clustering algorithm. The improvements are stable across domains and for all the evaluation measures. In just one case, on the SW dataset, the hierarchical clustering method achieved better results than topical hierarchies, but only in terms of MRR. Furthermore, a manual analysis of the constructed hierarchies showed that topical hierarchies are more intuitive, because the pruning algorithm favours closely related terms. Hierarchies constructed through clustering are more difficult to understand, as they rely on similarities in a high-dimensional space, which are more difficult to trace.

4.5 Summary

In this chapter, we considered an approach for constructing topical hierarchies from text and their application for Expertise Mining, addressing the research question RQ2. After analysing some of the caveats of manually constructed taxonomies for our application context, we concluded that topical hierarchies are better suited for improving Expertise Mining. We introduced a graph-based algorithm for constructing topical hierarchies

using a domain corpus. At the core of this algorithm is a global generality measure that can be used together with co-occurrence based relatedness measures to identify SKOS *broader* relations between expertise topics. To this end, an existing graph-based algorithm for taxonomy construction was adapted for constructing topical hierarchies. Topical hierarchies are further used for expert profiling and expert finding. We proposed a method for constructing expert profiles based on automatically extracted expertise topics, discussing how a topical hierarchy can be used to improve the quality of expert profiles. Several methods for expert finding were introduced, including Area Coverage, a measure of expertise that relies on structured information provided by a topical hierarchy. The main contributions presented here are:

- a graph-based algorithm for constructing topical hierarchies using a global generality measure
- a novel measure of expertise based on topical hierarchies, that estimates the knowledge of an expert based on their knowledge of subtopics in a field
- an application-based evaluation of domain taxonomies in the context of Expertise Mining

The experimental results presented in this chapter show that our term extraction approach, that extracts terms of an intermediate level of specificity, is better suited for extracting expertise topics than previously used methods based on tf-idf. We showed that expert profiles can be constructed using automatically extracted expertise topics, without the need for controlled vocabularies. Our methods outperform state-of-the-art information retrieval approaches for expert profiling, when applied to domain-specific corpora. Concerning the use of topical hierarchies for expert finding, our method achieves better results compared to language-modelling approaches on all domain-specific datasets. Topical hierarchies proved to be more informative for Expertise Mining than automatically constructed taxonomies and than structures constructed through hierarchical clustering.

5

Application context

This section highlights several applications of the techniques proposed in this thesis, starting with Saffron, a system that provides insights in a research community or organization, in Section 5.1. Topical hierarchies can be used as a tool for analysing the main topics addressed by a research community and the relations between them. In Section 5.2, we present the results of a study of interdisciplinarity in the WebScience community. Section 5.3 shows how Expertise Mining can be used for semantic enrichment of experimental data, as produced by the Information Retrieval community. Next, we turn to the enterprise environment and we discuss the role of Expertise Mining within existing Content Management solutions.

5.1 Saffron. An Expert Search system for exploration and discovery of experts and expertise

The techniques proposed in this work are integrated in Saffron [1], a system that provides insights in a research community or organisation by analysing their main topics of investigation and the individuals associated with them. Currently, Saffron analyses mainly Computer Science areas, including Natural Language Processing, Information Retrieval, and Semantic Web, but there is an on going effort to extend this to other research domains. We start by giving an overview of the Saffron architecture and then we describe in more detail the main components of the architecture and the connections between them. This section is partially based on [MBSB10].

[1] Saffron:`http://saffron.deri.ie/`

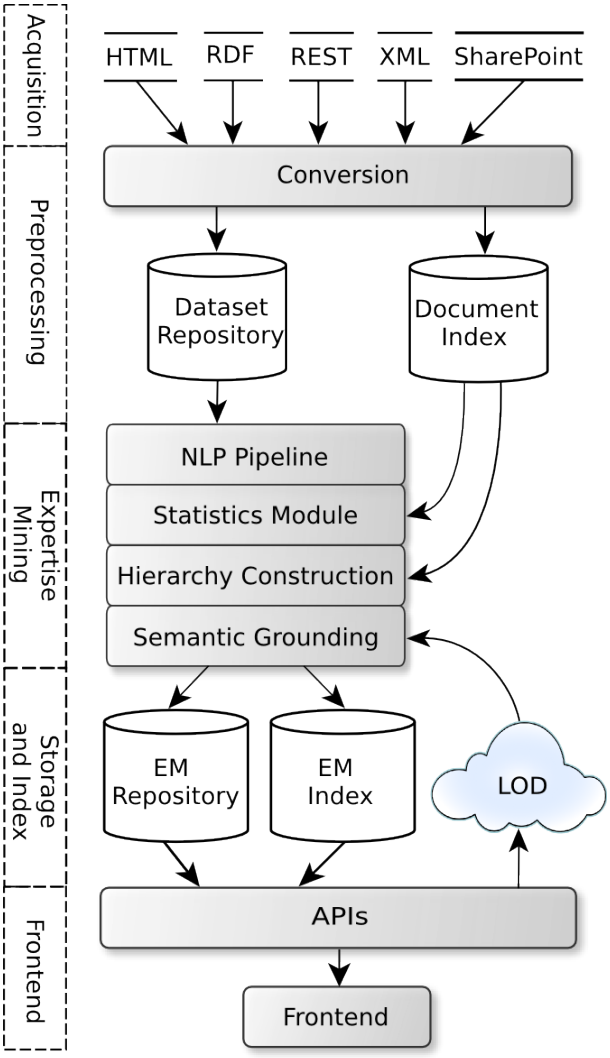

Figure 5.1: Overview of the Saffron infrastructure stack

5.1.1 System overview

Saffron is developed by DERI's Unit for Natural Language Processing (UNLP)[1], and was applied for several domains and application scenarios, including two organisations as well as several academic conferences and online communities. This system is a joint effort that involved several members of the UNLP. The system architecture, the backend and the NLP algorithms were developed as part of this thesis, while the frontend was designed and developed by Krystian Samp. The component for matching DBpedia

[1]UNLP:`http://www.deri.ie/nlp`

URIs was implemented in collaboration with Bianca Pereira.

As discussed in Chapter 3 and Chapter 4, Saffron provides support for several functional requirements including:

1. Automatic extraction of expertise topics
2. Automatic construction of expert profiles
3. Expert search functionality
4. Exploratory search of documents, expertise topics and experts
5. Support for finding similar researchers

Additionally, Saffron is designed to fulfil three main non-functional requirements:

1. Ability to cope with datasets acquired from different sources and represented in various formats
2. Minimal need for human interaction
3. Adaptive extraction algorithms that can cover a wide range of domains

These requirements guided the design of the Saffron infrastructure that can be seen in Figure 5.1. The first level of the infrastructure deals with the acquisition of documents and people associated with them for a given domain.The next layer is the preprocessing layer that converts and stores metadata and indexes documents. The Expertise Mining layer contains the core components of the system. Finally, results are stored, indexed and then made available for further integration with other knowledge management applications through APIs. The final layer in the infrastructure is the Saffron interface.

Data acquisition and data preprocessing

The first layer of the infrastructure is the acquisition of a suitable dataset about individuals and documents authored by them. As discussed in Section 4.3.1.2, relatively clean metadata about documents and associated people is already available for several domains. This metadata is most often represented in XML, but there is an increasing number of datasources available in RDF, through a public SPARQL endpoint. The CL

dataset makes use of the XML format to represent data about scientific publications, researchers and academic events, while the SW dataset represents the same types of information by making use of standardised vocabularies in RDF. RESTful Web Services are another way to provide access to expertise datasets, and this is the case of the DIRECT Infrastructure [1], which is discussed in more detail in Section5.3.

In the case of academic events such as conferences and workshops, it is often the case that information about publications and authors is not readily available, and has to be collected from dedicated HTML websites through web scraping. In the enterprise environment, most information about documents and organisation members is not public, and has to be accessed from content management tools, such as SharePoint. Depending on the dataset, people are identified using methods that are more or less ambiguous. Many of these datasets use personal names to identify the author of a document, therefore a name disambiguation and name consolidation component is required. Saffron uses a popular open source relational database, MySQL, as backend, and Lucene, an information retrieval library, is used for indexing full-content documents.

Expertise Mining components

This layer addresses the core tasks of Expertise Mining, including expertise topic extraction, topical hierarchy construction, expert profiling, and expert finding. Candidate terms are identified using a NLP pipeline based on the GATE natural language processing framework [CMBT02] and the ANNIE information extraction system, included in the standard GATE distribution. The NLP pipeline is depicted in more detail in Figure 5.2. We use several off-the-shelf components available in ANNIE for text tokenisation, sentence splitting and part-of-speech tagging. In the figure, components provided by GATE are represented in a lighter shade than the last two components, which are customized components for Expertise Mining.

A gazetteer, called DM Gazetteer[2] in the figure, annotates domain model words extracted from a domain-specific corpus using the method introduced in Section 3.3. Saffron identifies candidate terms using extraction patterns constructed starting from a domain model, as described in Section 3.4.4. Finally, candidate terms are annotated using a finite state transducer, called TE transducer[3]. Several extraction patterns are

[1]DIRECT: `http://direct.dei.unipd.it/`
[2]The acronym DM stands for Domain Model
[3]The acronym TE stands for Term Extraction

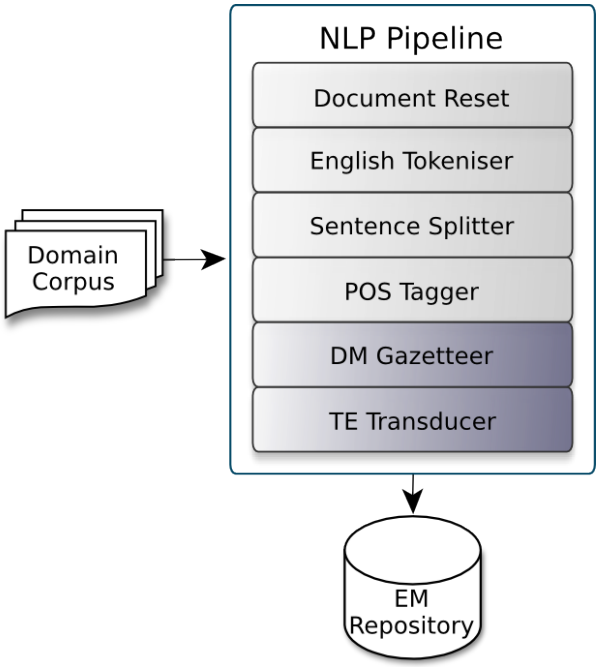

Figure 5.2: Overview of the NLP Pipeline for expertise topic extraction

used to select terms that contain a word from the domain model, or terms that are introduced by one, as discussed in Section 3.4.4. The candidates terms are stored in the Expertise Mining (EM) Repository for further analysis.

The Statistics Module is responsible for ranking and filtering candidate terms to identify expertise topics. Word occurrences, as well as relevance measures for expert finding and expert profiling, are computed using a Lucene index. The relations between expertise topics are identified by the Hierarchy Construction component, using the graph-based algorithm for constructing topical hierarchies introduced in Section 4.1.3. Again, Saffron relies on the Lucene index to measure co-occurrences between two expertise topics, making use of the span search functionality available in Lucene. This functionality allows us to perform proximity searches within a predefined window of words. The similarity between two experts is computed as well, to provide support for the case when the main expert is not available. We rely on the cosine similarity between expert profiles to identify people with a similar set of skills.

The final core Expertise Mining component is Semantic Grounding, that is responsible for identifying DBpedia URIs and descriptions of expertise topics. In this way Saffron provides an entry point in the Linked Open Data (LOD) cloud, as well as

descriptions for expertise topics that can be directly used by the end user.

Expertise storage and index

The next layer of the architecture, the Storage and Index layer prepares the data for high performance access by other applications. This is done either directly through APIs or by making the data available on the LOD cloud through a SPARQL endpoint. The EM Repository is a MySQL based solution for storing data. The Expertise Mining results are also indexed by the EM Index component, using Solr, a higly scalable enterprise search engine.

Frontend

Saffron supports users that have different information needs and varying levels of knowledge of a field to search for experts in a community or organisation. Several scenarios are considered, including novice members trying to establish connections, expert members looking for collaborators, as well as outsiders interested in an overview of the main areas of investigation or activity. The main functionalities of the system allow exploratory search and discovery of expertise topics, experts and expert profiles. Saffron provides keyphrase based search, enhanced with an autocomplete feature, for searching experts and expertise topics. Users that are not familiar with the domain are guided by a list of representative expertise topics that are listed on the start page. Table 5.1 shows the top ranked topics for three instances of Saffron, for Computational Linguistics, Semantic Web, and Information Retrieval. Users can select any of these expertise topics to find out more information about documents that mention them and associated experts.

Exploratory search is supported by a search interface which allows users to browse related expertise topics. Expertise topics are organised using our method for constructing a topical hierarchy presented in Section 4.1.3. Figure 5.3 shows the exploratory search interface for a topical hierarchy in the Semantic Web domain. Users can browse the Semantic Web domain starting from the most broad concept, the *Semantic Web* node, which can be seen at the center of the image. More specific concepts can be browsed by investigating subtrees, such as the *Linked Data* subtree, where each node and node label is linked to a corresponding topic page.

CL	SW	IR
training data	Semantic Web	information retrieval
word sense disambiguation	Web services	language models
speech recognition	search engine	search engine
target language	knowledge base	query expansion
syntactic structure	data set	relevance feedback
test data	web pages	retrieval system
relative clause	information retrieval	QA system
speech recognition system	social network	document retrieval
spoken language systems	data sources	answer type
source language	user interface	retrieval model
spoken language	web search	named page finding task
statistical machine translation	search results	Question Answering
translation system	web search engines	retrieval task
search engine	RDF data	web pages
semantic classes	natural language	web search engine
word classes	language model	relevant documents
translation model	Semantic Web technologies	search results
natural language processing	semantic web services	named entity
syntactic analysis	linked data	text retrieval
Text generation	RDF graph	expert search task

Table 5.1: Top ranked topics from the start page of the Saffron interface for Computational Linguistics (CL), Semantic Web (SW), and Information Retrieval (IR)

The Saffron interface is designed around three types of pages, based on the type of resource they describe: topical page, expert page, and document page. A topical page shows additional information about a topic such as occurrence trends across the time, a description of the topic, and related topics. Additional information includes a list of main experts that work on the topic, and the most relevant documents. An expert page presents the profile of that person, a list of similar experts that can be used if the expert cannot be contacted, and a list of documents authored by the expert. Figure 5.4 shows an extract from the expert page of a researcher in the Semantic Web community. Finally, a document page shows the authors of the document and several topics that describe the content of the document. Saffron maintains a web of connections between topics, experts and documents, enabling users to navigate from one type of resource to another.

The system has been applied inside organisations as well as at conferences. Further usability studies are required, but this is beyond the scope of this work. Future work will integrate topical hierarchies in the frontend, and use them as a tool for browsing documents and experts.

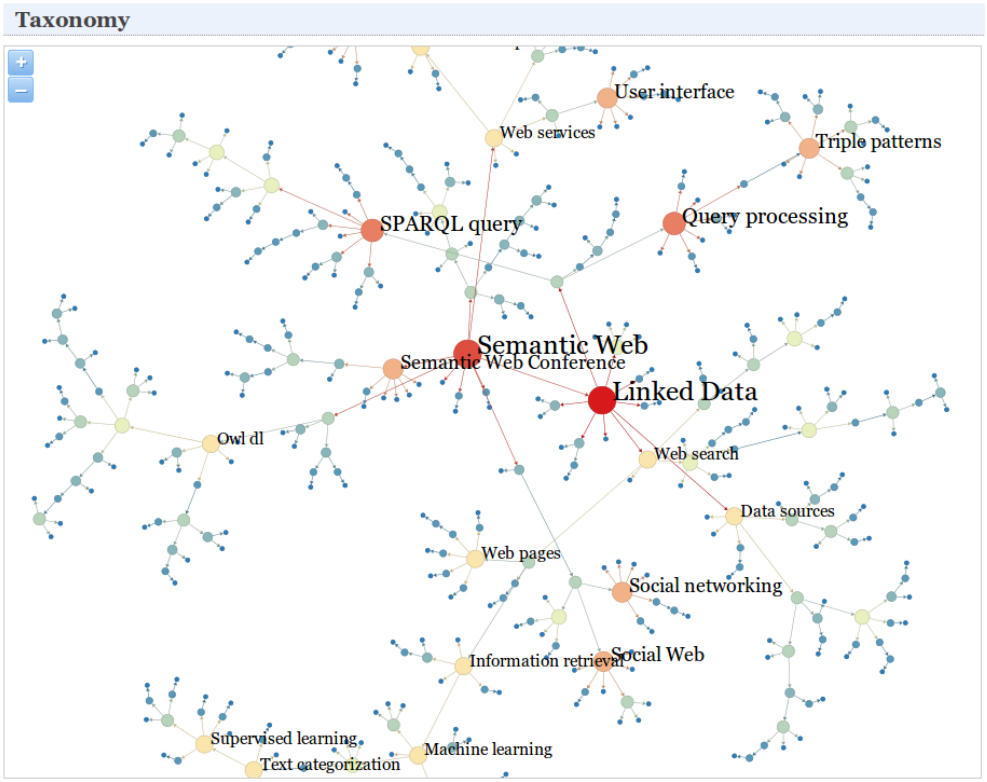

Figure 5.3: Exploratory search interface for the Semantic Web domain using an automatically constructed topical hierarchy

5.2 Analysing interdisciplinarity in the Web Science community

This section is partially based on [HBB13] and is the result of a collaboration with the IT Innovation Centre, of the University of Southampton [1]. Web Science documents were analysed using techniques presented in this thesis, while the interpretation was done in collaboration with the domain expert, Clare Hooper. She was also mainly responsible for organising the expert survey.

The Web Science community was created as part of a deliberate effort to bridge and formalise the social and technical aspects of the World Wide Web. Initial areas of interest covered research topics such as Social Networks, Collaboration, Understanding online communities, Analysing human interactions, Enhancing privacy and trust on the Web. As such, defining Web Science can be a difficult endeavour, as the community

[1] IT Innovation Centre: `http://it-innovation.soton.ac.uk/`

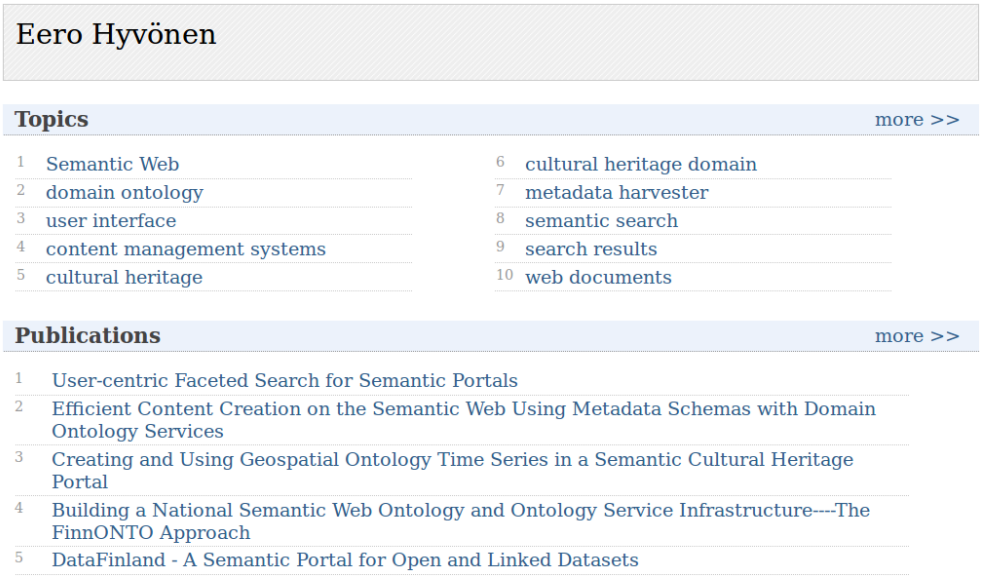

Figure 5.4: Saffron interface for an expert profile in the Semantic Web domain

aims to engage a rich variety of disciplines.

Tools to describe the field include the "Web Science butterfly" diagram, used early in the life of Web Science to convey the vision, but this diagram is a vision rather than an accurate depiction of the state of the field [HMK12]. Similar, the Web Science Subject Categorisation [Vaf11] only offers a vision and structure, not information on subjects prevalence within the community. Understanding the actual presence of different disciplines within Web Science offers several advantages. The community can better communicate what work is done under the WebSci flag; ground dialogue about Web Science diversity and disciplinary representation with data; identify under- and over-represented disciplines, and absent disciplines; identify problems that need addressing, and take action by seeking collaborations and communities that would remediate current weaknesses within Web Science.

This analysis is based on a corpus of papers from past Web Science conference proceedings, journal.webscience.org, and other sources. We used the term extraction approach presented in Chapter 3 and the algorithm for constructing topical hierarchies presented in Chapter 4. In this section, we present an analysis and discussion concerning:

1. Communities found within the corpus

2. Changes in the Web Science conference series over time
3. Changes in Web Science conference publications according to format
4. An expert survey regarding the mapping of terms to disciplines

We analysed communities within the corpus to gain insight into the relationship between different parts of the Web Science community. Our decision to analyse the conference series was a conscious one: the Web Science conference is in many ways the heart of Web Science, being the main annual gathering for Web Science researchers and practitioners. As such, the balance of disciplines at these conferences holds strong implications for Web Science as a whole: we therefore conducted an additional analysis examining these conferences.

We are interested in what differences, if any, can be discerned between poster and paper contributions. WebSci historically hosts extremely high quality poster sessions, but nonetheless, poster submissions are typically subject to somewhat less rigorous standards than paper submissions. Our assumption is that the distribution of disciplines represented by posters (as measured by terms) is broader than that of disciplines represented by papers. Finally, an expert survey was conducted. The task of relating terms to disciplines is a difficult one, and to avoid issues of subjectivity we pursue this path.

5.2.1 Background and related work

Previous work in bibliometrics ranges from co-citation analysis [CC99, WM98], to examination of multiple conference series [HGEF07], to geospatial visualisations of collaboration [NDH11]. Excepting a Web Science paper [HMK12], little prior work analyses the disciplinarity of conferences, although Web Science students at the University of Southampton produced an illustration of their own disciplines (based on supervisor disciplines) in March 2011. According to [BCB03], a bibliometric map can be constructed by analysing various types of items including journals, papers, authors, and descriptive terms. Our work is based on a basic assumption in bibliometric mapping [BCB03], which states that a research field can be described by a list of important keywords. While previous work made use of author assigned keyphrases and already built domain taxonomies [CMK98], we applied the automatic method described in Chapter 3 for term extraction as such resources are not readily available for our dataset.

Implicit relations between the extracted topical descriptors can be discovered and described through word co-occurrence analysis, a content analysis technique that was effectively applied to analyse interactions in different scientific fields [CCTB83, CMK98]. This technique was applied to analyse the interconnections between a main field, i.e., fuzzy logic theory, and other computing techniques [LHCHVH10], a setting that is similar to our analysis of the Web Science field. A more recent work on co-word analysis [WLLL12] outlines several limitations related to the use of keywords and proposes a method to integrate expert knowledge into the process. A main issue with this approach is that it requires a considerable amount of human intervention for the construction of domain specific thesauri. We alleviate this challenge by completely automating the process of identifying topical descriptors and by automatically constructing a domain taxonomy using the algorithm proposed in Section 4.1.3.

5.2.2 Method

We took the overall approach of automatically extracting terms from a large corpus of Web Science publications. We then analysed this data via graphing and visualisation. The communities present within the corpus of data and data concerning the Web Science conference series specifically are analysed. A survey is conducted to gain insight into perceptions of the linkage between different disciplines and key terms from the corpus. Co-authorship or co-citation data was not analysed, since our focus was on disciplines, which are better identified by term and not author: many authors particularly in the WebSci community have written within a diversity of disciplines. Nevertheless, co-citation data can be a useful basis for comparison with the language-based approach proposed in this work but we employ but we leave this direction of research for future work.

5.2.2.1 Data gathering

The focal point of our corpus was journal.webscience.org, which aims to collect and highlight Web Science literature and to provide a location for the research outputs of the Web Science community. Table 5.2 shows the data provided from this source, which spanned papers, posters, panels and keynotes (excluding 4 articles from a British Library Workshop on Ethics and the Web, as none of those articles could be processed by Saffron). Occasionally, articles on journal.webscience.org were listed without an

associated PDF. When this was the case, we instead included a text file containing the paragraph abstract on the webpage. Table 5.3 shows the additional publications that we included in the analysed Web Science corpus.

Source	Number of items (number usable)
Web Science Conference 2009	147 (133)
Web Science Conference 2010	109 (109)
Web Science Conference 2011	116 (116)
World Wide Web Conference 2001	1 (1)
Oxford Internet Institute Symposium: Dynamics of the Internet and Society	42 (42)
Web Evolution Workshop 2008	16 (16)
Royal Society Discussion Meeting	9 (9)
PLE (Personal Learning Environment) Conference 2011	75 (43)

Table 5.2: Publications (papers, posters, etc.) analysed from journal.webscience.org

Source	Number of items (number usable)
WebSci 2012	N/A
Key papers on Web Science	8 (8)

Table 5.3: Additional publications subject to analysis

The WebSci 2012 proceedings were processed as 3 PDFs representing posters, papers and panels; as each article was not processed separately it is hard to count how many were usable. These documents were not manually separated into their constituent parts due to time constraints, but this pre-processing step would improve the results of our analysis. 366 terms were extracted in total, showing a very noisy baseline: on average in the rest of the corpus we extracted 54 terms per article. Not all items were usable; the Data Processing section gives detail about this. We included WebSci 2012 publications (sourced from the WebSci 2012 webpage) and publications from Foundations & Trends in Web Science for the obvious reason: these publications are clearly a relevant part of the Web Science corpus. Note that the 7 files for Foundations & Trends constitute a large mass of data, with a total of 798 pages.

We included 6 key papers on Web Science that were surprisingly not already in our corpus, as they were not present on journal.webscience.org. Papers were chosen based on the recommended reading list of a forthcoming encyclopedia article on Web Science. These are: "Creating a Science of the Web" [BLHH+06], "Linked Data the Story so Far" [BHBL09], "Web Science: An Interdisciplinary Approach to Understanding the Web" [HSH+08], "The Semantic Web Revisited" [SHBL06], "Web Science Emerges" [SBL08], and "Web Science: a provocative invitation to computer science" [Shn07].

5.2.2.2 Data Processing: Saffron and Gephi

The number of files processed does not equal the number of publications, since some files contained multiple articles (e.g. the 3 PDF files for WebSci 2012). In total, we handled 552 files. We used the topic extraction component developed in Saffron, limiting the length of the candidate topics to 5 words. Candidate topics are further filtered using the web filter described in Section 3.4.1, discarding the candidates that have less than a minimum of 5 hits or that have more than a maximum 1 billion hits on the web. We used the ACM Subject Classification to build linguistic patterns for extracting terms in Computer Science as described in Section 3.4.4. Of 552 files, 491 were included in the analysis. 61 files were not processed, due to being of a format that Saffron, our processing tool, could not use. Saffron can only process plaintext and PDF files, meaning that Word documents and PowerPoint files were excluded.

The Saffron analysis yielded 5371 phrases that were identified as research term candidates, with an average of 54 candidates per document (although no term was extracted for 6 of the analysed documents [1]). Only the top 20% of terms are considered in our analysis. This threshold was necessary because the quality of terms influences the taxonomy of concepts: it is important to choose meaningful terms before analysing the relations between them. Another reason for limiting the number of analysed terms is because these terms were used for a manual analysis of the domain and a larger number would have been infeasible for the domain expert. The research terms are not manually curated, therefore they include incorrect terms such as *future research*, which is not a Web Science term. Like any other tool, term extraction and analysis has some

[1] At a closer analysis this was because the full text of the publication was not available for these documents

limitations, and the appearance of *future research* as an important term exemplifies the issue of incorrectly extracted terms.

We used the algorithm described in Section 4.1.3 to construct a topical hierarchy for the Web Science community from our corpus. Edges are added in the research terms graph for all the pairs that appear together in at least 3 documents. Again, Gephi is used to visualise a graph showing links between terms, allowing us to identify clusters of closely related terms. The Yifan Hu algorithm [Hu05] is used to layout the graph, and betweenness centrality to weight node importance. Betweenness centrality measures the fraction of shortest paths going through a node [Bar04]: a high value indicates that nodes play an important bridging role in a network. Finally, we ran the Louvain method [BGLL08] with resolution 12 to detect communities of closely related terms.

We examined the Web Science conference proceedings, by tracking four variables: keyword, year (2009, 2010, 2011), type (poster, paper), and count type (number of documents to contain keyword, overall keyword occurrence). The formatting of the WebSci 2012 proceedings meant that the data was largely unsuitable for processing using our methods; this analysis concerns the proceedings of WebSci 2009, WebSci 2010 and WebSci 2011.

5.2.3 Results

Figure 5.5 shows a visualisation of the WebSci topical hierarchy, where larger nodes and label fonts indicate terms with a higher betweenness centrality. This figure shows that *semantic web* is a central topic within WebSci, followed by other topics such as *search engine*, *information retrieval*, *learning network*, *social networking*, and *social science*. Overall, the algorithm for constructing topical hierarchies results in coherent subtrees, but semantic drift over longer paths is still an issue. Take for example the *social science* subtree shown on the upper-left part of the topical hierarchy presented in Figure 5.5. The node *resource description framework* is incorrectly classified under the *social science* subtree because it is considered a child of the node *mobile device*, which is not labelled in the provided image.

Table 5.4 lists terms with a high betweenness centrality:

The community detection algorithm found 9 communities, shown in Figure 5.5. Each community had its own subset of terms, which had been ranked by the topic

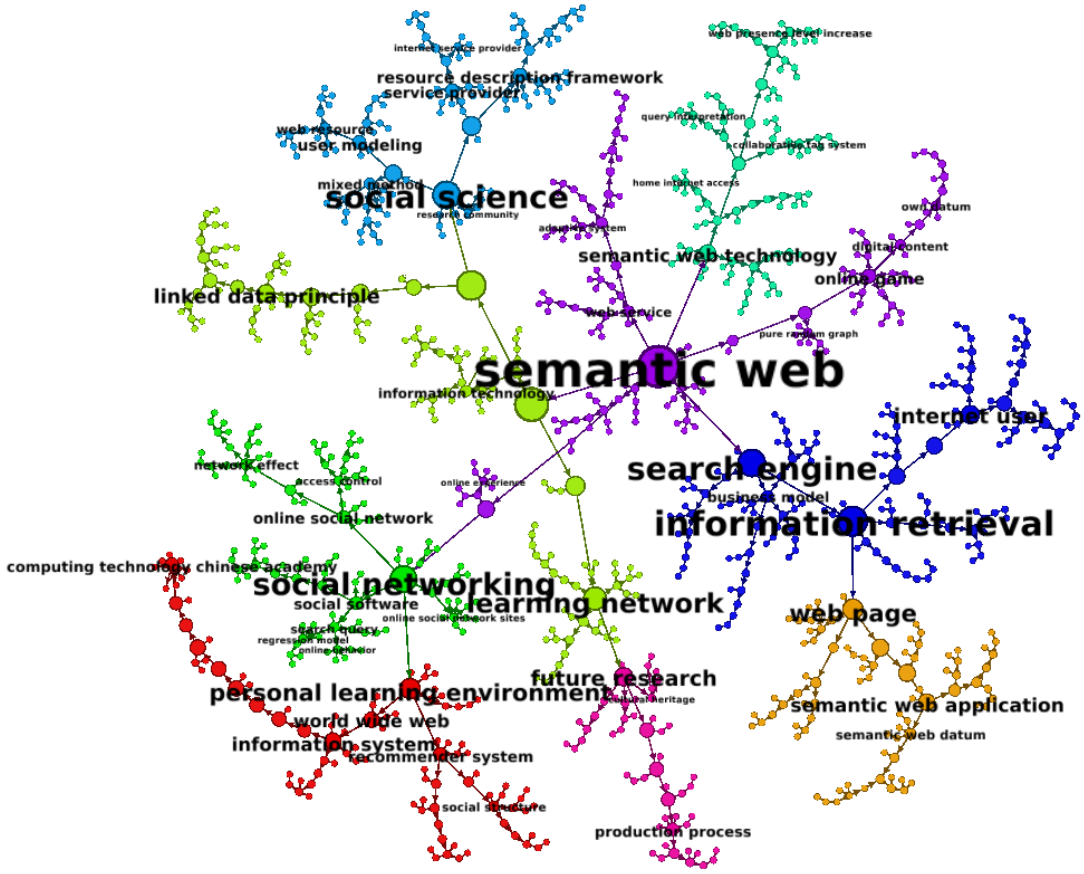

Figure 5.5: Topical hierarchy of Web Science - A visualisation of the extracted terms. Colours indicate communities

Betweenness centrality value	Term
758	semantic web
590	social media
504	information retrieval
495	social networking site
456	social science
454	search engine
434	social networking
360	learning network
304	web page
297	personal learning environment
282	social interaction
270	mobile device
260	future research
258	internet user
246	uniform resource identifier
235	web science research
235	user interface
235	web community
234	web application
231	linked data principle

Table 5.4: The 20 terms with highest betweenness centrality

extraction function described in Section 3.4.2 (rank range: 0 to 22). Table 5.5 details the 9 communities, including for each community its most highly ranked 5 terms and how many "hot" terms (terms with a score above 10) it contains.

To understand term diversity over time, we analysed the distribution of selected terms for each year. When comparing the documents published between 2009 and 2011, each year mentioned between 61% and 71% of the analysed terms, indicating that a relatively small number of topics vary from one year to another. These variations are nevertheless interesting, as they allow us to identify emerging trends by analysing "peak" terms. A "peak" term is defined as a term which occurs in five or more publications in a year compared to the previous year. It was considered that a "peak term" should displays this variation in both papers and posters.

Initial results considered as "peak terms" all terms that occurred more frequently in one year than in the previous year, but this condition was not sufficient to indicate an emerging trend. For example, the term "public sector" occurred in 3 papers in 2009 and 2010 and in 4 papers in 2011, but this difference in occurrence was only incidental. We discarded such results, keeping only peaks where the overall variation is greater than 5 papers when comparing subsequent years. The peak term "commercial advantage"

Root Node	# hot terms	% of graph	Top 5 terms
Search Engine	22	5	search result; open data; web search; natural language; information retrieval
Semantic Web	12	10	web science; data source; random graph; graph pattern; data set
Personal Learning Environment	9	10	world wide web; mobile web; information system; web archive; research information system
Social Science	9	15	web science research; p2p network; service provider; mobile device; user modeling
Social Media	8	8	social networking site; data mining; web site; information technology; social interaction Social
Web Page	6	13	web technology; user interface; web application; rdf data; semantic web application
Semantic Web Technology	3	12	social web; linked data; social bookmarking systems; information source; public sector information
Future Research	1	15	cultural heritage; learning network; production process; credibility evaluation; open source blogging platform
Social Networking	1	12	social software; search query; operating system; data management; analyzing social bookmarking systems

Table 5.5: Summary of the 9 WebSci communities

that appears in 2011 was discarded because this arose from a change in the wording of the copyright statement and not from the content of the papers.

This yielded 2 peaks in 2009, 10 peaks in 2010 and 1 peak in 2011 which can be used to identify emerging trends:

- 2009: machine learning; real world
- 2010: available online; information exchange; information retrieval; information sharing; natural language; RDF graph; real time; semantic web; share information; SPARQL query

- 2011: social media

According to this analysis, in 2009 the early Web Science community had a particular focus on broad topics such as machine learning, shifting its interest in 2010 to topics such as "information retrieval" and "semantic web". In more recent years, the peak term analysis shows an emerging trend around "social media" indicating a marked rise of interest around this area.

We tracked term diversity across papers and posters, examining how many of the considered WebSci terms are mentioned in each document type. Papers cover 70% of top terms, while posters are more diverse, covering 83% of terms. Peak terms average "height" (the difference between their minimum and maximum occurrence over time) is relevant. Average height in posters is 4.8, and in papers is 3.9.

5.2.4 Expert survey

The methodology for this expert survey was designed in collaboration with Clare Hooper, but she was mainly responsible for identifying Web Science experts and gathering their input because she is a domain expert. She approached experts in the field of Web Science and asked them to map disciplines to terms. Experts were recruited by email with the goal of targeting experts from a wide range of disciplines. These experts are provided with top 20 extracted terms (ranked by betweenness centrality). One incorrectly extracted term ("future research") was removed from this list. The experts are asked to map those terms to disciplines provided in a list, but they are not required to keep to that list. This controlled list of disciplines is made up of every discipline mentioned in the past 5 CFPs for the Web Science conference (2009-2013): communication; computer and information sciences; criminology; design; digital humanities; economics; geography; language and communication; law; linguistics; management; political science; sociology; philosophy; psychology.

13 experts responded that came almost entirely from academia (12/13); the industrial responder is Chief Scientist at a relevant company. The academics consisted of 2 professors, 4 lecturers, 3 post doctoral researchers and 3 PhD students. 12 respondents had worked in WebSci (1 described as having done "related work"), and 11 had published at the WebSci conference. 4 respondents described their main discipline as

WebSci, with the other main disciplines described as Archaeology (2), Computer Science/Software Development (3), Digital Humanities (1), Health Sciences (1), Law (1), NLP (1). All respondents reported working in additional fields, which is unsurprising given that we specifically targeted Web Science researchers and practitioners.

Table 5.6 summarises the results, showing the number of disciplines suggested in relation to each term, and enumerating disciplines that were mentioned in relation to each term by at least three experts. A deep analysis of the survey results is beyond the scope of this paper, but the results clearly show a high variance in the number of disciplines to be associated with a term.

Term	#Associated disciplines	Disciplines named by at least 3 experts
linked data principle	1	Computer and Information Sciences (CompSci)
information retrieval	2	CompSci
uniform resource identifier	4	CompSci
web science research	4	Any/all; CompSci; Web Science
semantic web	7	CompSci
user interface	7	CompSci; Design
search engine	8	CompSci
web application	8	CompSci
web page	8	Any/all; CompSci
internet user	9	CompSci; Psychology; Sociology
social science	9	Sociology
personal learning environment	10	CompSci; Education
web community	10	CompSci; Psychology; Sociology
learning network	11	CompSci; Pedagogy
mobile device	11	Any/all; CompSci; (Industrial) design
social networking	11	Communication; CompSci; Sociology
social networking site	11	Communication; CompSci
social interaction	12	Sociology; Psychology
social media	12	Communication; CompSci; Network Science; Sociology

Table 5.6: Disciplines associated by domain experts to extracted terms

Information retrieval and *uniform resource locator* are prime examples of terms where the majority of respondents immediately associated the term with Computer Science and nothing else. By contrast, some terms yielded wildly diverse discipline lists: examples included all terms to do with social interaction, media and networking (each yielding over 10 disciplines), and also *learning network* and *mobile device*. Many of the suggested disciplines were only suggested by one or two separate experts, and so are not enumerated in 5.6: nonetheless, it can be seen that relating a discipline to these terms is controversial. While *information retrieval* is a generic name for a set of

techniques from computer science, *social media* can be the object of study of multiple disciplines: it is probably this key difference that explains why some terms had more diverse connections to other disciplines.

When examining disciplines named by at least three experts, we see a preponderance of responses naming Computer Science and Sociology, with a majority of topics being associated with Computer Science. Given Web Science's traditional foundation upon these two terms, this is perhaps no big surprise. Our approach highlights topics actually discussed by the community, showing a discrepancy between the stated mission of the Web Science community, which is to bring together research from diverse disciplines, and ongoing research which is still driven by the Computer Science community.

Other strongly present disciplines were Psychology and Communication, which were both named by at least three experts in relation to three separate terms. There is no relationship between how highly ranked terms were and how controversial they were when experts related disciplines to them: the 5 top ranked terms (*semantic web, social media, information retrieval, social networking site, social science*) had 7, 12, 2, 11 and 9 disciplines associated to them respectively.

It is perhaps disheartening to see that at least three experts associated Computer Science with the term *web science research*, but that the same was not the case for any other discipline (except Web Science itself!) - even Sociology, which was otherwise frequently named by the experts. Unsolicited comments from the experts are informative. Some experts criticised the lists ("the [term] list seems to be very much slanted towards technology and away from anything like law, economics, sociology"; "you need to add all the [humanities] disciplines if you're going to add philosophy [...] And what about art, design, media studies, gender studies?"; "There are some startling absences, e.g. business studies, art, culture [...] and education"). This has implications for a) the meaning of the top-rated terms (do they imply that WebSci is in fact only the study of technology?) and b) the decisions made regarding what disciplines to enumerate in WebSci conference CFPs.

There is a larger debate unfolding here: can terms - whether extracted via NLP or by hand - ever reflect or represent particular disciplines? Some terms, such as information retrieval, mapped clearly to a single discipline, but many did not, occurring across many disciplines: terms such as *social media*, *social networking*, and *mobile device* might occur in any field of study, in very different ways. Indeed, some terms will mean

wholly different things according to context (consider *social networking* in Sociology, and in Computer Science). When a term is in situ in a publication, it has much contextual information (arguments made, methods used, authors' backgrounds) that the topic extraction process strips out. Solutions include: displaying clusters of related terms; identifying related terms through co-occurrence; using related terms from the taxonomy; semantic grounding via definitions; showing the context of the term (i.e., the paragraph surrounding the term).

Although we used most of the above techniques, the experts did not have access to this information: we kept the survey short to elicit more responses. Thus, the survey showed the terms extracted but not the taxonomy: for example *web page* is a term that was assigned to any discipline by some experts. The taxonomy reveals that it is mainly used as a subtopic of *information retrieval*, meaning we can conclude that it comes from computer science.

5.2.5 Discussion

The use of several different methods to analyse Web Science data allows us to corroborate our results. For example, the partition algorithm identified 9 communities, which on inspection mapped to 4 key communities: information retrieval; personalised learning/elearning; semantic web; social networking. We can see these communities at the Web Science conferences: WebSci10 clearly included the semantic web and information retrieval communities (its peak terms included those very phrases), while WebSci11 presumably had stronger presence from the social networking community, with its peak term of "social media". We noticed a dearth of peak terms in the WebSci conference series related to the Personalised Learning Environment community in the lower left corner of the Web Science topical hierarchy. This further suggests that that community arose from the PLE conference included in the corpus at journal.webscience.org.

The expert survey presented here included terms related to the 4 communities. Of these, information retrieval was uncontroversial and mapped straight to Computer Science. Computer Science was the only discipline named by more than 3 experts in relation to the term "semantic web", but the term did elicit a total of 7 named disciplines. The terms relating to the remaining two communities, personalised learning environment and social networking, were both controversial, eliciting 10 and 11 named disciplines apiece. We suggest that it is the controversial terms, the ones that elicit

many named disciplines, which are the most important to Web Science: although technologies like linked data and information retrieval techniques are clearly necessary to Web Science, they are perhaps tools of Web Science, rather than its heart. By contrast, the controversial terms such as social networking, web community and social interaction are the terms that truly reflect the ethos of Web Science, the goals of understanding and engineering the webs impact on our society and the impact of societies upon the web.

5.3 Semantic enrichment for Information Retrieval experimental data

Information Retrieval (IR) experimental evaluation is a process based on a conceptual framework relying on the Cranfield methodology [Cle67]. This process is carried out in the context of large-scale international evaluation campaigns which aim to guarantee an impartial comparison between different systems, reproducibility of the experiments, and re-use of the data adopted and produced during the campaigns. The Cranfield methodology makes use of shared experimental collections in order to create comparable experiments and evaluate the performances of different IR systems.

Evaluation campaigns contribute considerably to the advancement of information retrieval systems by providing an infrastructure and resources for researchers to test, tune, evaluate and compare new approaches. Large amounts of data are regularly produced in this process. This includes structured data about test collections, evaluation activities, evaluation measures, and visual analytics. At the same time, descriptions of shared tasks and reports about experimental results are made available as free text. Currently, evaluation tasks span application areas as diverse as cultural heritage, eHealth, intellectual property, image retrieval, XML retrieval, plagiarism detection, question answering, and entity recognition. Each new campaign sees the introduction of new application areas, while others become outdated.

Describing and annotating scientific datasets is essential for their interpretation, sharing, and reuse [Bow12]. A main goal of an evaluation campaign is to create reusable test collections for benchmarking information access systems, but without concerted effort to semantically annotate and interlink these datasets, impact is limited to the participants of a shared task and to a limited timeframe. The techniques proposed in

this thesis can be used to automatically annotate various resources of an evaluation campaign. In a first stage, documents are processed and expertise topics are extracted, which are further used to construct expertise profiles for authors. These profiles are used to inform users about the experts for a given task or application domain. In this way, a network of semantic relations between the main resources of an evaluation infrastructure can be created, that enables users to quickly locate relevant data for their experiments.

5.4 Expertise mining for Enterprise Content Management

This section is partially based on [BKBP12] and was the result of a collaboration with Storm Technology[1], Galway, Ireland. Enterprise documents were analysed using the techniques presented in this thesis, but we had no direct access to the documents themselves due to security reasons. These documents were processed by Sabrina Kirrane using Saffron, but we had direct access to the results of the extraction.

The Enterprise Content Management (ECM) concept has been slowly evolving over the past two decades. Today, this all encompassing term is commonly used to refer to enterprise document management, content management, records management, collaboration, portal technologies, workflow and search [Dil11]. The business benefits attributed to the deployment of an ECM system are compliance, efficiency, customer service and lower costs [Sco11], [Dil11]. Nevertheless, the aforementioned benefits are highly dependent on the effectiveness of both the taxonomy and metadata used to describe the enterprise content [MH06], [Sco11]. Significant groundwork goes into the initial enterprise content analysis and platform configuration, therefore organisations regularly seek help carrying out this activity from specialist consultants known as Information Architects. Because Information Architects are not necessarily experts in the organisation's business domain it is important that they identify key individuals to be involved in analysis interviews that will result in the construction of the ECM taxonomy.

Recent advances in expert finding [BdRA06], [MO06], [SRH08], automatic term recognition and automatic taxonomy construction [NVF11], [KH10], as well as the in-

[1]Storm Technology: `http://www.storm.ie/Pages/home.aspx`

creasing richness of structured data openly available on the Linked Data cloud [1] address some of these challenges. But these approaches still suffer from various limitations such as exact matches of expertise topics, lack of expert profiles needed in the selection process and generality of extracted terms and taxonomical relations. The work presented in this section addresses automatic techniques to support the initial content analysis, taxonomy generation and the selection of experts who can validate the knowledge obtained from enterprise repositories. Expertise mining complements the traditional task of expert finding with expertise topic extraction and expert profiling to automatically link expertise, Information Workers and documents. The ECM taxonomy can be used both for organising the enterprise contents and for improving expert finding. We integrate data driven approaches for expert search with knowledge resources such as domain taxonomies and Linked Data expertise traces.

Experimental setup and results

The enterprise dataset under analysis consists of 11,319 files subdivided into 3,319 folders and is composed of both structured and unstructured documents. The corpus contains a combination of word documents, excel spreadsheets, power point presentations, pdf documents and plain text files that span several years from 2003 to 2009 inclusive. Excel documents do not introduce noise because we rely on lexical patterns for selecting candidate terms for term extraction. In our first experiment we evaluate the extraction of expertise topics through a user study, then we present the results of expertise topics grounding on DBpedia.

A user study with three participants who are domain experts is set up to evaluate the topic extraction method. The objective of this experiment is to investigate whether our term extraction approach achieves acceptable results to be included as part of a Content Management system in a real enterprise use case. Due to limited availability and strict time constraints, domain experts are asked to evaluate a reduced list of 100 topics from top, middle and bottom of the ranked list of expertise topics. We expect a high number of correct topics at the top of the list and a high number of incorrect topics at the bottom of the list. Domain experts are given a list of shuffled topics selected in

[1] Linked Data cloud: a freely available collection of structured data from different domains that provides us with a gateway to additional information about expertise topics and people `http://richard.cyganiak.de/2007/10/lod/`

the following way: 34 topics from top ranked topics, 33 from middle ranked topics and 33 from bottom ranked topics.

The three judges are instructed to rate the expertise topics for the given domain by selecting one of three possible options for each topic: "good", "bad" and "undecided". Table 5.7 gives an overview of the user study results where all three annotators were in agreement. We only present topics considered correct and incorrect including the position where the topics appear in the ranked list (i.e. Top, Middle, Bottom). Almost 80% of the topics that are ranked high by our system are confirmed to be correct by all the three judges but only 30% of the topics ranked low are confirmed bad.

Answer	Top	Middle	Bottom
Good	**0.79**	0.18	0.09
Bad	0	0.06	**0.30**

Table 5.7: Perfect agreement results for expertise topic extraction

The kappa statistic is used to measure the agreement between the three judges. Only the expertise topics that are judged as good or bad (62 out of 100) are considered, ignoring all topics that are ranked as "undecided" by at least one participant. Kappa is in the moderate agreement range (0.61), but much higher agreement (perfect agreement for almost 80%) can be observed for the expertise topics at the top of the ranked list. The agreement rate is much lower for the topics ranked lower, indicating that the human judges have more difficulties distinguishing the quality of lower ranked topics.

5.5 Summary

We presented Saffron, an expert search system that allows users to browse through representative topics of a domain and identify associated experts. We explained the functional and non-functional requirements that guide the design of the system and we presented an overview of its underlying architecture. Next, several applications that include analysing the interdisciplinarity of a research community using topical hierarchies and the semantic enrichment of experimental data showed the importance of Expertise Mining in the academic domain. With respect to the applications of Saffron in the enterprise environment, we discussed its integration with a Content Management system. Finally, we considered Expertise Mining from an emerging type of documents

characterised by small incremental contributions, as produced by users of knowledge-curation platforms.

Further work needs to be done to combine knowledge extracted from text with background knowledge available from the LOD cloud. This would improve the quality of extracted expertise topics and the quality of topical hierarchies.

6

Conclusions

In this thesis, we investigated the automatic construction of topical hierarchies from text, focusing on an application-driven evaluation in the context of Expertise Mining. This work is founded on several findings in cognitive psychology, that link the level of expertise of a person with knowledge representation abilities. These studies also found a connection between the expertise topics known by an expert and their position in a domain hierarchy. We mainly relied on documents written by a person as an approximation of their expertise, making use of content analysis techniques to extract and structure expertise descriptions, which are further used for profiling and ranking experts. More specifically, we considered four different tasks which are relevant to expert search: expertise topic extraction, topical hierarchy construction, expert profiling, and expert finding.

6.1 Summary of the thesis

Throughout this thesis we advocate for using knowledge structures extracted from text for exploring and discovering experts and expertise, and for deriving semantically motivated measures of expertise. We introduced a method for constructing domain models from domain-specific corpora, and we proposed several ways to improve term extraction using domain models. Our quest for methods that balance domain specificity with the generality of a term within a domain has lead us to find a global generality measure, with applications in the construction of topical hierarchies.

In this thesis, we tackled the following specific issues related to Expertise Mining from text:

Expertise topic extraction using automatically constructed domain models (Chapter 3). A domain model provides an elegant solution for the extraction of terms of an intermediate level of specificity, specific to a domain, but general enough to be used for summarising expertise areas. This approach has applications beyond Expertise Mining, for keyphrase extraction and index term extraction (Section 3.7.2.3). We proposed a method to automatically extract a domain model from domain corpora and we introduced two approaches for term extraction that make use of a domain model. We thoroughly evaluated these approaches across various domains and we showed that domain modelling achieves better results for expertise topic extraction than the state-of-the-art in the field (Section 4.4).

Topical hierarchy construction through a global generality measure (Section 4.1.2) and a graph-based algorithm (Section 4.1.3). Topical hierarchies have several advantages compared to *is-a* taxonomies, when applied to Expertise Mining. Our method for constructing topical hierarchies relies on domain corpora alone, without the need for external corpora. In this work we provided a qualitative comparison with taxonomies constructed using previous approaches, discussing their limitations in the context of Expertise Mining.

Expert profiling using automatically extracted expertise topics (Section 4.2.1). We proposed a method to construct expert profiles using automatically extracted expertise topics. This approach avoids the need for controlled vocabularies or for manual identification of knowledge areas. Through a series of experiments, we showed that this approach is performing best when applied to specialised domains (Section 4.4.3), outperforming state-of-the-art methods for expert profiling.

Expert finding by analysing the coverage of subordinated topics in a topical hierarchy (Section 4.2.2). A novel measure of expertise was proposed, which measures in-depth knowledge of a field based on a topical hierarchy. Experimental results support the conclusion that topical hierarchies are more informative for Expertise Mining than similar structures constructed through hierarchical clustering (Section 4.4.4).

Finally, we presented a system architecture that integrates the techniques proposed in this thesis (Section 5.1). This system proved to be a valuable tool for analysing

research communities (Section 5.2), for enriching experimental data (Section 5.3), and for Enterprise Content Management (Section 5.4). Additionally, we proposed an exploratory search interface based on an automatically-constructed topical hierarchy that allows novice users to discover experts and expertise (Section 5.1.1). Together, these contributions demonstrate the feasibility of extracting high-coverage, informative knowledge structures from text, and applying them to improve real-world applications such as Expertise Mining.

6.2 Discussion

In this section we present each of our contributions in more detail, discussing their limitations and how they can be addressed.

6.2.1 Expertise topic extraction

Motivated by the lack of controlled vocabularies for knowledge areas and their low coverage for most domains, we proposed adapting existing term extraction techniques to automatically extract expertise topics from domain corpora. We made the observation that expertise topics are a subset of the terms in a domain and that the specificity level of a term is a main characteristic that allows us to distinguish expertise topics. The drawbacks of contrastive approaches for term extraction, which generally favour more specialised terms within a domain, can be overcome by using a domain model. The main challenge when constructing a domain model is to select generic words that have a high distribution in a domain, but which are specific enough to be representative for that domain.

We proposed a method to automatically construct a domain model using several heuristics to select candidate words from a domain corpus and a measure for domain coherence used to rank these candidates. An information-theoretic approach was used to measure the similarity between terms, but there are many other approaches that can be considered here. This work can be extended with a more exhaustive comparison of similarity measures for measuring coherence. Another limitation of our approach is that we did not consider using manually extracted terms from background knowledge, because we were interested in a completely automated approach.

Next, we investigated methods for applying a domain model to guide term extraction and we proposed a pattern-based approach to select candidate terms. A limitation of this approach is that recall is decreased because we consider a smaller list of candidates. To address this limitation we proposed a method for ranking candidates using domain coherence. A disadvantage of the ranking approach is increased computation time, therefore the choice should be done depending on specific requirements of the domain. The size of the domain model can be adjusted to decrease computation time, making a compromise between speed and precision. Another limitation of this approach is that each document has to be processed twice, by two separate NLP pipelines, one for domain modelling, and another one for term extraction. Our assumption is that the domain model can be constructed by computing only a subset of the available documents but further experiments are required to compute the optimal size of this subset. Multiword terms were mainly considered, because single-word terms are more challenging to extract and also more ambiguous.

Furthermore, we made a first step towards grounding expertise topics on the LOD cloud. The disambiguation of expertise topics was not addressed, but an approach using the domain model to select domain-relevant senses can be envisioned. Also, we considered DBpedia alone as a source of background knowledge, but a much larger number of domain-specific datasets are available. The selection of relevant knowledge resources for a domain is a challenging problem by itself, but again a domain model could provide a feasible solution.

6.2.2 Topical hierarchy construction

In Chapter 4 we analysed existing taxonomies based on their application to Expertise Mining and we identified a need for automatically constructed topical hierarchies. We proposed a method for constructing a topical hierarchy from text that has several advantages compared to previous approaches for constructing lexical taxonomies:

- the resulting hierarchies are informative and have a high coverage of domain-specific concepts
- a much larger number of technical terms can be connected in the hierarchy
- automatic identification of the hierarchy root

- no human input, in the form of seed terms or upper level concepts, is required

This method is based on a global generality measure that takes in consideration the co-occurrence of a term with all the other terms in a domain. A graph-pruning algorithm was used to derive a tree-like structure from the dense, noisy graph constructed using the semantic relatedness of terms. Then, a novel expertise measure was proposed based on a topical hierarchy. Topical hierarchies were mainly evaluated through their contribution in the task of expert finding, by comparing them with hierarchies constructed using hierarchical clustering. Topical hierarchies are more coherent that hierarchical clusters, consistently achieving the best results when applied to expert finding.

Topical hierarchies also proved to be intuitive enough to be directly used by humans for analysing large corpora, as we could see in Section 5.2. But there is still place for improvement, for example a manual analysis showed that topics tend to drift from the main knowledge area on longer paths. This limitation could be addressed by clustering topics as a pre-processing step. This step is also important when constructing topical hierarchies for more heterogeneous domains, such as the documents in a university, which cover a wide range of topics, with a small overlap. We did not limit the depth or the width of the hierarchy based on manually provided parameters, but this could be desirable when constructing hierarchies for human users. There are cognitive limitations which make less complex structures more usable and easier to comprehend if they are designed directly for the end user.

Knowledge structures can be evaluated through comparison with existing gold standards or through manual evaluation with domain experts. Further studies are needed to prove the applicability of topical hierarchies beyond Expertise Mining.

6.2.3 Expert profiling

Expert profiles provide useful information when searching for an expert, as they put the expertise of a person in the context of other expertise topics. A main motivation of our work was to completely automate the construction of expert profiles, by eliminating the need for manually-identified expertise topics. This was achieved by constructing expert profiles using automatically-extracted expertise topics. Expertise topics extracted from documents authored by a person, were ranked based on their overall quality and based

on their relevance to that person. Our approach for expert profiling was evaluated on a dataset about workshop committee members from different fields of Computer Science. It was assumed that each committee member is an expert on all the expertise topics mentioned in the call-for-papers of a workshop.

This evaluation dataset has its limitations, especially for workshops that intend to bring together researchers from different research areas. Another limitation is that the resulting profiles are much larger than what would normally be displayed in an expert search system. A main challenge remains to construct profiles that are as complete as possible while being still concise. A possible solution to this problem can be to make use of a topical hierarchy to select broad topics that can summarise expertise, while filtering more specialised topics. Additionally, further work is required to represent expert profiles in machine readable formats which are compatible with existing standards for competency.

6.2.4 Expert finding

Expert finding has many applications, in industry as well as in research communities. To tackle this problem, we relied mainly on documents associated with a person, investigating several content-based methods for expertise. We relied on the number of times a person mentions an expertise topic to measure relevance, and the number of different documents written about a topic to measure experience. Additionally, we proposed a measure based on a topical hierarchy, which takes into consideration how well a person knows specialised topics in an area. The method used to combine these scoring functions is rather simplistic, leaving place for improvement either by considering a linear combination or a more sophisticated learning-to-rank approach.

We showed that these measures outperform state-of-the-art methods for expert finding in terms of both precision and recall, when applied to domain-specific datasets. Nevertheless, the overall results are still low enough to require additional sources of information beyond textual documents. A solution to this problem is to combine expertise extracted from the content written by a person with more structured information about the number of citations of a document or about education or training, which can be extracted from CVs.

Another limitation of this work is that our expert ranking approach is static, and does not take into consideration the profile of the person that is doing the search. The

final goal of an expert search system is to enable collaboration, therefore it is just as important to map the expertise of the user that is doing the search as it is to identify experts. Therefore, expert finding should be seen as a recommendation problem, and the results should be personalised.

6.3 Directions for Future Research

The results presented in this work open several avenues for future research. In the following, we present a list of future directions of research related to the construction of topical hierarchies and their application in Expertise Mining:

Background knowledge The LOD cloud is a rich source of manually curated hierarchical relations, that are available for an increasing number of domains. Existing hierarchical relations from background knowledge should inform the construction of topical hierarchies whenever possible. The challenge is to identify which concepts and relations from background knowledge are relevant for a given domain and task, and to disambiguate concepts based on them.

Integration of pattern-based and definition-based relations A hybrid approach that combines high-quality taxonomical relations extracted using patterns and/or from definitions, with high-coverage, data-driven relations extracted using the techniques proposed in this work could benefit from the advantages of these different approaches.

Application-based evaluation of concept hierarchies In this thesis we evaluated concept hierarchies based on the benefits they bring in Expertise Mining. Other applications where evaluation datasets are readily available such as document browsing, document clustering, and question answering can be considered as well for an automatic evaluation of concept hierarchies. This would enable measurable progress toward automatically constructed hierarchies that are beneficial for real-world applications.

Heterogeneous datasets This work does not attempt to solve the problem of constructing topical hierarchies for heterogeneous datasets that cover multiple loosely-related domains. A possible solution can come from the area of document clustering, but this direction has to be explored further.

Exploratory search interface In this thesis, we discussed a first prototype for an exploratory search interface based on topical hierarchies. This prototype was applied for browsing hierarchically-organised expertise topics, but the interface could be extended to display expert profiles as well. Further usability studies are required to evaluate this novel search interface.

7

Appendix

Appendix 1. Domain modelling survey

The participants of the domain modelling survey received the following instructions by email, followed by a list of 40 candidate words:

> *Select from the list provided below words that are likely to be used in any scientific publication in the Computer Science field by marking them with an X. You should reject words that are only used in a subfield of the domain as well as words that are too specific to be used often.*

Table 7.1 summarises the answers given by survey participants when asked to select words that are likely to be used in Computer Science publications. The columns represent the answers given by each participant, which are anonymised using a numeric identifier from 1 to 27. Each row contains the answer given by a participant for the candidate words listed in the first column. Each correct word for a domain model was marked by 1, while incorrect words were marked with 0.

Appendix 2. Automatically extracted domain model

In this appendix we include a selected list of words from a domain model constructed using the method presented in Section 3.3 for the Computer Science domain. Table 7.2 presents the top 50 ranked words extracted from the *Krapivin* corpus, described in Section 3.6.2.2. The DM Score presented in this table is the score described in Equation 3.3.

7. APPENDIX

Candidate	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27
agent	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0	1	0	0
algorithm	1	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0
analysis	1	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
approximation	0	1	1	0	1	0	1	0	1	1	0	0	1	0	1	1	0	1	0	1	0	0	0	0	0	0	0
circuit-switching	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0
concept	0	1	0	1	1	0	1	0	1	1	1	1	1	1	1	0	0	1	0	0	1	1	1	0	1	1	1
connection	0	0	0	1	0	0	0	1	0	1	1	0	0	0	1	0	0	0	0	0	1	0	1	0	0	0	0
data	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
debuggers	0	0	0	1	0	0	1	1	0	0	0	0	0	0	1	0	0	0	1	0	1	0	0	0	0	0	0
definition	0	1	1	0	1	1	1	0	1	1	1	1	1	1	1	0	1	1	0	0	1	1	1	1	0	1	1
depth	0	0	0	0	0	0	0	0	0	1	0	0	1	0	1	0	0	1	0	0	1	1	1	0	0	0	0
directory	0	0	0	1	0	0	0	1	0	0	0	0	0	0	1	1	0	0	0	0	1	0	0	0	0	0	0
execution	1	1	0	1	0	0	1	0	0	0	1	0	1	1	1	1	0	1	0	0	1	1	0	0	0	0	0
feature	0	0	0	0	1	1	1	0	1	1	1	1	1	1	0	1	1	1	0	0	1	1	1	1	0	1	0
framebuffer	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
frames	0	0	0	0	0	0	0	1	1	0	0	0	0	0	0	0	0	1	1	0	1	0	1	0	0	0	0
generation	1	1	0	0	0	0	1	0	1	1	0	0	1	0	0	0	0	1	0	0	1	1	0	0	0	0	0
grayscale	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
integration	1	1	1	0	0	0	1	0	1	1	1	1	1	0	1	0	1	1	0	1	1	1	1	0	1	0	0
key	0	0	0	0	0	0	1	1	1	1	0	0	0	0	1	0	0	1	0	0	0	0	1	0	0	1	0
language	0	1	0	1	0	0	1	1	0	1	0	1	1	1	1	1	0	1	0	1	1	1	1	0	1	0	0
machine	1	1	0	1	0	1	1	1	0	1	0	0	1	1	0	1	0	1	0	0	1	1	1	1	1	1	1
mathematics	0	0	0	0	0	1	1	0	0	1	0	1	0	0	0	1	0	1	1	0	0	0	1	0	0	0	0
moment	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
multimedia	0	0	0	0	0	0	1	0	1	1	0	0	0	0	0	0	0	1	1	0	1	0	0	0	1	0	0
optimisation	1	1	0	0	1	0	1	0	1	1	1	1	1	0	1	1	0	1	1	1	1	1	0	0	1	0	1
probability	0	0	0	1	0	0	1	0	0	1	1	1	1	0	1	0	0	1	1	1	1	0	1	0	0	0	1
processing	1	1	0	1	1	1	1	1	1	1	1	0	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1
protocol	0	1	1	1	0	0	0	1	1	0	0	0	1	0	1	1	0	1	1	0	1	0	1	0	1	0	1
service	1	1	0	1	0	0	0	0	1	1	1	0	1	0	1	0	0	1	1	0	1	0	1	1	1	0	1
solution	1	1	0	1	1	1	1	0	1	1	1	1	1	1	1	1	0	1	0	0	1	0	1	1	1	1	1
sorting	0	1	0	1	0	0	0	0	0	1	0	0	0	0	1	0	0	1	1	1	1	0	1	0	0	0	0
speech	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0
standard	0	0	1	1	1	1	1	1	0	1	0	1	1	1	1	1	0	1	1	0	1	1	1	1	1	0	0
supplier	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
technology	1	1	1	1	0	1	1	1	1	1	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1	1
text	0	1	0	0	0	0	0	1	1	0	0	1	0	0	1	1	0	1	0	1	1	0	1	0	1	0	0
theory	0	1	0	1	0	1	1	0	1	1	1	1	1	0	0	1	0	1	1	1	1	1	1	0	0	1	1
user	1	1	1	1	0	1	0	0	1	1	0	1	1	1	1	1	0	1	1	1	1	1	1	1	1	0	1
word	0	0	0	0	0	0	1	0	0	0	0	1	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0

Table 7.1: Participant answers for the domain modelling survey

Rank	Word	DM Score
1	base	3.8949542609
2	efficiency	3.8592498038
3	efficient	3.8592498038
4	technique	3.7884674328
5	introduction	3.7393271442
6	analysis	3.7269523981
7	algorithm	3.7001272799
8	algorithms	3.7001272799
9	optimal	3.6258621677
10	optimization	3.6258621677
11	computer	3.6136895738
12	computational	3.6136895738
13	computation	3.6136895738
14	computing	3.6136895738
15	problem	3.5608560708
16	system	3.5568410382
17	systems	3.5568410382
18	approach	3.5463389032
19	methods	3.5434121714
20	method	3.5434121714
21	ctr	3.5377950901
22	application	3.5187717837
23	performance	3.4919803439
24	study	3.4701165522
25	modeling	3.4305648049
26	model	3.4305648049
27	improvement	3.4054494531
28	generality	3.3986252447
29	generalization	3.3986252447
30	implementation	3.3601326944
31	time	3.3527342872
32	distribution	3.3408537137
33	design	3.3354773111
34	requirement	3.3333685715
35	data	3.332403392
36	development	3.3292013228
37	programming	3.2630147149
38	program	3.2630147149
39	evaluation	3.2205254309
40	practice	3.2044215755
41	paper	3.1930022553
42	complexity	3.1627362248
43	relation	3.1540331505
44	effect	3.1284869784
45	section	3.1210284024
46	generation	3.1087926733
47	result	3.0822165792
48	proceedings	3.0816172287
49	difference	3.0774241496
50	control	3.0766234694

Table 7.2: Top 50 ranked words from an automatically constructed domain model

References

[AMBB+07] Boanerges Aleman-Meza, Uldis Bojars, Harold Boley, John G. Breslin, Malgorzata Mochol, Lyndon Jb Nixon, Axel Polleres, and Anna V. Zhdanova. Combining rdf vocabularies for expert finding. In *Proceedings of the 4th European Semantic Web Conference (ESWC2007)*, number 4519 in Lecture Notes in Computer Science, pages 235–250. Springer, 2007. 20

[Ana94] Sophia Ananiadou. A methodology for automatic term recognition. In *Proceedings of the 15th conference on Computational linguistics-Volume 2*, pages 1034–1038. Association for Computational Linguistics, 1994. 27

[AW02] R.B. Allen and Yejun Wu. Generality of texts. In Ee-Peng Lim, Schubert Foo, Chris Khoo, Hsinchun Chen, Edward Fox, Shalini Urs, and Thanos Costantino, editors, *Digital Libraries: People, Knowledge, and Technology*, volume 2555 of *Lecture Notes in Computer Science*, pages 111–116. Springer Berlin Heidelberg, 2002. 5

[BAdR09] Krisztian Balog, Leif Azzopardi, and Maarten de Rijke. A language modeling framework for expert finding. *Information Processing & Management*, 45(1):1–19, 2009. 99

[Bar04] Marc Barthelemy. Betweenness centrality in large complex networks. *The European Physical Journal B-Condensed Matter and Complex Systems*, 38(2):163–168, 2004. 130

[BB10a] Georgeta Bordea and Paul Buitelaar. Deriunlp: A context based approach to automatic keyphrase extraction. In *Proceedings of the 5th*

international workshop on semantic evaluation, pages 146–149. Association for Computational Linguistics, 2010. 52, 57

[BB10b] Georgeta Bordea and Paul Buitelaar. Expertise mining. In *Proceedings of the 21st National Conference on Artificial Intelligence and Cognitive Science, Galway, Ireland*, 2010. 85

[BBA+07] Krisztian Balog, Toine Bogers, Leif Azzopardi, Maarten de Rijke, and Antal van den Bosch. Broad expertise retrieval in sparse data environments. In *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '07, pages 551–558, New York, NY, USA, 2007. ACM. 6, 10, 32, 88, 91, 92, 98

[BBB+13] Richard Berendsen, Krisztian Balog, Toine Bogers, Antal van den Bosch, and Maarten de Rijke. On the assessment of expertise profiles. *Journal of the American Society for Information Science and Technology (JASIST)*, 2013. 86

[BBQ04] C. Biemann, S. Bordag, and U. Quasthoff. Automatic acquisition of paradigmatic relations using iterated co-occurrences. In *Proceedings of the 4th International Conference on Language Resources and Evaluation*, Lisbon, Portugal, 2004. 31

[BC00] Ken Barker and Nadia Cornacchia. Using Noun Phrase Heads to Extract Document Keyphrases. In *Canadian Conference on AI*, pages 40–52. Springer, 2000. 27

[BCB03] Katy Börner, Chaomei Chen, and Kevin W Boyack. Visualizing knowledge domains. *Annual review of information science and technology*, 37(1):179–255, 2003. 126

[BCdVS07] Peter Bailey, Nick Craswell, Arjen P de Vries, and Ian Soboroff. Overview of the trec 2007 enterprise track draft. In *TREC 2007 Working notes*, 2007. 16, 91, 92

[BCM05] Paul Buitelaar, Philipp Cimiano, and Bernardo Magnini. Ontology learning from text: An overview. In *Paul Buitelaar, P., Cimiano, P., Magnini B. (Eds.), Ontology Learning from Text: Methods, Applications and Evaluation*, pages 3–12. IOS Press, 2005. 32

[BdR07] Krisztian Balog and Maarten de Rijke. Determining expert profiles (with an application to expert finding). In *proc. of the International Joint Conferences on Artificial Intelligence (IJCAI 2007)*, 2007. 4, 9, 18, 85

[BdRA06] Krisztian Balog, Maarten de Rijke, and Leif Azzopardi. Formal models for expert finding in enterprise corpora. In *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval - SIGIR '06*, pages 43–50, New York, New York, USA, 2006. ACM Press. 17, 88, 139

[BdRW08] Krisztian Balog, Maarten de Rijke, and Wouter Weerkamp. Bloggers as experts: feed distillation using expert retrieval models. In *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '08, pages 753–754, New York, NY, USA, 2008. ACM. 16

[BE08] P. Buitelaar and T. Eigner. Topic extraction from scientific literature for competency management. In *Personal Identification and Collaborations: Knowledge Mediation and Extraction (PICKME2008)*, 2008. 15

[BF00] Irma Becerra-Fernandez. The role of artificial intelligence technologies in the implementation of people-finder knowledge management systems. *Knowledge-Based Systems*, (13(5)):315–320, 2000. 6, 13

[BF10] Gabor Berend and Richard Farkas. SZTERGAK : Feature Engineering for Keyphrase Extraction. In *Proceedings of the ACL 2010 Workshop on Evaluation Exercises on Semantic Evaluation (SemEval 2010)*, number July, pages 186–189, 2010. 28, 29, 30

[BGJT04] David M. Blei, Thomas L. Griffiths, Michael I. Jordan, and Joshua B. Tenenbaum. Hierarchical topic models and the nested chinese restaurant

process. In *Advances in Neural Information Processing Systems*, page 2003. MIT Press, 2004. 56

[BGLL08] Vincent D Blondel, Jean-Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre. Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2008(10):P10008, 2008. 130

[BHBL09] Christian Bizer, Tom Heath, and Tim Berners-Lee. Linked data-the story so far. *International journal on semantic web and information systems*, 5(3):1–22, 2009. 129

[BJ99] Didier Bourigault and Christian Jacquemin. Term extraction + term clustering: an integrated platform for computer-aided terminology. In *Proceedings of the ninth conference on European chapter of the Association for Computational Linguistics*, EACL '99, pages 15–22, Stroudsburg, PA, USA, 1999. Association for Computational Linguistics. 27, 43

[BKBP12] Georgeta Bordea, Sabrina Kirrane, Paul Buitelaar, and Bianca O Pereira. Expertise mining for enterprise content management. In *The International Conference on Language Resources and Evaluation (LREC 2012), Istanbul, Turkey*, pages 3495–3498, 2012. 139

[BKH+11] Dominik Benz, Christian Krner, Andreas Hotho, Gerd Stumme, and Markus Strohmaier. One tag to bind them all: Measuring term abstractness in social metadata. In Grigoris Antoniou, Marko Grobelnik, Elena Simperl, Bijan Parsia, Dimitris Plexousakis, Pieter Leenheer, and Jeff Pan, editors, *The Semanic Web: Research and Applications*, volume 6644 of *Lecture Notes in Computer Science*, pages 360–374. Springer Berlin Heidelberg, 2011. 5, 39

[BLHH+06] Tim Berners-Lee, Wendy Hall, James Hendler, Nigel Shadbolt, and Danny Weitzner. Creating a science of the web. *Science*, 313(5788):769–771, 2006. 129

[BMP02] Roberto Basili, Alessandro Moschitti, and Maria Teresa Pazienza. Empirical investigation of fast text classification over linguistic features. In Frank van Harmelen, editor, *ECAI*, pages 485–489. IOS Press, 2002. 22

[BMWM12] Anton Bakalov, Andrew McCallum, Hanna Wallach, and David Mimno. Topic models for taxonomies. In *Proceedings of the 12th ACM/IEEE-CS joint conference on Digital Libraries*, JCDL '12, pages 237–240, New York, NY, USA, 2012. ACM. 32

[BNC00] G. Bisson, C. Nédellec, and D. Canamero. Designing clustering methods for ontology building: The Mo'K workbench. In S. Staab, A. Maedche, C. Nédellec, and P. Wiemer-Hastings, editors, *Proceedings of the First Workshop on Ontology Learning OL'2000*, Berlin, Germany, August 2000. 31

[BNJ03] David M Blei, Andrew Y Ng, and Michael I Jordan. Latent dirichlet allocation. *the Journal of Machine Learning research*, 3:993–1022, 2003. 19, 38, 56

[Bor10] Georgeta Bordea. Concept extraction applied to the task of expert finding. In *The Semantic Web: Research and Applications*, pages 451–456. Springer, 2010. 85

[Bou92] Didier Bourigault. Surface grammatical analysis for the extraction of terminological noun phrases. In *14th International Conference on Computational Linguistics - COLING 92*, pages 977–981, Nantes, France, 1992. 23

[Bow12] Shawn Bowers. Scientific workflow, provenance, and data modeling challenges and approaches. *Journal on Data Semantics*, 1(1):19–30, 2012. 138

[BPB13] Georgeta Bordea, Tamara Polajnar, and Paul Buitelaar. Domain-independent term extraction through domain modelling. In *10th International Conference on Terminology and Artificial Intelligence*, 2013. 35

[BV00] Chris Buckley and Ellen M. Voorhees. Evaluating evaluation measure stability. In *Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '00, pages 33–40, New York, NY, USA, 2000. ACM. 48

[BYRN99] Ricardo Baeza-Yates and Berthier Ribeiro-Neto. Modern information retrieval, 1999. 86

[CAMA+10] Delroy Cameron, Boanerges Aleman-Meza, Ismailcem Budak Arpinar, Sheron L Decker, and Amit P Sheth. A taxonomy-based model for expertise extrapolation. In *Semantic Computing (ICSC), 2010 IEEE Fourth International Conference on*, pages 333–340. IEEE, 2010. 6, 10

[CC99] Chaomei Chen and Les Carr. Trailblazing the literature of hypertext: author co-citation analysis. In *Proceedings of the tenth ACM Conference on Hypertext and hypermedia: returning to our diverse roots: returning to our diverse roots*, pages 51–60. ACM, 1999. 126

[CCTB83] Michel Callon, Jean-Pierre Courtial, William A Turner, and Serge Bauin. From translations to problematic networks: An introduction to co-word analysis. *Social Science Information*, 22(2):191–235, 1983. 127

[CdVS06] Nick Craswell, Arjen P de Vries, and Ian Soboroff. Overview of the trec-2005 enterprise track. In *The fourteenth Text REtrieval Conference Proceedings (TREC 2005)*, 2006. 16

[CGHH91] Kenneth Church, William Gale, Patrick Hanks, and Donald Hindle. Using statistics in lexical analysis. In *Lexical Acquisition: Exploiting On-Line Resources to Build a Lexicon*, pages 115–164. Erlbaum, 1991. 23

[CHS04] Philipp Cimiano, Andreas Hotho, and Steffen Staab. Comparing conceptual, divise and agglomerative clustering for learning taxonomies from text. In *Proceedings of the European Conference on Artificial Intelligence (ECAI)*, pages 435–439, 2004. 113

[CHS05] Philipp Cimiano, Andreas Hotho, and Steffen Staab. Learning concept hierarchies from text corpora using formal concept analysis. *J. Artif. Intell. Res.(JAIR)*, 24:305–339, 2005. 75

[CHVW01] Nick Craswell, David Hawking, Anne-Marie Vercoustre, and Peter Wilkins. P@ noptic expert: Searching for experts not just for documents. In *Ausweb Poster Proceedings, Queensland, Australia*, 2001. 15, 17

[CL65] Yoeng-Jin Chu and Tseng-Hong Liu. On the shortest arborescence of a directed graph. *Science Sinica*, 14(1396-1400):270, 1965. 82

[Cle67] Cyril Cleverdon. The cranfield tests on index language devices. In *Aslib proceedings*, volume 19, pages 173–194. MCB UP Ltd, 1967. 138

[CMBT02] H. Cunningham, D. Maynard, K. Bontcheva, and V. Tablan. GATE: A Framework and Graphical Development Environment for Robust NLP Tools and Applications. In *Proceedings of the 40th Anniversary Meeting of the Association for Computational Linguistics*, 2002. 120

[CMCD03] 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. ACM, 2003. 16

[CMK98] Neal Coulter, Ira Monarch, and Suresh Konda. Software engineering as seen through its research literature: A study in co-word analysis. *Journal of the American Society for Information Science*, 49(13):1206–1223, 1998. 126, 127

[Dai96] Béatrice Daille. Study and implementation of combined techniques for automatic extraction of terminology. In Judith Klavans and Philip Resnik, editors, *The Balancing Act: Combining Symbolic and Statistical Approaches to Language*, pages 49–66. The MIT Press, Cambridge, Massachusetts, 1996. 23

[Dam90] Fred J. Damerau. Evaluating computer-generated domain-oriented vocabularies. *Information Processing & Management*, 26(6):791 – 801, 1990. 23

[Dam93] Fred J. Damerau. Generating and evaluating domain-oriented multi-word terms from texts. *Information Processing & Management*, 29(4):433 – 447, 1993. 23

[DB07] Jrg Diederich and Wolf-Tilo Balke. The semantic growbag algorithm: Automatically deriving categorization systems. In Lszl Kovcs, Norbert Fuhr, and Carlo Meghini, editors, *Research and Advanced Technology for Digital Libraries*, volume 4675 of *Lecture Notes in Computer Science*, pages 1–13. Springer Berlin Heidelberg, 2007. 32

[DDL+90] Scott C. Deerwester, Susan T Dumais, Thomas K. Landauer, George W. Furnas, and Richard A. Harshman. Indexing by latent semantic analysis. *JASIS*, 41(6):391–407, 1990. 88

[DE84] William HE Day and Herbert Edelsbrunner. Efficient algorithms for agglomerative hierarchical clustering methods. *Journal of classification*, 1(1):7–24, 1984. 113

[Dem07] Gianluca Demartini. Finding experts using wikipedia. In *Proceedings of the Workshop on Finding Experts on the Web with Semantics (FEWS2007) at ISWC/ASWC2007*, pages 33–41, 2007. 46

[DFI+08] Gianluca Demartini, Claudiu S Firan, Tereza Iofciu, Ralf Krestel, and Wolfgang Nejdl. A model for ranking entities and its application to wikipedia. In *Latin American Web Conference, 2008. LA-WEB'08.*, pages 29–38. IEEE, 2008. 16

[Dil11] Rod. Dilnutt. Surviving the Information explosion? *IEE Engineering Management*, 118(2):59, February 2011. 139

[DKLK08] Hongbo Deng, Irwin King, Michael R Lyu, and Hong Kong. Formal models for expert finding on dblp bibliography data. In *Proceedings of the Eighth IEEE International Conference on Data Mining, 2008. ICDM '08.*, pages 163–172, 2008. 19

[DM06] F. Draganidis and G. Metzas. Competency based management: A review of systems and approaches. *Information Management and Computer Security*, 14(1):51–64, 2006. 14

[DN08] Gianluca Demartini and Claudia Niedere. Finding experts on the semantic desktop. In *Proceedings of the the 7th International Semantic Web Conference*, 2008. 16

[Dun94] T. Dunning. Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics*, 19(1):61–74, 1994. 23

[EbR09] Samhaa R El-beltagy and Ahmed Rafea. KP-Miner : A keyphrase extraction system for English and Arabic documents. *Information Systems Journal*, 34(1):132– 144, 2009. 28, 29

[EBR10] Samhaa R El-Beltagy and Ahmed Rafea. Kp-miner: Participation in semeval-2. In *Proceedings of the 5th international workshop on semantic evaluation*, pages 190–193. Association for Computational Linguistics, 2010. 59

[Edm67] Jack Edmonds. Optimum branchings. *Journal of Research of the National Bureau of Standards B*, 71(4):233–240, 1967. 82

[EH09] Kathrin Eichler and Holmer Hemsen. Unsupervised and domain-independent extraction of technical terms from scientific articles in digital libraries. *Proceedings pf the LWA Information Retrieval Workshop*, 2009. 29, 30

[EN10] Kathrin Eichler and Gunter Neumann. DFKI KeyWE : Ranking keyphrases extracted from scientific articles. *Proceedings of the ACL 2010 Workshop on Evaluation Exercises on Semantic Evaluation (SemEval 2010)*, 2010. 29

[FAM00] Katerina Frantzi, Sophia Ananiadou, and Hideki Mima. Automatic recognition of multi-word terms : the C-value / NC-value method. *Journal on Digital Libraries, Natural language processing for digital libraries*, 3 (2):115–130, 2000. 23, 24, 28, 42, 43, 44

[FDW06] Trudi Farrington-Darby and John R Wilson. The nature of expertise: A review. *Applied Ergonomics*, 37(1):17–32, 2006. 7

[FPW+99] Eibe Frank, Gordon W Paynter, Ian H Witten, Carl Gutwin, and Craig G Nevill-Manning. Domain-Specific Keyphrase Extraction. In *Proceedings of the 16th International Joint Conference on Aritfiicial Intelligence*, pages 668–673, 1999. 28, 29

[FZFY12] Maryam Fazel-Zarandi, Mark S Fox, and Eric Yu. Ontologies in expertise finding systems: Modeling, analysis, and design. *Ontology-Based Applications for Enterprise Systems and Knowledge Management*, page 158, 2012. 14

[Gau98] Éric Gaussier. Flow network models for word alignment and terminology extraction from bilingual corpora. In *Proceedings of the 17th international conference on Computational linguistics-Volume 1*, pages 444–450. Association for Computational Linguistics, 1998. 22

[Gre05] Rebecca Green. Vocabulary alignment via basic level concepts. In *Final Report, 2003 OCLC/ALISE Library and Information Science Research Grant Project*, Dublin, OH: OCLC, 2005. 39

[Haj13] Lala Hajibayova. Basic-level categories: A review. *Journal of Information Science*, 2013. 39

[Har68] Zellig Harris. *Mathematical Structures of Language.* John Wiley and Son, New York, 1968. 31

[HBB13] Clare J Hooper, Georgeta Bordea, and Paul Buitelaar. Web science and the two (hundred) cultures: Representation of disciplines publishing in web science. In *Proceedings of Web Science 2013*, 2013. 124

[HBBdR10] Katja Hofmann, Krisztian Balog, Toine Bogers, and Maarten de Rijke. Contextual Factors for Finding Similar Experts. *Journal of the American Society for Information Science*, 61(5):994–1014, 2010. 1

[Hea92] Marti A. Hearst. Automatic acquisition of hyponyms from large text corpora. In *Proceedings of the 14th International Conference on Computational Linguistics*, pages 539–545, 1992. 31, 32, 80

[HGEF07] Nathalie Henry, Howard Goodell, Niklas Elmqvist, and Jean-Daniel Fekete. 20 years of four hci conferences: A visual exploration. *International Journal of Human-Computer Interaction*, 23(3):239–285, 2007. 126

[HMK12] Clare J. Hooper, Nicolas Marie, and Evangelos Kalampokis. Dissecting the butterfly: representation of disciplines publishing at the web science conference series. In Noshir S. Contractor, Brian Uzzi, Michael W. Macy, and Wolfgang Nejdl, editors, *WebSci*, pages 137–140. ACM, 2012. 125, 126

[Hof99] Thomas Hofmann. Probabilistic latent semantic indexing. In *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*, pages 50–57. ACM, 1999. 88

[HSH+08] James Hendler, Nigel Shadbolt, Wendy Hall, Tim Berners-Lee, and Daniel Weitzner. Web science: an interdisciplinary approach to understanding the web. *Communications of the ACM*, 51(7):60–69, 2008. 129

[Hu05] Yifan Hu. Efficient, high-quality force-directed graph drawing. *Mathematica Journal*, 10(1):37–71, 2005. 74, 130

[Hui86] Y Huizhong. A new technique for identifying scientific/technical terms and describing science texts. *Lit. Linguist. Comput.*, 1:93–103, April 1986. 23

[Hul04] Anette Hulth. Enhancing Linguistically Oriented Automatic Keyword Extraction. In *Proceedings of HLT/NAACL: Short Papers*, pages 17–20, 2004. 27, 28

[JK95] John S. Justeson and Slava M. Katz. Technical terminology: some linguistic properties and an algorithm for identification in text. *Natural Language Engineering*, 1(01):9–27, 1995. 23

[JLsK+07] Hanmin Jung, Mikyoung Lee, In su Kang, Seung woo Lee, and Won kyung Sung. Finding topic-centric identified experts based on full text

analysis. In *Proceedings of the 2nd International Workshop on Finding Experts on the Web with Semantics (FEWS07)*, pages 56–63, 2007. 101

[JR10] Nikhil Johri and Dan Roth. Experts retrieval with multiword-enhanced author topic model. *Computational Linguistics*, (June):10–18, 2010. 19

[JS07] Hideo Joho and Mark Sanderson. Document frequency and term specificity. In *Large Scale Semantic Access to Content (Text, Image, Video, and Sound)*, pages 350–359. Le centre de hautes etudes internationales d'informatique documentaire, 2007. 61

[KAM09] Mikalai Krapivin, Aliaksandr Autayeu, and Maurizio Marchese. Large dataset for keyphrases extraction. In *Technical Report DISI-09-055, DISI*. University of Trento, Italy, 2009. 53, 54

[KH10] Zornitsa Kozareva and Eduard Hovy. A semi-supervised method to learn and construct taxonomies using the web. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, EMNLP '10, pages 1110–1118, Stroudsburg, PA, USA, 2010. Association for Computational Linguistics. 9, 33, 72, 76, 78, 103, 139

[KMKB10] Su Nam Kim, Alyona Medelyan, Min-Yen Kan, and Timothy Baldwin. SemEval-2010 Task 5: Automatic Keyphrase Extraction from Scientific Articles. In *Proceedings of the ACL 2010 Workshop on Evaluation Exercises on Semantic Evaluation (SemEval 2010)*, 2010. 52

[KRH08] Zornitsa Kozareva, Ellen Riloff, and Eduard Hovy. Semantic Class Learning from the Web with Hyponym Pattern Linkage Graphs. In *Proceedings of ACL-08: HLT*, pages 1048–1056, Columbus, Ohio, June 2008. Association for Computational Linguistics. 32

[KS96] Henry Kautz and Bart Selman. Agent amplified communication. In *Proceedings of the thirteenth national conference on Artificial intelligence - Volume 1*, AAAI'96, pages 3–9. AAAI Press, 1996. 14

[KU96] K. Kageura and B. Umino. Methods of automatic term recognition: A review. *Terminology*, 3(2):259–289, 1996. 23, 37

[KVM00] Jörg-Uwe Kietz, Raphael Volz, and Alexander Maedche. Extracting a domain-specific ontology from a corporate intranet. In *Proceedings of the 2nd workshop on Learning language in logic and the 4th conference on Computational natural language learning - Volume 7*, ConLL '00, pages 167–175, Stroudsburg, PA, USA, 2000. Association for Computational Linguistics. 22

[LAT10] Atif Latif, Muhammad Tanvir Afzal, and Klaus Tochtermann. Constructing experts profiles from linked open data. In *Proceedings of the 6th International Conference on Emerging Technologies (ICET)*, pages 33 – 38, 2010. 21

[LHCHVH10] AG López-Herrera, MJ Cobo, E Herrera-Viedma, and F Herrera. A bibliometric study about the research based on hybridating the fuzzy logic field and the other computational intelligent techniques: A visual approach. *International Journal of Hybrid Intelligent Systems*, 7(1):17–32, 2010. 127

[LR10] Patrice Lopez and Laurent Romary. HUMB : Automatic Key Term Extraction from Scientific Articles in GROBID. In *Proceedings of the ACL 2010 Workshop on Evaluation Exercises on Semantic Evaluation (SemEval 2010)*, number July, pages 248–251, 2010. 22, 27, 28, 29, 30, 31, 59

[LWY+05] Tao Liu, X Wang, Guan Yi, Zhi-Ming Xu, and Qiang Wang. *Domain-Specific Term Extraction and Its Application in Text Classification*, volume 1481, pages 1481–1484. 2005. 23

[LYL08] Ping Liu, Yan Ye, and Kan Liu. Building a semantic repository of academic experts. In *Wireless Communications, Networking and Mobile Computing, 2008. WiCOM '08. 4th International Conference on*, pages 1–6, 2008. 101

[MAM06] Hideki Mima, Sophia Ananiadou, and Katsumori Matsushima. Terminology-based knowledge mining for new knowledge discovery. *ACM Trans. Asian Lang. Inf. Process.*, 5(1):74–88, 2006. 22

[MBSB10] Fergal Monaghan, Georgeta Bordea, Krystian Samp, and Paul Buitelaar. Exploring your research: Sprinkling some saffron on semantic web dog food. In *Semantic Web Challenge at the International Semantic Web Conference*, 2010. 117

[McC02] Andrew Kachites McCallum. Mallet: A machine learning for language toolkit. http://mallet.cs.umass.edu, 2002. 56

[Med09] Olena Medelyan. *Human-competitive automatic topic indexing.* PhD thesis, The University of Waikato, 2009. 30

[MFW09] Olena Medelyan, Eibe Frank, and Ian H Witten. Human-competitive tagging using automatic keyphrase extraction. *Proceedings of the International Conference for Empirical Methods in Natural Language Processing (EMNLP)*, 2009. 28

[MH06] Bjørn Erik Munkvold and Anne Kristine Hodne. Contemporary Issues of Enterprise Content Management : The Case of Statoil. *Journal Scandinavian Journal of Information Systems*, 2(18), 2006. 139

[MM07a] David Mimno and Andrew McCallum. Expertise Modeling for Matching Papers with Reviewers. In *SIGKDD '07: Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 500–509, 2007. 1

[MM07b] David Mimno and Andrew Mccallum. Expertise modelling for matching papers with reviewers. In *Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 500–509, 2007. 16, 19, 85

[MO06] Craig Macdonald and Iadh Ounis. Voting for candidates: adapting data fusion techniques for an expert search task. In *CIKM '06: Proceedings of the 15th ACM international conference on Information and knowledge management*, pages 387–396, New York, NY, USA, 2006. ACM. 17, 139

[Mos03] Alessandro Moschitti. A study on optimal parameter tuning for rocchio text classifier. In Fabrizio Sebastiani, editor, *Advances in Information*

Retrieval, volume 2633 of *Lecture Notes in Computer Science*, pages 420–435. Springer Berlin Heidelberg, 2003. 37

[MPG02] Bernardo Magnini, Giovanni Pezzulo, and Alfio Gliozzo. The role of domain information in word sense disambiguation. *Natural Language Engineering*, 8:359–373, 2002. 38

[MT04] Rada Mihalcea and Paul Tarau. Textrank: Bringing order into texts. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 404–411, 2004. 27

[MW08] Olena Medelyan and Ian H. Witten. Domain independent automatic keyphrase indexing with small training sets. *J. Am. Soc. Information Science and Technology*, 2008. 53

[Nav12] Roberto Navigli. A quick tour of word sense disambiguation, induction and related approaches. In *SOFSEM 2012: Theory and practice of computer science*, pages 115–129. Springer, 2012. 2

[NDH11] T Nagel, E. Duval, and F. Heidmann. Exploring the geospatial network of scientific collaboration on a multitouch table. In *Proceedings of the 22nd ACM Conference on Hypertext and Hypermedia (demo)*, pages 51–60, 2011. 126

[NFS+11] Roberto Navigli, Stefano Faralli, Aitor Soroa, Oier de Lacalle, and Eneko Agirre. Two birds with one stone: learning semantic models for text categorization and word sense disambiguation. In *Proceedings of the 20th ACM international conference on Information and knowledge management*, CIKM '11, pages 2317–2320, New York, NY, USA, 2011. ACM. 37

[NL10] Thuy Dung Nguyen and Minh-Thang Luong. Wingnus: Keyphrase extraction utilizing document logical structure. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, SemEval '10, pages 166–169, Stroudsburg, PA, USA, 2010. Association for Computational Linguistics. 28, 29

[NLGB10] David Newman, Jey Han Lau, Karl Grieser, and Timothy Baldwin. Automatic evaluation of topic coherence. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 100–108. Association for Computational Linguistics, 2010. 40

[NMWP99] Craig G. Nevill-Manning, Ian H. Witten, and Gordon W. Paynter. Lexically-generated subject hierarchies for browsing large collections. *International Journal on Digital Libraries*, 2:111–123, 1999. 32, 82

[NS+72] Allen Newell, Herbert Alexander Simon, et al. *Human problem solving*, volume 14. Prentice-Hall Englewood Cliffs, NJ, 1972. 5

[NVF11] Roberto Navigli, Paola Velardi, and Stefano Faralli. A graph-based algorithm for inducing lexical taxonomies from scratch. In *Proceedings of the Twenty-Second international joint conference on Artificial Intelligence - Volume Volume Three*, IJCAI'11, pages 1872–1877. AAAI Press, 2011. 9, 33, 38, 72, 80, 81, 103, 104, 139

[NZI+09] Shinsuke Nakajima, Jianwei Zhang, Yoichi Inagaki, Tomoaki Kusano, and Reyn Nakamoto. Blog ranking based on bloggers knowledge level for providing credible information. In Gottfried Vossen, Darrell Long, and Jeffrey Yu, editors, *Web Information Systems Engineering - WISE 2009*, volume 5802 of *Lecture Notes in Computer Science*, pages 227–234. Springer Berlin / Heidelberg, 2009. 16

[OTK+01] Tomoko Ohta, Yuka Tateisi, Jin-Dong Kim, Sang-Zoo Lee, and Jun'ichi Tsujii. Genia corpus: A semantically annotated corpus in molecular biology domain. In *Proceedings of the ninth International Conference on Intelligent Systems for Molecular Biology (ISMB 2001) poster session*, page 68, July 2001. 53

[Paq07] Gilbert Paquette. An ontology and a software framework for competency modeling and management. *Educational Technology & Society*, 10(3):1–21, 2007. 4, 5, 14, 20, 89

[PBB02] Youngja Park, Roy J. Byrd, and Branimir Boguraev. Automatic glossary extraction: Beyond terminology identification. In *19th International Conference on Computational Linguistics - COLING 02*, Taipei, Taiwan, August-September 2002. Howard International House and Academia Sinica. 23, 28, 29

[PC98] Jay M. Ponte and W. Bruce Croft. A language modeling approach to information retrieval. In *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '98, pages 275–281, New York, NY, USA, 1998. ACM. 16

[PC06] Desislava Petkova and W. Bruce Croft. Hierarchical language models for expert finding in enterprise corpora. In *Proceedings of the 18th IEEE International Conference on Tools with Artificial Intelligence*, ICTAI '06, pages 599–608, Washington, DC, USA, 2006. IEEE Computer Society. 17, 88

[PH04] V. Posea and M. Harzallah. Building a competence ontology. In *proc. of the workshop Enterprise modelling and Ontology of the International Conference on Practical Aspects of Knowledge Management (PAKM 2004)*, 2004. 20

[PNMH08] Mari-Sanna Paukkeri, Ilari T. Nieminen, Polla Matti, and Timo Honkela. A Language-Independent Approach to Keyphrase Extraction and Evaluation. In *Coling 2008 Posters*, number August, pages 83–86, 2008. 29, 43

[PS11] F. Piazza and S. Strohmeier. Domain-driven data mining in human resource management: A review. In *Data Mining Workshops (ICDMW), 2011 IEEE 11th International Conference on*, pages 458 –465, dec. 2011. 15

[PT10] Emanuele Pianta and Sara Tonelli. KX : A flexible system for Keyphrase eXtraction. In *Proceedings of the ACL 2010 Workshop on Evaluation*

Exercises on Semantic Evaluation (SemEval 2010), number July, pages 170–173, 2010. 28, 59

[RB08] Marko A. Rodriguez and Johan Bollen. An Algorithm to Determine Peer-Reviewers. In *'08: Proceedings of the Seventeenth International Conference on Information and Knowledge Management*, pages 319–328. ACM, 2008. 1

[RS97] Ellen Riloff and Jessica Shepherd. A corpus-based approach for building semantic lexicons. In *Proceedings of the Second Conference on Empirical Methods in Natural Language Processing*, pages 117–124, 1997. 45

[RSN96] Maxine Robertson, Jacky Swan, and Sue Newell. The role of networks in the diffusion of technological innovation*. *Journal of Management Studies*, 33(3):333–359, 1996. 1

[RZCG+10] Michal Rosen-Zvi, Chaitanya Chemudugunta, Thomas Griffiths, Padhraic Smyth, and Mark Steyvers. Learning author-topic models from text corpora. *ACM Transactions on Information Systems*, 28(1):1–38, 2010. 19

[SB08] Andreas Schmidt and Simone Braun. People tagging & ontology maturing: Towards collaborative competence management. In *8th International Conference on the Design of Cooperative Systems (COOP 2008), Carry-le-Rouet*, 2008. 21

[SB11] Elena Smirnova and Krisztian Balog. A user-oriented model for expert finding. In *Proceedings of the 33rd European conference on Advances in information retrieval*, ECIR'11, pages 580–592, Berlin, Heidelberg, 2011. Springer-Verlag. 92

[SBL08] Nigel Shadbolt and Tim Berners-Lee. Web science emerges. *Scientific American*, 299(4):76–81, 2008. 129

[SC99] Mark Sanderson and Bruce Croft. Deriving concept hierarchies from text. In *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '99, pages 206–213, New York, NY, USA, 1999. ACM. 32, 33, 55, 103

[Sco11] J.E. Scott. User Perceptions of an Enterprise Content Management System. In *hicss*, pages 1–9. IEEE Computer Society, 2011. 139

[SdVC07] Ian Soboroff, Arjen P de Vries, and Nick Craswell. Overview of the trec 2006 enterprise track. In *The fifteenth Text REtrieval Conference Proceedings (TREC 2006)*, 2007. 16, 91, 92

[SH08] Pavel Serdyukov and Djoerd Hiemstra. Modeling documents as mixtures of persons for expert finding. In *Proceedings of the IR research, 30th European conference on Advances in information retrieval*, ECIR'08, pages 309–320, Berlin, Heidelberg, 2008. Springer-Verlag. 17, 88

[SHBL06] Nigel Shadbolt, Wendy Hall, and Tim Berners-Lee. The semantic web revisited. *Intelligent Systems, IEEE*, 21(3):96–101, 2006. 129

[Shn07] Ben Shneiderman. Web science: a provocative invitation to computer science. *Communications of the ACM*, 50(6):25–27, 2007. 129

[SHR07] Emilia Stoica, Marti A Hearst, and Megan Richardson. Automating creation of hierarchical faceted metadata structures. In *HLT-NAACL*, pages 244–251, 2007. 72

[SJL11] Milan Stankovic, Jelena Jovanovic, and Philippe Laublet. Linked Data Metrics for Flexible Expert Search on the Open Web. In *Proceedings of the 8th extended semantic web conference on The semantic web: research and applications - Volume Part I*, ESWC'11, pages 108–123, Berlin, Heidelberg, 2011. Springer-Verlag. 1, 21

[SJN04] Rion Snow, Daniel Jurafsky, and Andrew Y. Ng. Learning syntactic patterns for automatic hypernym discovery. In *Advances in Neural Information Processing Systems (NIPS 2004)*, November 2004. This is a draft version from the NIPS preproceedings; the final version will be published by April 2005. 32

[SJN06] Rion Snow, Daniel Jurafsky, and Andrew Y. Ng. Semantic taxonomy induction from heterogenous evidence. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual*

meeting of the Association for Computational Linguistics, ACL-44, pages 801–808, Stroudsburg, PA, USA, 2006. Association for Computational Linguistics. 33

[SRH08] Pavel Serdyukov, Henning Rode, and Djoerd Hiemstra. Exploiting sequential dependencies for expert finding. In *SIGIR '08: Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*, pages 795–796, New York, NY, USA, 2008. ACM. 16, 139

[STV+11] Pavel Serdyukov, Mike Taylor, Vishwa Vinay, Matthew Richardson, and RyenW. White. Automatic people tagging for expertise profiling in the enterprise. In Paul Clough, Colum Foley, Cathal Gurrin, GarethJ.F. Jones, Wessel Kraaij, Hyowon Lee, and Vanessa Mudoch, editors, *Advances in Information Retrieval*, volume 6611 of *Lecture Notes in Computer Science*, pages 399–410. Springer Berlin Heidelberg, 2011. 4, 18

[SV07] Francesco Sclano and Paola Velardi. Termextractor: a web application to learn the shared terminology of emergent web communities. In *Enterprise Interoperability II*, pages 287–290. Springer, 2007. 24

[TDO07] Giovanni Tummarello, Renaud Delbru, and Eyal Oren. Sindice. com: Weaving the open linked data. In *The Semantic Web*, pages 552–565. Springer, 2007. 47

[TH03] Takashi Tomokiyo and Matthew Hurst. A Language Model Approach to Keyphrase Extraction. In *Proceedings of the ACL 2003 workshop on Multiword expressions*, pages 33–40, 2003. 43

[TM02] Simone Teufel and Marc Moens. Summarizing scientific articles - experiments with relevance and rhetorical status. *Computational Linguistics*, 28:2002, 2002. 38

[TMS08] Rajesh Thiagarajan, Geetha Manjunath, and Markus Stumptner. *Finding experts by semantic matching of user profiles.* PhD thesis, CEUR-WS, 2008. 46

[TT91] James W. Tanaka and Marjorie Taylor. Object categories and expertise: Is the basic level in the eye of the beholder? *Cognitive Psychology*, 23(3):457–482, July 1991. 5, 89

[TTHG10] Pucktada Treeratpituk, Pradeep Teregowda, Jian Huang, and C. Lee Giles. Seerlab: A system for extracting key phrases from scholarly documents. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, SemEval '10, pages 182–185, Stroudsburg, PA, USA, 2010. Association for Computational Linguistics. 28, 29

[Tur00] Peter D Turney. Learning algorithms for keyphrase extraction. *Information Retrieval*, 2:303–336, 2000. 28

[Tur03] Peter D. Turney. Coherent keyphrase extraction via web mining. In *Proceedings of International Joint Conferences on Artificial Intelligence (IJCAI)*, pages 434–439, 2003. 29, 39

[TZY+08] Jie Tang, Jing Zhang, Limin Yao, Juanzi Li, Li Zhang, and Zhong Su. ArnetMiner: Extraction and Mining of Academic Social Networks. In *KDD '08: Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 990–998, New York, NY, USA, 2008. ACM. 93

[Vaf11] Michalis Vafopoulos. Web science subject categorization (wssc). *Proceedings of ACM WebSci*, pages 1–13, 2011. 125

[VMB01] Paola Velardi, Michele Missikoff, and Roberto Basili. Identification of relevant terms to support the construction of domain ontologies. In *Proceedings of the ACL 2001 Workshop on Human Language Technology and Knowledge Management*, Toulouse, July 2001. 23, 29

[WBB10] Wei Wang, P. Barnaghi, and Andrzej Bargiela. Probabilistic topic models for learning terminological ontologies. *Knowledge and Data Engineering, IEEE Transactions on*, 22(7):1028–1040, 2010. 32

[WC06] Xing Wei and W. Bruce Croft. Lda-based document models for ad-hoc retrieval. In *Proceedings of the 29th annual international ACM SIGIR*

conference on Research and development in information retrieval, SIGIR '06, pages 178–185, New York, NY, USA, 2006. ACM. 19, 88

[WH06] Joachim Wermter and Udo Hahn. You can't beat frequency (unless you use linguistic knowledge): a qualitative evaluation of association measures for collocation and term extraction. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics*, ACL-44, pages 785–792, Stroudsburg, PA, USA, 2006. Association for Computational Linguistics. 44

[WLLL12] Zhong-Yi Wang, Gang Li, Chun-Ya Li, and Ang Li. Research on the semantic-based co-word analysis. *Scientometrics*, 90(3):855–875, 2012. 127

[WM98] Howard D White and Katherine W McCain. Visualizing a discipline: An author co-citation analysis of information science. *Journal of the American society for information science*, 49(4):327–355, 1998. 126

[WUH+08] Robert Wetzker, Winfried Umbrath, Leonhard Hennig, Christian Bauckhage, Tansu Alpcan, and Florian Metze. Tailoring taxonomies for efficient text categorization and expert finding. In *Web Intelligence and Intelligent Agent Technology, 2008. WI-IAT'08. IEEE/WIC/ACM International Conference on*, volume 3, pages 459–462. IEEE, 2008. 72

[YC09] Hui Yang and Jamie Callan. A metric-based framework for automatic taxonomy induction. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 1 - Volume 1*, ACL '09, pages 271–279, Stroudsburg, PA, USA, 2009. Association for Computational Linguistics. 33

[YGTH00] Roman Yangarber, Ralph Grishman, Pasi Tapanainen, and Silja Huttunen. Automatic acquisition of domain knowledge for information extraction. In *Proceedings of the 18th conference on Computational linguis-*

tics - Volume 2, COLING '00, pages 940–946, Stroudsburg, PA, USA, 2000. Association for Computational Linguistics. 22

[YJZY05] Lingpeng Yang, Dong-Hong Ji, Guodong Zhou, and Nie Yu. Improving retrieval effectiveness by using key terms in top retrieved documents. In David E. Losada and Juan M. Fernndez-Luna, editors, *ECIR*, volume 3408 of *Lecture Notes in Computer Science*, pages 169–184. Springer, 2005. 22

[YK00] D. Yimam and A. Kobsa. Demoir: a hybrid architecture for expertise modeling and recommender systems. In *Enabling Technologies: Infrastructure for Collaborative Enterprises, 2000. (WET ICE 2000). Proceedings. IEEE 9th International Workshops on*, pages 67 –74, 2000. 13

[ZAAN07] Jun Zhang, Mark S Ackerman, Lada Adamic, and Kevin Kyung Nam. Qume: a mechanism to support expertise finding in online help-seeking communities. In *Proceedings of the 20th annual ACM symposium on User interface software and technology*, pages 111–114. ACM, 2007. 16

[ZCMZ09] Wei Zhang, Lei Chang, Jianqing Ma, and Yiping Zhong. Aggregation models for people finding in enterprise corpora. In *Knowledge Science, Engineering and Management*, pages 180–191. Springer, 2009. 15

[ZGU+05] J. Zhu, A. L. Goncalves, V. S. Uren, E. Motta, and R. Pacheco. Mining web data for competency management. In *Proceedings of Web Intelligence (WI 2005*, pages 94–100. IEEE Computer Society, 2005. 15

[Zha98] Qiao Zhang. Fuzziness - vagueness - generality - ambiguity. *Journal of Pragmatics*, 29(1):13 – 31, 1998. 5

[ZIBC08] Ziqi Zhang, Jose Iria, Christopher Brewster, and Fabio Ciravegna. A comparative evaluation of term recognition algorithms. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC08), Marrakech, Morocco*, 2008. 23

[ZPVP07] Elias Zavitsanos, Georgios Paliouras, George A. Vouros, and Sergios Petridis. Discovering subsumption hierarchies of ontology concepts from

text corpora. In *Proceedings of the IEEE/WIC/ACM International Conference on Web Intelligence*, WI '07, pages 402–408, Washington, DC, USA, 2007. IEEE Computer Society. 32

[ZTL07] Jing Zhang, Jie Tang, and Juanzi Li. Expert finding in a social network. In *Advances in Databases: Concepts, Systems and Applications*, pages 1066–1069. Springer, 2007. 7