Automatically Structuring Domain Knowledge from Text: an Overview of Current Research

Malcolm Clark[a], Yunhyong Kim[a], Udo Kruschwitz[b], Dawei Song[a], Dyaa Albakour[b], Stephen Dignum[b], Ulises Cerviño Beresi[a], Maria Fasli[b], Anne De Roeck[c]

[a]*School of Computing and IDEAS Institute, Robert Gordon University, Aberdeen, United Kingdom*
[b]*School of Computer Science and Electronic Engineering, University of Essex, Colchester, United Kingdom*
[c]*Departments of Mathematics and Computing, Open University, Milton Keynes, United Kingdom*

Abstract

This paper presents an overview of automatic methods for building domain knowledge structures (domain models) from text collections. Applications of domain models have a long history within knowledge engineering and artificial intelligence. In the last couple of decades they have surfaced noticeably as a useful tool within natural language processing, information retrieval and semantic web technology. Inspired by the ubiquitous propagation of domain model structures that are emerging in several research disciplines, we give an overview of the current research landscape and some techniques and approaches. We will also discuss trade-offs between different approaches and point to some recent trends.

Key words: Domain models, Information retrieval, Natural language processing, Artificial intelligence

1. Introduction

This paper presents an overview of the research landscape in the automated construction of domain models from text collections. The aim of the paper is to facilitate the general understanding of domain models over multiple disciplines. Instead of giving a systematic review, we aim to illustrate current work and recent trends in three distinct communities in which

domain modelling has for decades been a line of research, namely artificial intelligence (and more recently semantic web), natural language processing and information retrieval. There has traditionally been little overlap between these communities, but increasingly there are problem domains, such as biomedical information retrieval and text mining, that make use of hybrid approaches and techniques developed by all these communities.

Domain modelling can generally be defined as the process of capturing and structuring knowledge embedded within a selected domain (for example, a collection of documents, a community, an area of interest). Domain models can be realised in many ways, for example, as an organisation of documents into a classification schema, as a linked network of information objects, e.g., documents or concepts, as a relational database, and as a hierarchical or partially ordered graph comprising domain-relevant entities as nodes. We will focus on the most general formulation of a domain model, described as a selection of concepts (usually terms) judged to be salient within a given collection (whether the collection be a single document, an entire document collection, or a collection of other textual data underlying a domain) and/or relations between these concepts.

Domain models have been developed in a variety of research disciplines and for various different reasons. As a result, numerous (sometimes synonymous) terms have emerged which are all used to refer to the concept of a domain model, such as: *Semantic Network, Ontology, Concept Map, Conceptual Graph, Term Association Graph, Taxonomy*, etc. While these names have been created to convey slightly different notions in the literature, they often overlap in their usages and are employed to refer to an underlying homomorphic structure, characterised by the general formulation of selected vocabulary and relations between concepts (usually terms) in the vocabulary. Note that the notion of "concept" varies and there is no consensus across different communities. It is beyond the scope of this paper to go deep into this issue, and we will not adopt any specific definition of "concept". Instead we will use it in an abstract sense.

Domain modelling has long been a key research area in *artificial intelligence (AI)*, in particular in the field of *knowledge representation (KR)* (for example, Quillian [1967], Woods [1975]). Cyc, a large-scale knowledge representation project aimed at conceptually capturing *Common Sense Knowledge*, goes back to 1984 and the work on this project is still ongoing, for

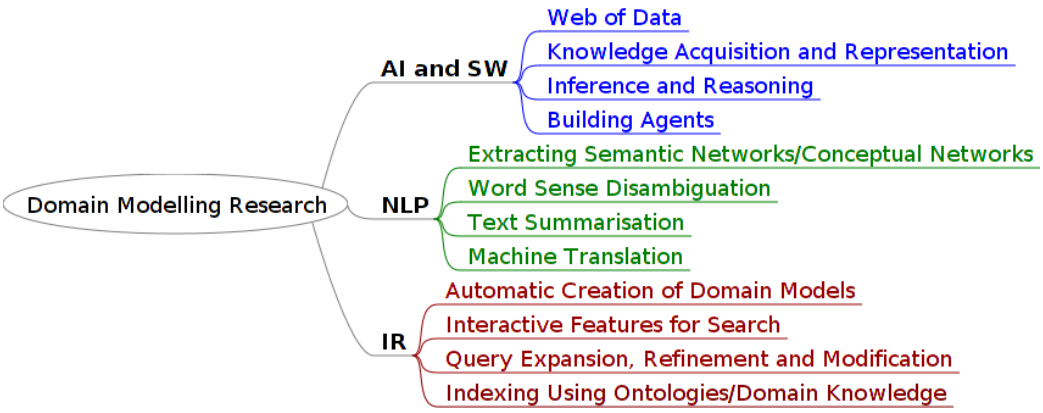

Figure 1: Examples of Domain Modelling in Various Disciplines

example, OpenCyc[1] and ResearchCyc[2] (Lenat et al. [1985]). Domain models aim not only to provide a valid and meaningful representation of the world, but also to facilitate reasoning and inference. The development of domain models further evolved with the emergence of the *semantic web (SW)*, and as part of the ongoing research in *information retrieval (IR)* and *natural language processing (NLP)* applications. Figure 1 illustrates a number of application areas where domain modelling is being employed in these three major research disciplines, i.e. AI/SW, IR and NLP. A more detailed discussion will be given in Section 2

Some of the differences and similarities in the various approaches can be illustrated through the example shown in Figure 2. This provides a small snapshot of a domain model that has been built from a text collection. As a simple term association graph/network, it shows the links between different terms that are in some way related but the relations between the terms are not formally specified. If, however, the nodes in the model were treated as concepts and the specific types of relations between these concepts were identified, this model could be used to develop *Conceptual Graphs*, a *Semantic Network* or part of an *Ontology*. For example, with reference to the Cyc project, in OpenCyc the concept, "Mozart" is of type *Individual*. This concept includes a number of aliases referring to the same individual ("Wolfgang A. Mozart" and "Wolfgang Amadeus Mozart"). This particu-

[1]http://opencyc.org/
[2]http://research.cyc.com/

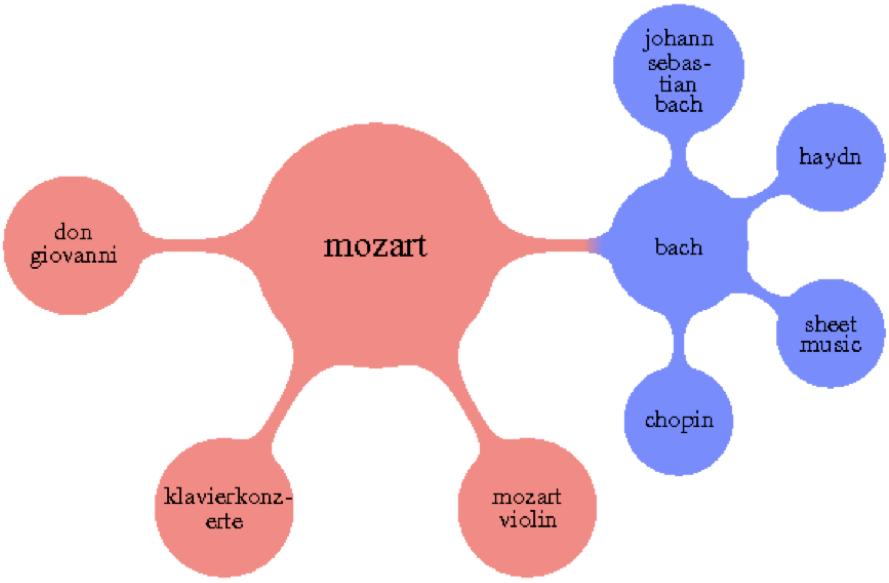

Figure 2: Partial Domain Model.

lar concept is of types *classical music performer*, *Austrian*, *composer*, etc. Adding specific interlinking relations, such as the information that Mozart composed the opera "Don Giovanni" would represent a move towards developing the original simple term association network into a broader, more semantically-enriched structure. A different approach would involve representing only hierarchical relations between terms, for example, the fact that Mozart is a composer, a composer is a musician, a musician is an artist, etc. This would result in a different, somewhat simpler knowledge structure. In fact, for this example, these are exactly the relations that can be found in WordNet[3] (Fellbaum [1998]), a large-scale *lexical* knowledge base.

Domain models are, of course, built and used for different purposes. Cyc is an AI project that encodes knowledge which can be used, for example, in automatic reasoning. It is thus very closely related to the idea of the SW which is aimed at bringing *"structure to the meaningful content of Web*

[3]http://wordnet.princeton.edu/

pages, creating an environment where software agents roaming from page to page can readily carry out sophisticated tasks for users." (Berners-Lee et al. [2001]). The NLP and IR communities on the other hand have very different priorities and research questions. *WordNet*, for example, conceptualises knowledge about the English language which can be applied in NLP, e.g., to disambiguate word senses (Navigli [2009b]). NLP techniques have also been used to extract semantic networks or conceptual networks for text summarisation (Lin and Hovy [2000]) and adapting general lexicons to specific domains (Widdows and Dorow [2002]). In IR, domain models have been extracted automatically from text collections and query log files within a search engine, to suggest query expansion and modification terms. The incorporation of such models as a means of visualising a domain for navigational support is an area of growing importance, clearly reflected in the fact that the prominent search engines have started introducing more and more such interactive features, for example, Google's Wonderwheel[4]. This is not restricted to Web searches, and the success of Aquabrowser[5] as a tool enabling broader exploration of digital libraries using a network of related terms is further evidence of this trend. The partial domain model shown in Figure 2 is an IR example. This model was actually extracted from the query log files that collect user interactions with a library catalogue search engine. It has been built automatically to suggest query expansion and modification terms in an IR context.

Automatic domain model acquisition typically relies heavily on a variety of NLP steps that turn plain text into structured knowledge. We will look at this in more detail. This paper will also examine the various approaches that have been employed towards making the automatically-acquired models adaptive, able to update, improve and change automatically. Adaptive models are unlike static (traditional AI-style) knowledge sources such as WordNet or Cyc. The advances of automatic construction and adaptation of domain models are addressing the so-called knowledge acquisition bottleneck (KAB), including problems such as acquisition latency, knowledge inaccuracy and maintenance of the acquired knowledge (Cullen and Bryman [1988], Tang et al. [1994], Wagner [2006]). To break through the KAB, various research communities have been seeking more effective solutions to the automatic

[4]http://www.googlewonderwheel.com/

[5]http://www.serialssolutions.com/aquabrowser/

construction and adaptation of domain models.

Domain model acquisition seeks to learn a model from data, and one way of categorising them is by looking at the approach it takes to learning and what kind of data it takes to learn them. The overall aim of this paper is to draw contrasts between different approaches and point to trade-offs and some recent trends. The rest of this paper is structured as follows. In Section 2, we first take a general look at various relevant research disciplines within the context of their attempts to create domain models and the ways in which these influence the types of concepts and relationships they include in their models. Section 3 gives details of the learning algorithms that are commonly employed in automatic domain model construction. We distinguish unsupervised, weakly supervised and supervised approaches. Section 4 will discuss domain model construction approaches that make use of existing knowledge sources. Assessing the quality and usefulness of automatically acquired domain knowledge is also a difficult task. Section 5 looks into this issue in detail. The final section of the paper offers some concluding remarks and observations.

2. Mapping the Landscape

Several research communities have shown an interest in the field of domain modelling. To provide a constructive reference point for this discussion, we have focused on three research streams: (1) AI and SW technology, (2) NLP, and (3) IR. There has been surprisingly little overlap between these communities despite the range of shared interests.

It should be pointed out that this categorisation is not intended to be definitive. For example, it could be argued that AI and SW deserve to be treated as two separate areas, whereas in other cases the borders are not so clear-cut. For example, work in information extraction inherits from both NLP and IR. Furthermore, we assume NLP to be an umbrella term that also includes the areas of computational linguistics, human language technology and natural language engineering.

The simplified categorisation into three fields is intended to help demonstrate the spectrum of characteristics that arise as a consequence of the particular vision within different research areas. These research communities can often be characterised by the types of concepts and relationships between concepts in which they tend to be interested, and this seems to be heavily influenced by the over-arching objectives within each of these communities.

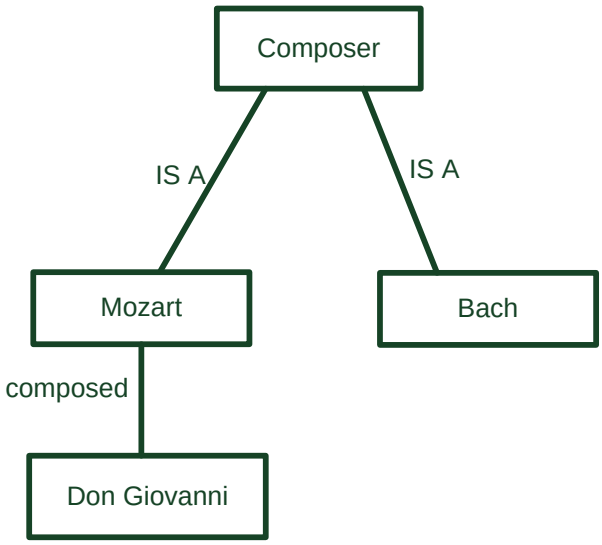

Figure 3: Partial Ontology Example.

To illustrate the different visions and representations of models in different disciplines, consider the partial models in Figures 2 and 3. The first model (Figure 2) is a simple term association network in which nodes, that is, the model concepts, represent query terms and the relations between these nodes are not defined. As mentioned earlier, these models are very common in the IR community. The second model (Figure 3) is part of an ontology in which the nodes refer to entities and the relationships between these entities are semantically defined (*Mozart* composed *Don Giovanni*). In the SW and AI communities such knowledge representation is necessary to allow automatic reasoning and enable Web agents to understand the content on the Web.

In this paper, we distinguish two main paradigms of building domain models: *data-driven* and *knowledge-driven* approaches. More generally speaking, these could be referred to as statistical and symbolic models.

The data-driven approaches are defined by the emphasis they place on extracting key words or phrases that capture concepts. The relationships included in a data-driven model tend to vary widely in type and granularity reflecting only a loose notion of *relatedness* based on the topic of the text. Some approaches do not attempt to generate relationships at all while oth-

ers generate relationships between concepts based on degrees of specificity and subsumption. The relationships are often extracted using co-occurrence frequency within the collection or using inferred attributes of the concepts.

The *knowledge-driven* approaches, on the other hand, tend to target specific types of relationships (such as hyponymy, meronymy and synonymy) that are defined a priori to the extraction process. Entities with the corresponding relationships are extracted based on the specified types. For this purpose, lexical databases such as WordNet are widely used. The integration of a manually engineered knowledge source into the process introduces more control over the relationships extracted, but may not be able to cover a sufficient number of domain-specific concepts, which can affect the adaptability of the framework to very specialised domains.

Data-driven domain modelling approaches have been widely used in IR and NLP, which are examples of research fields that have seen a shift from mainly symbolic ideas to a strong preference for the statistical approaches. AI, and SW technology, on the other hand, is an example where the knowledge-driven approach is more prominent.

We will now look at these disciplines in more detail.

2.1. Artificial Intelligence and the Semantic Web

AI researchers have always been interested in representing knowledge in such a way that it can be utilised by automatic reasoning systems (Sowa [2008]). We can see the idea of the SW as a natural extension to this long tradition.

The main objective of the SW lies in extending the Web to include content currently outside the immediate scope of linked pages, to enable agents to use this content in a variety of applications across different platforms (Berners-Lee et al. [2001]). As such, creating common formats and links between databases and their content is at the core of their many tasks. Consequently, domain knowledge representation together with the extraction of fine-grained metadata to describe content form one of the many important areas of research within the SW community. In particular, the ability to extract formal terminology and identify various types of semantic relationships between the terms (a.k.a. *ontology*) from unstructured text is considered to be of critical importance (Navigli and Velardi [2008], Buitelaar et al. [2005]). The

PASCAL Ontology Learning Challenge[6] was an example initiative aiming to address this issue (e.g., (Dagan et al. [2005], Giampiccolo et al. [2008]).

At the heart of the semantic web is the desire to enable different applications to *understand* and use the same data. This drives domain concepts and relationships between concepts be defined as explicitly as possible. The concepts are often required to express the same level of detail that would be found in a relational database comprising abstractions of inclusion, aggregation and association. This encourages domain models developed within this community to have a strong foundation in knowledge-driven approaches. Such knowledge is frequently specified using a machine readable description language format (e.g. Resource Description Framework[7] (RDF)) and a machine readable knowledge representation language (e.g. Web Ontology Language[8] (OWL); and the DARPA Agent Markup Language[9] (DAML) plus Ontology Inference Layer[10] (OIL)) to enable web-based applications in communicating across different domains.

The reliance of applications on well-designed data structure leads the research in this community to be largely dominated by semi-automatic and manual approaches (e.g., Flouris et al. [2008], Maedche et al. [2003]).

Some tools, however, such as the Karlsruhe Ontology (KAON) framework[11] and OntoLearn (Navigli et al. [2004]), actively support language processing for automatically extracting and selecting keywords representative of domain concepts from natural language texts.

Understanding and extracting knowledge from data requires a fine-grained representation of the semantic relationships between entities found within the text. The research in AI and the SW tends to reflect this by focusing heavily on knowledge-driven approaches to domain modelling. Data-driven approaches do, however, also find their way into this area, primarily those that extract relations using NLP methods (Wilks and Brewster [2006]).

[6] http://olc.ijs.si/
[7] http://www.w3.org/TR/PR-rdf-syntax/
[8] http://www.w3.org/TR/owl-guide/
[9] http://www.daml.org/
[10] http://www.ontoknowledge.org/oil/
[11] http://kaon.semanticweb.org/

2.2. Natural Language Processing

Research in NLP holds the basic standpoint that relationships between words are important both to capture in domain models and for NLP applications. Hence, researchers in NLP have shared a long-standing interest in constructing domain models or semantic networks to characterise textual structure, to find terms related to each other (e.g., Ceccato [1961], Doyle [1961], Phillips [1985], Hearst [1992], Widdows and Dorow [2002], Pantel and Lin [2002] and Kozareva and Hovy [2010]). A thorough overview of NLP approaches to the construction of conceptual networks can be found in Widdows [2004].

A typical data-driven example that illustrates the difference to the AI and SW approaches is introduced in Widdows and Dorow [2002]. The algorithm can be used for *"assembling semantic knowledge for any domain or application"*, is based on grammatical relationships such as co-occurrence of nouns or noun phrases, and needs only a corpus tagged for part-of-speech. The underlying motivation is the extraction of term relationships that do not need to strictly follow fully specified semantic relations but which can, for example, be used for query modification in a search context. In other words, the underlying idea is *"to observe word meanings with no prior agenda: to hear the corpus speak with its own voice"* (Widdows et al. [2002]).

One example NLP area that profits from the extraction of conceptual graphs from textual documents is word sense disambiguation (Navigli [2009b]). It is often an objective in itself in natural language processing, but at the same time it is an essential component in a variety of applications (for example, in question-answering). Remarkably, large-scale conceptual networks have been applied and evaluated in the literature as part of the word sense disambiguation and induction tasks (e.g., Navigli and Lapata [2010], Cuadros and Rigau [2006], Navigli [2009b], Widdows and Dorow [2002], Pantel and Lin [2002]). NLP techniques have also been used to extract semantic networks (Mintz et al. [2009], Snow et al. [2006], Richardson et al. [1998]), for example, for text summarisation (Lin and Hovy [2000]) and adapting general lexicons to specific domains (Toumouth et al. [2006], Widdows and Dorow [2002]), and so on.

While the main paradigm for current research appears to be data-driven, the emergence of powerful NLP toolkits such as GATE[12] has been a signifi-

[12]http://gate.ac.uk/

cant development not just in the area of NLP but also because they offer ways of bridging different areas (such as NLP and SW) by combining data-driven and knowledge-driven approaches within the same framework.

2.3. Information Retrieval

The domain modelling research in the IR community aims to build systems that assist users in retrieving information through information spaces. In contrast to NLP, relationships between words are lost altogether when simply looking at frequency analysis. The common IR scenario takes the form of a user submitting queries, formulated as a number of keywords, to a search engine that is expected to return relevant information from a collection, normally an indexed collection of documents, by computing a numeric score based on the original query. This experimental setup referred to as the "Cranfield Paradigm" (Cleverdon [1960]) offered a formalised methodology for pre-existing retrieval researchers to evaluate an IR system against a test collection of documents.

However, the search process is becoming more complex and interactive than the traditional IR evaluation discussed above, because of the extension of the IR paradigm as the research on human behaviour during the user's interactive information seeking, browsing and navigation expands (Case [2007], Wilson [1999], Golovchinsky et al. [2009], Marchionini and White [2009]). We have already mentioned the interactive features introduced by standard Web search engines, but faceted searching has also become popular in recent years (Ben-Yitzhak et al. [2008]). This asks for knowledge structures that can assist a user in the search process.

As explicitly engineered ontologies, semantic networks and document annotation, appropriate for selected domains, are often unavailable and expensive to create, automatically created domain models from textual documents are increasingly attracting interest within the IR community not least because one of the features being language independence. Some efforts have already been made to use these in query expansion, reformulations and suggestions (Sanderson and Croft [1999], Kruschwitz [2003]), Lau et al. [2008]), as well as filtering information (for example, Nanas and de Roeck [2009]).

The relationships between domain concepts are recognised as being important within IR. For example, networks of hyponym/hypernym relations, and other forms of relatedness have been used to expand, refine, and modify queries and to score document relevance to the given query (for example, Hovy et al. [2009], Grefenstette [1992], Sanderson and Croft [1999], Gürkök

et al. [2008], Nanas [2003]). However, for IR, the emphasis lies in capturing a set of terms that are *closely related* to each other as co-occurring within the same context (whether topical or semantic). As such, the research tends to focus more on clustering concepts for word discrimination rather than on distinguishing between each relationship explicitly to achieve word disambiguation (cf. discussion in Schütze [1998]).

Like in NLP, the data-driven approach appears to be dominant in IR. However, it is not surprising that some researchers have also turned to formal ontologies that capture domain knowledge (Navigli and Velardi [2003], Hsu et al. [2006]) for query expansion, and have exploited semantic relations from selected texts or queries (for example, van der Plas and Tiedemann [2008], Hollink et al. [2007]) for question-answering and/or query modification. Vallet et al. [2005], for example, exploited knowledge bases by creating an ontology-based scheme for the semi-automatic annotation of documents and the creation of an IR system using an annotation-weighting and ranking algorithm.

2.4. Summarising Remarks

To summarise, different research communities have shown a continuing interest in domain modelling, but it appears that there has been little overlap between different disciplines. From the discussion in this section, it is clear that the approaches adopted by different communities can be complementary to each other.

This section has provided a flavour of the different communities and the respective domain modelling approaches. We will now look in a bit more detail into data-driven approaches. These approaches range from unsupervised to supervised techniques, each of which with their own strengths and weaknesses. Once we have discussed data-driven approaches we will turn to knowledge-driven methods in Section 4.

3. Data-Driven Domain Modelling

In this section we will describe the mainstream data-driven algorithms that have been employed in the context of automatic domain model construction. The algorithmic approaches can be divided roughly into three strands: unsupervised learning (cf. Section 3.1), weakly supervised learning (cf. Section 3.2), and supervised learning (cf. Section 3.3). The advantage of unsupervised methods is that little human labour is required to produce

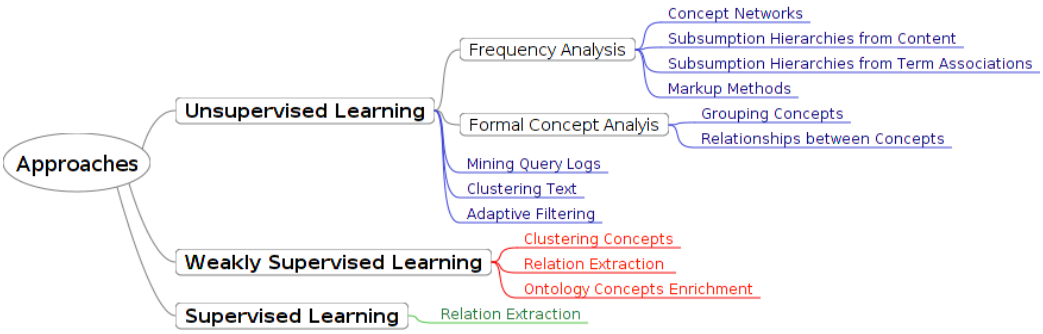

Figure 4: Data-Driven Domain Modelling.

well annotated training data, but the main drawback is the difficulty in producing annotations of explicit concept classes and relationships. Weakly supervised learning methods of domain modelling require some manual effort, for example, to identify seed patterns, templates or specific concepts and relationships. Supervised learning methods on the other hand typically require substantial human annotation effort but the main advantage is that the annotation of training data can be of high quality and specific to the domain. The obvious disadvantage of this approach is that such annotated data is often unavailable and expensive to create.

Figure 4 lists some typical example methods for each of the strands. This section will give an overview of these approaches.

3.1. Unsupervised Learning Methods

Unsupervised learning takes raw data to learn a model, it therefore requires no prior annotation effort (e.g. to classify input into a number of different categories). This is a very active research area and we will distinguish two types of input for the domain modelling step, first of all actual text as found in documents (cf. Section 3.1.1) and secondly implicit data such as query logs, relevance feedback information etc. which contribute to a quickly growing area of domain modelling (cf. Section 3.1.2).

3.1.1. Unsupervised Learning from Text

Frequency analysis has long been employed in text processing. In particular, the extraction of *concepts* from text followed by an analysis of co-occurrence statistics (that is the counts of two concepts occurring within close proximity within selected text) as an approach to information search

and seeking was already being mentioned by Doyle [1961] at a time when computational resources were limited.

Phillips [1985] used the study of co-occurrence to build what he called *conceptual structures* and *syntagmatic lexical networks* from different types of books e.g. science books. Words found in collocations with content words were extracted and clustered. A network of stemmed concept words was produced for each chapter, and the macro structures for the whole volumes were inferred by examining the extent of overlap between selected networks.

Schütze [1998] took context analysis to a more formal level. He mapped each occurrence of an ambiguous word w to a high-dimensional word space using collocated words and their co-occurrence frequency. He clustered them using the EM algorithm, initialised by group average agglomerative clustering on a random sample. Singular value decomposition was also used to identify the major axes of variation. Sanderson and Croft [1999] took a highly *query-centric* approach. In contrast to the clusters of the previous methods which did not attempt to label relationships between terms or concepts, they introduced a hierarchical relationship, imposing a subsumption relation between concepts extracted from top matching documents retrieved for a given query.

Lau et al. [2007] also followed an extraction process (later applied to the e-learning task (Lau et al. [2009]) similar to that of Schütze [1998] and Sanderson and Croft [1999]. They processed a corpus with stopword removal, part-of-speech tagging, stemming, linguistic patterns selection (for example, patterns such as *noun-noun* or *adjective-noun*), and statistical analysis for concept extraction. They further used information theory measures such as *mutual information* and *balanced mutual information* and term frequency within selected domains to refine the selection of concepts that represent domain concepts. Fuzzy subsumption relations were derived from term associations. The resultant domain ontology was further smoothed by including concepts from WordNet. This is different from the approach in (Sanderson and Croft [1999]) where term relationships were derived based on frequency counts of retrieved document passages.

Rather than taking unstructured text as input for the domain modelling process one could also make use of existing markup structure within the text to guide the process. Web document structure as represented by hypertext markup language (HTML), extensible hypertext markup language (XHTML) or extensible markup language (XML) has been exploited in conjunction with frequency analysis. For example, Kruschwitz [2003] used the count of different structural contexts as a guide for extracting concepts and subsequently

building a domain model based on these concepts whereas Brunzel [2008] used XHTML tag paths for text (that is, the Web page markup that leads to a given piece of text) as context for finding synonyms while Shinzato and Torisawa [2004] used list itemisation for locating hyponyms.

Formal concept analysis (FCA) focuses on building a lattice derived from concepts as defined by a set of attributes (Cimiano et al. [2005]). It creates a one-to-one mapping between groups of similar concepts and a set of attributes so that attribute inheritance from concept group C_1 to concept group C_2 determines a partially ordered relationship similar to subsumption. Some of the research discussed later on (Hattori and Tanaka [2008], Poesio and Almuhareb [2008], Paşca and Alfonseca [2009]) is closely related to FCA, in that they place focus on the attribute sets of concepts as determining relationship between concepts.

In addition to the approaches outlined above, there is also a strand of research that focuses on grouping texts, so that examples found within the clusters are more closely related to each other than those outside the cluster. For example, Chuang and Chien [2005] grouped short text segments from top search results using agglomerative clustering to create a hierarchical tree of text clusters. Also included in this line of research are the suffix tree methods of text clustering discussed by Zamir and Etzioni [1998], Branson and Greenberg [2002], Chim and Deng [2007], and Crabtree et al. [2005]. Zhang and Wu [2008] used topical clustering as a visualisation technique for digital libraries. Self-organising maps have also been applied to enrich the relationships between concepts (Dittenbach et al. [2004], Chen et al. [2008]).

3.1.2. Unsupervised Learning from Implicit Data

Unsupervised learning methods that do not exploit the actual text documents but instead make use of search log files, click data, implicit relevance feedback etc. have emerged over recent years. In particular, graph-based domain models incorporating user search behaviour by examining query and click logs have started to appear more and more frequently in the literature in recent years (Baeza-Yates [2007], Baeza-Yates and Tiberi [2007], Boldi et al. [2008]). Similarly, bipartite graphs that include both queries and URLs as nodes can be used to identify a domain model of closely related terminology, that is, phrases that have resulted in the retrieval of the same documents (Deng et al. [2009], Craswell and Szummer [2007]). This research has developed into an entire research area of Web data mining. The premise of much of the work is that queries and documents selected by users constitute concept

terms and term sources preferred by the user community of the underlying collection.

Much research is aimed at exploiting implicit feedback in one way or another and much of it derives from the concept of relative relevance where user clicks are not treated as *relevant per se*; but instead clicked links are seen as *more relevant* than other links that have not been clicked (Radlinski and Joachims [2005]). This has sparked a lot of further research in recent years.

Implicit feedback has also been used to extract conceptual structures, which *"expresses declarative knowledge by implementing it as a connected multilabeled bipartite oriented graph"* (Sowa [1984]). For example, Lungley and Kruschwitz [2009], built a domain model on collection-wide formal concept analysis followed by an adaptation process to reflect implicit feedback inferred from user-clicked documents.

Lau et al. [2008] examined concept relations adapted as part of a belief revision framework incorporating document relevance feedback. Their findings showed that the belief-based system was as effective as a classical adaptive IR system. Some approaches in domain modelling have emerged in the context of user's viewpoint, for example, concept hierarchies as user profiles, subsequently adapted using immune system inspired approaches (Nanas [2003], Cayzer and Aickelin [2005]). User models can be seen as a special type of domain model that reflects an individual user's or a group of users' view on the search domain. Nanas et al. [2010] discussed other user viewpoint based domain modelling methods that use genetic algorithms, clonal selection algorithms, negative selection, co-stimulation, and immune inspired self-organising networks. They used documents judged relevant by users to construct and update a domain concept model. Terms in the query and documents are linked to nodes in the concept model network. An initial level of energy is disseminated through the query nodes, then distributed through the network, and, finally, accumulated as a document energy or relevance score. When new user relevance feedback becomes available the network is updated by a similar process of energy distribution.

More user-centric methods were suggested by Paşca and Alfonseca [2009], where query logs were analysed to derive likely attributes for identified objects in order to refine the concepts in the model with associated attribute hierarchies. Query and document history have been used to model short-term and long-term user interests, in the form of domain models. This research defines a topical similarity measure so that if the topical similarity of user

interest context changes at the point of any query submission, a new user interest is constructed or the previous interest model is revised.

Unsupervised methods using implicit data can also be applied to *first* building a domain model and *subsequently* adapting the model in an ongoing adaptation cycle. Examples include adapting domain models based on user feedback on the relevance of documents (that is, no explicit judgement on the domain model itself) (Nanas [2003]) as well as adaptive domain models that learn from user query modifications in interactive search (Kruschwitz et al. [2011]).

3.2. Weakly Supervised Learning Methods

Unlike in unsupervised learning, in weakly supervised learning some annotation is required, e.g. phrase and/or syntactical patterns are identified empirically and used on a large textual corpora to harvest entities satisfying the pattern.

Hearst [1992] used this approach to build a network of hyponyms from text. In this study, for example, phrases "A such as B" and "A, especially B" were used to establish B as a hyponym of A. Grefenstette [1992] extended methods based on lexical patterns by quantifying the similarity of syntactic dependencies (for example, modifiers) associated to a word, to cluster similar words together. Grefenstette [1992], however, did produce explicit relationship tags for his clusters.

Others, such as Thelen and Riloff [2002] and Snow et al. [2005], took this further, using entities already identified as being in a semantic class, or taking pairs of entities identified within WordNet as being in a hypernym/hyponym relation, as seeds for the identification of new phrasal patterns and entities belonging to that category or relationship. Here, it might be needed to first extract a pool of patterns that are likely to extract the seed entities or relationships. The pool of patterns is used to extract candidate entities and those candidates that are associated with patterns most likely to extract the seeds are added to the network.

Some researchers formulated document-template pairs to induce pattern matching rules (Califf and Mooney [2003]). The patterns induced, within this framework, can have constraints not only on surface patterns such as lexicon and part-of-speech, but also constraints on the semantic classes of the words in the pattern. Morin and Jacquemin [2004] inferred multi-word variants from single word hypernym relations based on the lexical patterns of the single word hypernym network.

Approaches that harvest concepts based on relational patterns tend to extract concepts across many domains. Hovy et al. [2009] tried to better define the concept domain by examining the network produced as hyponyms of one seed term. On the other hand, Valarakos et al. [2004] developed a semi-automated ontology enhancement workflow that starts with a seed domain ontology used to annotate a domain corpus, and extract and cluster further candidates for inclusion in the ontology. The candidates are examined by a domain expert for final quality control.

Pantel and Pennacchiotti [2006] induced generic patterns to retrieve a wide range of concept pairs and then made use of a large sampling space such as the Web to filter the results to retain those associated to high precision patterns. They used, for example, "A of B" as a pattern for meronymy (*part-of* relation).

Hattori and Tanaka [2008] looked at property inheritance and aggregation as a means of hierarchical knowledge organisation from the Web. They used two types of lexical patterns (for example, patterns such as "X's Y" as an instance of "an attribute Y of a concept X") to harvest, first, a set of candidate hyponyms in relation to a given concept, and, second, a set of properties for each target concept. The weight of each candidate as a hyponym would be weighted on the basis of how many of the root concept's properties it inherits. Poesio and Almuhareb [2008] also discussed the importance of concept attributes and their values in extracting conceptual knowledge. They used lexical patterns as well as dependency parsers, to extract concept descriptions from the Web.

Another stream of methods that goes under this category is the extraction of arbitrary relations from text between named entities in the form of subject-predicate-object triplets. For example, the REXTOR system in (Katz and Lin [2000]) used a finite state language model to extract what they call *ternary expressions* that describe relations between entities. They argued that these structures are simple to extract and serve as a powerful tool for bridging the gap between NLP and IR, as they were able to cover a wide variety of relation types. There is a wealth of related work, often using Wikipedia. For example, Akbik and Broß [2009] used dependency link grammars to identify all the link paths that result in valid relationships. These paths were used to extract semantic relations from plain text in a subject-predicate-object triplet form, analogous to statements in RDF between entities (also known as *resources* in RDF) from Wikipedia articles.

3.3. Supervised Learning Methods

In the supervised approach, pre-labelled training examples are required. For example, lexical and syntactic association patterns have been used as features in a general learning algorithm (such as Support Vector Machines and Conditional Random Fields), which usually learns classifiers from a collection of pre-annotated relations.

Most of the supervised learning algorithms use varying levels of syntactic information, ranging from part-of-speech tagging to full parsing and, in some cases, additional information, such as named entity tagging (Mintz et al. [2009]).

The research presented in Girju et al. [2006], in particular, used supervised learning methods to determine whether the *part-whole* relation candidates retrieved using lexical patterns, constitute a true example of the relationship. The classification builds rules regarding noun phrase constituents (for example, regarding prepositional phrases in the noun phrase compound) to iteratively learn semantic specialisation instances.

Snow et al. [2006]'s algorithm incorporated evidence from multiple classifiers over diverse relationships to optimise the entire structure of a model. They used the algorithm to merge the predictions of coordinated term classifiers to add hypernymy to a pre-existing semantic taxonomy.

Tree kernel methods, a class of pattern analysis algorithms that can detect types of data and general types of relations, have also been suggested (Reichartz et al. [2009], Culotta and Sorencen [2004]) to learn the association patterns from the phrase grammar parse tree and dependency parse tree of the sentences containing the relationship to detect new instances. In addition, Giuliano et al. [2007] has used kernel functions on parse trees to learn relationships between named entities.

Mintz et al. [2009] presented *distant* supervised learning using sentences extracted from the Freebase[13] Wikipedia Extraction. This source is already seeded with a large database of relationships and instances extracted from Freebase itself. Distant supervised learning is an alternate extension of the paradigm to that introduced by [Snow et al., 2005] with the purpose of merging some of the positive aspects of supervised and unsupervised learning. In the case of Mintz et al. [2009] it was to extricate Hypernym (is-a) relationship pairs between entities for successful sentence extraction.

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

All supervised learning approaches depend on training data which does not always exist. This raises problems when trying to apply them to a specialised domain where large sets of data for training may not be available, and specialised concepts, which are not annotated elsewhere, might arise. Furthermore, the deeper level of linguistic features involved in these approaches brings the scalability of these methods into question for large data collections in interactive environments, such as the Web.

3.4. Summarising Remarks

The major advantage of unsupervised methods for domain model creation is that little human labour is required to produce well annotated training data. This is significant not only in terms of the costs in time and money involved, but also in terms of the methods' applicability to any data and domain. The main problem with regard to the unsupervised approaches is the difficulty in producing annotation of explicit concept classes and relationships. This is a major disadvantage with respect to the machine readability of the model by applications where this is paramount (for example, in SW and linguistic analysis). Unsupervised methods can benefit from approaches to adaptation that might elevate the model to a more rigorous standard, not only in terms of explicit annotation of concepts and relationships, but also in terms of consistency across the concept network (for example, with respect to types of relationships between siblings and between parents and siblings that populate the network). Unsupervised learning can be applied to the actual textual data sources to build a domain model or to implicit data sources such as query logs and click information associated with text collections.

The advantage of weakly supervised learning methods of domain modelling is that explicit concepts and relationships become available through targeted harvest. At the same time, the method does not require extensive manual annotation of training data. Depending on the context and application available, this might offer the best of both worlds, but the relationship classes that are covered within this framework tend to be narrow. While some efforts are being undertaken to broaden the coverage, these tend to support general semantic networks that are not optimised to assist users within focused domains or applications. Weakly supervised learning methods could benefit from research oriented towards broadening the coverage of relationship types in a way (for example, active learning) that actively selects new relationship types with respect to a selected application or domain.

The advantage of supervised learning methods is that the annotation of training data can be carried out in such a way that all relationships and concepts are selected to meet the needs of the selected domain or application. The annotated data may also serve as the gold standard against which any automatically constructed models can be compared. The obvious disadvantage of this approach is that such annotated data is often unavailable and expensive to create. It also often relies on a black and white scenario where experts agree completely on the important concepts and relationships of the domain. This might result in models that are not easily open to adaptation and evolution. Supervised learning methods for the construction of domain models might benefit from research directions that incorporate information from the interaction of users with the model in an application environment.

4. Knowledge-Driven Domain Modelling

The main data-driven approaches described in Section 3 focused on building domain models from scratch, simply using a collection of text or implicit query log and relevance feedback data. However, a range of external knowledge sources can also be used to build and enrich domain models. These knowledge sources can be fully structured (such as WordNet) or semi-structured (such as Wikipedia). We divide this research into two main strands according to the resources incorporated into the framework: those using explicit knowledge sources (cf. Section 4.1) and those that enrich existing domain models (cf. Section 4.2).

4.1. Using Explicit Knowledge Sources

A number of knowledge sources have been used to build domain models (among other things). Some of the most commonly used resources are large-scale, freely available and of high quality. The works described here are designed to assist general applications (e.g. word sense disambiguation), the creation of large scale knowledge bases (e.g. YAGO[14]), and the extension of general lexicons (e.g. WordNet) with semantic relations (e.g., Navigli [2009a], Pennacchiotti and Pantel [2006]).

WordNet is a popular knowledge source that has been used extensively in research in many different ways, primarily because it is a substantial linguistic

[14]http://www.mpi-inf.mpg.de/yago-naga/yago/

knowledge source of high quality and is freely available. Enriching WordNet with additional knowledge is one strand of work. For example, adding "topic signatures" (that is, a list of topically related words, such as *restaurant, menu* in relation to *waiter*) was proposed by Agirre et al. [2000] (and also their later work Agirre et al. [2001], Agirre and de Lacalle [2004]). Each WordNet concept is used to construct Web search queries that retrieve a collection of documents relevant to that concept from the Web. Words with high Chi-square (χ^2) values are selected as topic signatures.

Instead of using the Web as a knowledge source, a controlled vocabulary could be used to enrich WordNet. Longman Dictionary of Contemporary English[15], for example, has been used to locate corresponding representatives in WordNet that serve as good replacements for their descendants (for example, *restaurant* is a representative for *bistro* or *cybercafe*) (Navigli [2005]).

The use of Wikipedia's inherent structure is another growing strand of research. For example, Wikipedia's categories have been used to build a large-scale taxonomy as a conceptual network (Ponzetto and Strube [2007], Ponzetto and Navigli [2009]). A methodology for disambiguating Wikipedia categories with monosemous WordNet synsets was presented. The framework was evaluated using a *manual gold standard* (cf. Section 5) against manually tagged datasets.

Medelyan and Legg [2008] mapped groups from Cyc onto Wikipedia articles describing corresponding concepts. Their method calls on both Wikipedia's rich and sometimes messy hyperlink structure and Cyc's carefully defined taxonomic and common-sense knowledge.

Suchanek et al. [2007] created the knowledge base YAGO which currently contains more than 2 million entities (for example, person, location, and organisation) and 20 million facts about these entities (non-taxonomic relations between entities, such as *hasWonPrize* and *is-A* hierarchy). The facts have been automatically extracted from Wikipedia categories and redirections, in conjunction with WordNet semantic relations, using a carefully planned mix of rule-based/heuristic methods (for example, first concepts are extracted from Wikipedia categories then organised using WordNet hyponym relations to obtain the *subClassOf* relation). The knowledge base, according to the authors, *"is a major step beyond WordNet: in quality by adding knowledge about individuals like persons, organisations, products, etc. with their se-*

[15]http://www.ldoceonline.com/

mantic relationships – and in quantity by increasing the number of facts by more than an order of magnitude."

There exists a number of other large-scale explicit knowledge sources that can be used to build domain models, including commercially available products such as TrueKnowledge[16]; knowledge bases for academic purposes, e.g., Open Mind Common Sense[17] (Singh et al. [2002]), which can be accessed via ConceptNet[18], an open-source, multilingual semantic network (Liu and Singh [2004], Speer et al. [2008]); WikiNet, a large scale multilingual concept network (Nastase et al. [2010]); and BabelNet (Navigli and Ponzetto [2010]), a large scale multilingual semantic network.

DBPedia[19] is another massive database which makes Wikipedia content available as structured knowledge on the Web (Auer et al. [2007]). It uses a variety of vocabularies and knowledge schemas to represent facts between entities including the previously mentioned YAGO ontology and it also links entities and facts to external knowledge resources.

4.2. Enriching Existing Domain Models

The research described in this section aims to enhance knowledge representation within the context of existing domain-specific knowledge structures, by identifying the changes that arise within domain-specific environments and showing how these can be incorporated into a high level knowledge representation and enrichment framework.

Theoretical approaches have been developed that address the question of how new information can be incorporated into an existing domain models. For example, Chen et al. [2008] used the distances of a new term from the concept groups in the model to determine onto which group the new term should be mapped.

One of the domains that relies heavily on conceptual networks is the medical domain. Toumouth et al. [2006] used a fairly simple syntactical pattern to harvest nouns from the Oshumed corpus[20], which were then organised according to their common ancestors and the senses (as prescribed by WordNet) most likely to occur within the corpus. Diederich and Balke [2008] used

[16]http://www.trueknowledge.com/
[17]http://openmind.media.mit.edu/
[18]http://conceptnet.media.mit.edu/
[19]http://dbpedia.org/
[20]http://ir.ohsu.edu/ohsumed/ohsumed.html

keywords specified in Medline[21] articles to examine high order co-occurrence statistics of the keywords, subsequently mapped to a concept graph.

With regard to the area of ontology enrichment, a number of approaches have been proposed, often semi-automated rather than fully automated. OntoLearn is a semi-automated ontology creation tool which can also be used to automatically enrich a domain ontology by utilising WordNet and other online dictionaries for heuristics (Navigli et al. [2004]). Valarakos et al. [2004] developed a semi-automated ontology enhancement workflow that starts with a seed domain ontology. This is used to annotate a domain corpus, and to extract and cluster further candidates for inclusion in the ontology. The candidates are examined by a domain expert for final quality control. Navigli and Velardi [2006] described a pattern-based method to automatically enrich a core ontology with the definitions of a domain glossary. They applied the method to the cultural heritage domain and used available resources including WordNet and the Dmoz[22] taxonomy for named entities.

Working on a similar strand of research, Monachesi et al. [2009] proposed ontology enrichment with social tags for e-learning. The authors argued that social tagging systems have become a standard application of the Web. These applications can be considered as shared external knowledge structures of users on the Internet. They described how social tagging systems relate to individual semantic memory structures and how social tags affect individual processes of learning and information foraging. Furthermore, they presented an experiment consisting of an online study targeted at the evaluation of the interaction of external and internal structures of spreading activation.

Web logs have also been used in combination with ontologies and folksonomies, for example, Passant [2007], who addressed some of the problems originating from free-tagging classification when applied to information retrieval. The authors combined ontological knowledge on top of an existing folksonomy as a way of dispensing with *free-tagging classification flaws*.

4.3. Summarising Remarks

The types of resources described in this section that have been found to be in use for building, enriching and adapting domain models, reflect the objectives that underpin the research: to produce fine-grained description

[21] http://www.nlm.nih.gov/databases/databases_medline.html
[22] http://www.dmoz.org/

of textual structure, to enhance machine readability, to represent knowledge within a community to facilitate its extraction and re-use, and to assist users to find what they need from a large collection of material. This allows the work to be divided into different research areas based on the specific needs and tasks. Researchers have aimed to build models that can adapt to the selected needs of a user or community, or they have focused on general lexical and semantic knowledge bases (e.g., WordNet) and general knowledge sources (e.g., Wikipedia). Yet others have chosen to gear their work to the needs of a specialist community. The spectrum of needs and tasks that arise within these different groups are, however, merely iconic samples drawn from a continuum of granularities. Users often belong to different groups of specialist communities and will eventually be happy when their needs with respect to these different communities are met within the language and conceptual structures they have been trained to understand. As a next step, to consolidate the diverse array of research described in this section, future research should move into the direction of testing domain modelling approaches within vertically sampled scenarios, that is, a well-defined set of scenarios, each of which incorporates the continuum from users' specific interests.

5. Evaluation of Domain Models

The evaluation of a complex network structure such as a domain concept model is a challenging task. The diverse reasons (for example, the target application) for the development of the model have a direct influence on the way in which the model is evaluated. To some extent, this is reasonable, but this diversity can hinder the development of a commonly accepted evaluation methodology and the failure to establish such a methodology can present difficulties for researchers trying to compare the effectiveness of the different construction approaches available.

Apart from the target application the assumptions about what the network actually models are different, and therefore evaluation methods will also have to be different. For example, ontologies are concerned with the extraction of concepts and relations between them, and typically strip out lexical information from the network. Two ontologies covering the same domain may use different concept and relation designators in different configurations, which makes similarity comparisons difficult. Term-based models, on the other hand, will tend to reflect terminology similarities more closely. As

a consequence, ontology similarity tends to rely on terminological similarity, by comparing the terms associated with concepts and relations.

Overall, it can be said that there are three methods of evaluation: the *qualitative criteria-based user evaluation* carried out by users of the model (relatedness judgement, for example, used in Sanderson and Croft [1999]); *task-based user evaluation* of the model's effectiveness in assisting a given application or task (in IR, for example, Gürkök et al. [2008], Lau et al. [2007], Grefenstette [1992], Nanas [2003], Lawrie and Croft [2000]); and *quantitative evaluation* of the model against a gold standard model (Hovy et al. [2009], Maedche and Staab [2002]). In most cases, several of these approaches are combined (e.g., Lau et al. [2009]).

Evaluation techniques of ontology learning have been examined by a number of researchers. Dellschaft and Staab [2008] presented a comprehensive set of descriptions of the approaches and measures adopted by ontology developers and researchers, while Brank et al. [2005] gave a very concise overview. In contrast, Maedche et al. [2003] focused on quantifying the similarity between two ontologies.

5.1. Qualitative Criteria-based User Evaluation

There are comprehensive accounts of criteria-based evaluation, such as that presented by Chuang and Chien [2005]. The qualitative measures that they identified were:

1. Cohesiveness: used to make a decision on whether the clustered instances are similar in a semantic way.
2. Isolation: utilised to test whether the automatically-generated clusters at the same level are distinguishable and whether their concepts include one another.
3. Hierarchy: used to decide whether the generated topic hierarchy is traversed from broader concepts at the higher levels to narrower concepts at the lower levels.
4. Navigation Balance: used to make a decision on whether the fan-out at each level of the hierarchy is appropriate.
5. Readability: used to decide whether the concepts of clusters at all levels are easy to recognise with the composed clusters and instances.

Although many studies refer to some of these criteria within the framework of a user evaluation, we have been able to find very little research that offers

a thorough user evaluation based on all these criteria. The disadvantage of employing such an approach is the cost in terms of time and labour.

5.2. Task-based User Evaluation

The quality and usefulness of domain models should not only be assessed through the qualitative criteria-based measures discussed above, but also largely depend on the applications in which the domain models are used. Therefore, within the context of the given task, users can be asked to assess the quality of relations encoded in domain models (Sanderson and Croft [1999], Kruschwitz and Al-Bakour [2005]).

The task-based evaluation can also be conducted to evaluate the effect of using the domain model in a given application, e.g., query expansion and document re-ranking (Gürkök et al. [2008], Nanas and de Roeck [2009]), or in disambiguating words and supporting machine translation (Navigli and Velardi [2003]). The standard evaluation methodology and measures for the specific application, e.g., Mean Average Precision (MAP) commonly used in IR applications, can then be adopted to show how much benefit the domain model can bring compared with the baseline without using the domain model.

We would also like to emphasise that, at the time of writing this paper, we had not yet discovered any evaluation methods represented in the literature that try to investigate whether we might be able to infer the usefulness of a domain model through implicit feedback observed from the users' direct interactions with the concept graph. This is surprising since the incorporation of concept graphs as a means of visualising a domain for navigational support seems to be a growing trend (for example, Google Wonderwheel and Yahoo Correlator[23]). Research in this direction would be highly beneficial, not only in helping to overcome the necessity for time-consuming qualitative user evaluations but also in providing a pipeline for automated domain model adaptation.

5.3. Quantitative Evaluation against a Gold Standard

In addition to the user evaluation methods as described above, quantitative evaluation against a gold standard has also been widely adopted. Dellschaft and Staab [2008] illustrated the obvious progress in ontology evaluation that has been made in recent years, but these methods are still not

[23] http://correlator.sandbox.yahoo.net/

widely validated. They show a clear bias towards evaluation by comparison against a gold standard. Their reasoning follows the argument that the cost of building a gold standard is only incurred once and is therefore affordable (e.g., as in Bordag [2006], Dellschaft and Staab [2006], Ponzetto and Navigli [2009]). However, while this may hold true for static domains, it may not hold true for a dynamic environment such as Web-based search scenarios where user interests change rapidly and collections are in constant flux. Even within a fairly static environment, change is inevitable and essential (Flouris et al. [2008]), and it therefore seems vital to have an evaluation method which can reflect the dynamic information environment.

The definitions of all the quantitative measures used for the comparison of ontologies that are presented in this section are also detailed in Dellschaft and Staab [2006] and Dellschaft and Staab [2008], and the equations are sourced from their paper. These measures can be described as one of two types: that focusing on lexical precision and recall (cf. Section 5.3.1), and that focusing on the entire taxonomic similarity (cf. Section 5.3.2). The original IR based definition of the Precision, Recall and F-Measure can be sourced in van Rijsbergen [1979].

5.3.1. Lexical Precision, Recall and F-Measure

Given a gold standard reference taxonomy (*Ref*) and a taxonomy to be compared to the reference taxonomy (*Comp*), we can simply use lexical precision, recall, and F-measure to evaluate *Comp*. Lexical precision measures whether terms (purely on a lexical level) given in *Comp* are actually from the reference taxonomy, and recall measures how completely the terms in the reference taxonomy have been represented within *Comp*. More formally, precision and recall are defined in Equations 1 and 2, respectively. These measures are commonly combined to give an overall harmonic mean (or weighted average) of precision and recall (cf. Equation 3).

$$P(\text{Ref}, \text{Comp}) = \frac{|\text{Ref} \cap \text{Comp}|}{|\text{Comp}|} \tag{1}$$

$$R(\text{Ref}, \text{Comp}) = \frac{|\text{Ref} \cap \text{Comp}|}{|\text{Ref}|} \tag{2}$$

$$F(\text{Ref}, \text{Comp}) = \frac{2 \times P(\text{Ref}, \text{Comp}) \times R(\text{Ref}, \text{Comp})}{P(\text{Ref}, \text{Comp}) + R(\text{Ref} + \text{Comp})} \tag{3}$$

The precision, recall and F-measures above, could be criticised for inadequately reflecting conceptual relationships that may exist between terms. For example, if "car" is returned within *Comp*, and "auto" is within the reference taxonomy, *Comp* would not be rewarded, despite the obvious relationship between "car" and "auto" (Dellschaft and Staab [2008]).

5.3.2. Taxonomic Precision, Recall and F-Measure

Taxonomic Precision (TP) and Recall (TR) are developed to capture the similarity between two concepts even when there is little lexical similarity.

The similarity between two concepts c_1 and c_2 is computed based on the basis of a comparison of *characteristic extracts*, denoted $ce(c_1, O_1)$ and $ce(c_2, O_2)$, from the two conceptual graphs O_1 and O_2 being compared. For example, take the situation described at the end of the last section: in comparing "car" from *Comp* and "auto" from *Ref*, we could take the other terms identified as being related to these terms, that is, "van", "speed", "mileage", as the characteristic extracts to be compared. The premise is that, if the terms are conceptually linked, then there will be a large overlap in the extended extract.

Given a definition for the characteristic extract *ce*, the local taxonomic precision $tp_{ce}(c_1, c_2, O_C, O_R)$ of O_C with respect to concept c_1 and a given concept c_2 from a reference set O_R is defined as:

$$tp_{ce}(c_1, c_2, O_C, O_R) = \frac{|ce(c_1, O_C) \cap ce(c_2, O_R)|}{|ce(c_1, O_C)|} \tag{4}$$

Then we can define the global taxonomic precision of O_C with respect to reference taxonomy O_R to be:

$$tp(O_C, O_R) = \frac{1}{|O_C|} \sum_{c \in O_C} \begin{cases} tp_{ce}(c, c, O_C, O_R) & \text{if } c \in O_R \\ max_{c' \in O_R} tp(c, c', O_C, O_R) & \text{if } c \notin O_R \end{cases}$$

An example of extracting *ce* is the *semantic cotopy*. Semantic cotopy $sc(c, O)$ of concept c with respect to ontology O is defined to be the set of all super-concepts and sub-concepts of c. Semantic cotopy is heavily influenced by the lexical precision (cf. Section 5.3.1). Common semantic cotopy considers only the nodes in the semantic cotopy that are shared by both taxonomies to enhance independence with respect to lexical extraction performance. Some measures try to strengthen the independence by only considering terminology common to both taxonomies so that $tp_{ce}(c, c, O_C, O_R) = 0$ for $c \notin O_R$ or $c \notin O_C$.

Local taxonomic recall is defined using the characteristic extract:

$$tr_{ce}(c_1, c_2, O_C, O_R) = \frac{|ce(c_1, O_C) \cap ce(c_2, O_R)|}{|ce(c_2, O_R)|}. \quad (5)$$

This results in defining global taxonomic recall as the precision of the reference ontology O_R with respect to O_C. Taxonomic F-measure TF can then be defined in exactly the same way as the lexical F-measure, to produce a combined measure. Where TF is not influenced heavily by the lexical level performance, the harmonic mean of lexical recall and TF can be used to produce a second order F' value. In addition, the overlap, $TO(c1, c_2, O_1, O_2)$, between two taxonomies O_1 and O_2 for concepts c_1 and $c2$ (cf. Equation 6) has been suggested (instead of local taxonomic precision) as building blocks for comparing the taxonomies.

$$TO_{ce}(c_1, c_2, O_1, O_2) = \frac{|ce(c_1, O_1) \cap ce(c_2, O_2)|}{|ce(c_1, O_1) \cup ce(c_2, O_2)|} \quad (6)$$

Quantitative evaluations tend to compare systems (Giunchiglia et al. [2009]) across different similarity measures and varying sets of features and rarely involve a comparison of different approaches to construction (for example, a comparison between a hypernym-hyponym ontology developed by employing a concept-centric approach, and the same developed by employing a relationship-centric approach).

6. Concluding remarks

In this paper, we have presented an overview of domain modelling research within three separate disciplines: artificial intelligence and the semantic web (AI/SW), natural language processing (NLP) and information retrieval (IR). We have focused on automated methods for constructing domain models from text collections and knowledge resources. We have also outlined a number of evaluation methodologies that have been employed within the literature. Our main findings can be summarised as follows:

- Domain concept modelling with its roots in traditional AI technology has developed into a heterogeneous research area. Real progress could now be made, particularly in the area of adaptive domain modelling, by exploiting the different strengths of independent efforts in different research disciplines such as SW, NLP and IR.

- The literature reviewed seems to suggest a lack of research addressing the questions of which type of domain model is most suitable for what types of application.
- The evaluation of different domain models as well as different approaches for constructing these models is an ongoing research challenge.
- We see a lot of potential in combining data-driven and knowledge-driven approaches.

In this paper, we have not paid much attention to the effects that interactive user interfaces visualising domain models might have on implicit feedback for domain model adaptation. Most previous research strands have used implicit document relevance feedback within the traditional search interface setting. The scope for further research in this area looks promising, given the growing number of popular search engines that have started employing optional interactive visualisations of term relationships. The traversal of such domain model representations presents an opportunity to log and learn from direct user interaction with the model. For example, positive indicators such as a traversal followed by document selection and a long dwell time could, in future, be used to strengthen links whilst traversals which yield no results could be used to identify poorly performing areas of the model. Extensive research in this direction can also work to improve interfaces for applications other than search, for example, in the context of the domain concept structures increasingly being adopted by traditional libraries that use modern visualisation tools such as Aquabrowser (e.g. Queens Library[24]) as well as those libraries that rely on user tags, such as LibraryThing[25], by assisting users to engage with domain knowledge in an efficient way.

7. Acknowledgements

The research carried out for this review was conducted within the framework of the project, Automatic Adaptation of Knowledge Structures for Assisted Information Seeking (AutoAdapt), funded by the EPSRC grants EP/F035357/1 and EP/F035705/1. We are grateful to the anonymous reviewers who provided us with their invaluable and extensive feedback.

[24]http://aqua.queenslibrary.org/
[25]http://www.librarything.com/tagcloud.php

References

E. Agirre and O. de Lacalle. Publicly available topic signatures for all WordNet nominal senses. In *Proceedings of the 4th International Conference on Languages and Resources and Evaluations (LREC 2004)*, pages 1123–1126, Lisbon, Portugal, 2004. European Language Resources Association (ELRA).

E. Agirre, O. Ansa, E. Hovy, and D. Martinez. Enriching very large ontologies using the WWW. In *Proceedings of the ECAI Workshop on Ontology Learning*, Berlin, 2000.

E. Agirre, O. Ansa, D. Martinez, and E. Hovy. Enriching WordNet concepts with topic signatures. In *Proceedings of the NAACL Workshop on WordNet and Other Lexical Resources: Applications, Extensions and Customizations*, Pittsburgh, 2001.

A. Akbik and J. Broß. Wanderlust: Extracting semantic relations from natural language text using dependency grammar patterns. In *Proceedings of The WWW 2009 Workshop on Semantic Search*, pages 6–15, Madrid, Spain, 2009.

S. Auer, C. Bizer, G. Kobilarov, J. Lehmann, R. Cyganiak, and Z. Ives. Dbpedia: A nucleus for a web of open data. In *Proceedings of the 6th International Semantic Web Conference and 2nd Asian Semantic Web Conference (ISWC 2007 and ASWC 2007)*, pages 722–735, 2007.

R. Baeza-Yates. Graphs from search engine queries. In *Proceedings of the 33rd Conference on Current Trends in Theory and Practice in Computer Science (SOFSEM 2007)*, pages 1–8, 2007.

R. Baeza-Yates and A. Tiberi. Extracting semantic relations from query logs. In *Proceedings of the 13th ACM International Conference on Knowledge Discovery and Data Mining (SIGKDD 2007)*, pages 76–85. ACM, 2007.

O. Ben-Yitzhak, N. Golbandi, N. Har'El, R. Lempel, A. Neumann, S. Ofek-Koifman, D. Sheinwald, E. Shekita, B. Sznajder, and S. Yogev. Beyond basic faceted search. In *Proceedings of the International Conference on Web Search and Web Data Mining (WSDM 2008)*, pages 33–44, New York, NY, USA, 2008. ACM.

T. Berners-Lee, J. Hendler, and O. Lassila. The Semantic Web. *Scientific American*, 5:34–43, 2001.

P. Boldi, F. Bonchi, C. Castillo, D. Donato, A. Gionis, and S. Vigna. The query-flow graph: model and applications. In *Proceedings of the 17th ACM Conference on Information and Knowledge Management*, CIKM 2008, pages 609–618, New York, USA, 2008. ACM.

S. Bordag. Word sense induction: Triplet-based clustering and automatic evaluation. In *Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2006)*, pages 137–144, Trento, Italy, 2006.

J. Brank, M. Grobelnik, and D. Mladenic. A survey of ontology evaluation techniques. In *Proceedings of the Conference on Data Mining and Data Warehouses (SiKDD 2005)*, pages 166–170, 2005.

S. Branson and A. Greenberg. Clustering web search results using suffix tree methods. Technical Report CS276A, Stanford University, 2002.

M. Brunzel. The XTREEM methods for ontology learning from web documents. In *Proceedings of the 2008 Conference on Ontology Learning and Population: Bridging the Gap between Text and Knowledge*, pages 3–26, Amsterdam, The Netherlands, 2008. IOS Press.

P. Buitelaar, P. Cimiano, and B. Magnini. *Ontology Learning from Text: Methods, Evaluation and Applications*, volume 123 of *Frontiers in Artificial Intelligence and Applications Series*. IOS Press, 2005.

M. Califf and R. Mooney. Bottom-up relational learning of pattern matching rules for information extraction. *The Journal of Machine Learning Research*, 4(2):177–210, 2003.

D.O. Case. *Looking for information: A survey of research on information seeking, needs, and behavior*. Elsevier/Academic Press, Boston, USA, 2007.

S. Cayzer and U. Aickelin. A recommender system based on idiotypic artificial immune networks. *Journal of Mathematical Modelling and Algorithms*, 4:181–198, 2005.

S. Ceccato. *Linguistic Analysis and Programming for Mechanical Translation (Mechanical Translation and Thought)*. Gordon and Breach, 1961.

R. Chen, I. Lee, Y. Lee, and Y. Lo. Upgrading domain ontology based on latent semantic analysis and group center similarity calculation. In *Proceedings of the 2008 IEEE International Conference on Systems, Man, and Cybernetics (SMC 2008)*, pages 1495–1500. IEEE, 2008.

H. Chim and X. Deng. A new suffix tree similarity measure for document clustering. In *Proceedings of the 16th International Conference on World Wide Web (WWW 2007)*, pages 121–130, New York, NY, 2007. ACM.

S. L. Chuang and L. F. Chien. Taxonomy generation for text segments: A practical web-based approach. *ACM Transactions on Information Systems (TOIS)*, 23(4):363–396, 2005.

P. Cimiano, A. Hotho, and S. Staab. Learning concept hierarchies from text corpora using formal concept analysis. *Journal of Artificial Intelligence Research*, 24:305–339, 2005.

C.W. Cleverdon. ASLIB Cranfield research project on the comparative efficiency of indexing systems. In *ASLIB Proceedings*, volume 12, pages 421–431. College of Aeronautics, Cranfield, 1960.

D. Crabtree, X. Gao, and P. Andreae. Improving web clustering by cluster selection. In *Proceedings of the 2005 IEEE/WIC/ACM International Conference on Web Intelligence (WI 2005)*, pages 172–178, Washington, DC, USA, 2005. IEEE Computer Society.

N. Craswell and M. Szummer. Random walks on the click graph. In *Proceeding of the 30th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2007)*, pages 159–166, Amsterdam, The Netherlands, 2007. ACM.

M. Cuadros and G. Rigau. Quality assessment of large scale knowledge resources. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing (EMNLP 2006)*, pages 534–541, Sydney, Australia, 2006. ACM.

J. Cullen and A. Bryman. The knowledge acquisition bottleneck: time for reassessment? *Expert Systems*, 5(3):216–225, 1988.

A. Culotta and J. Sorencen. Dependency tree kernels for relation extraction. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL 2004)*, Barcelona, 2004. ACL.

I. Dagan, O. Glickman, and B. Magnini. The PASCAL recognizing textual entailment challenge. In *Proceedings of the PASCAL workshop on Recognizing Textual Entailment*, pages 1–8, Southampton, UK, 2005.

K. Dellschaft and S. Staab. On how to perform a gold standard based evaluation of ontology learning. In *Proceedings of the 5th International Semantic Web Conference (ISWC 2006)*, pages 228–241, Athens, GA, USA, 2006.

K. Dellschaft and S. Staab. Strategies for the evaluation of ontology learning. In *Proceedings of the 2008 Conference on Ontology Learning and Population: Bridging the Gap between Text and Knowledge*, volume 167, pages 253–272, Amsterdam, The Netherlands, 2008. IOS Press.

H. Deng, I. King, and M. Lyu. Entropy-biased models for query representation on the click graph. In *Proceedings of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2009)*, pages 339–346. ACM, 2009.

J. Diederich and W. Balke. Automatically created concept graphs using descriptive keywords in the medical domain. *Methods of Information in Medicine*, 47(3):241–250, 2008.

M. Dittenbach, H. Berger, and D. Merll. Improving domain ontologies by mining semantics from text. In *Proceedings of the 1st Asian-Pacific Conference on Conceptual Modelling (APCCM 2004)*, volume 31, pages 91–100. Australian Computer Society, 2004.

L. Doyle. Semantic road maps for literature searchers. *Journal of the ACM (JACM)*, 8(4):553–578, 1961.

C. Fellbaum. *WordNet: an electronic lexical database*. MIT Press, 1998.

G. Flouris, D. Manakanatas, H. Kondylakis, D. Plexousakis, and G. Antoniou. Ontology change: classification and survey. *The Knowledge Engineering Review*, 23(2):117–152, 2008.

D. Giampiccolo, H. Dang, B. Magnini, I. Dagan, and B. Dolan. The fourth PASCAL recognizing textual entailment challenge. In *Proceedings of the Text Analysis Conference(TAC 2010)*, pages 1–9, Gaithersburg, MD., 2008.

R. Girju, A. Badulescu, and D. Moldovan. Automatic discovery of part-whole relations. *Computational Linguistics*, 32(1):83–135, 2006.

C. Giuliano, A. Lavelli, D. Pighin, and L. Romano. FBK-IRST: Kernel methods for semantic relation extraction. In *Proceedings of the 4th International Workshop on Semantic Evaluations (SemEval 2007)*, pages 141–144. ACL, 2007.

F. Giunchiglia, M. Yatskevich, P. Avesani, and P. Shivaiko. A large dataset for the evaluation of ontology matching. *The Knowledge Engineering Review Archive*, 24(2):137–157, 2009.

G. Golovchinsky, P. Qvarfordt, and J. Pickens. Collaborative information seeking. *Information Seeking Support Systems*, 42(3):47–51, 2009.

G. Grefenstette. Use of syntactic context to produce term association lists for text retrieval. In *Proceedings of the 33rd international ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 1992)*, pages 89–97. ACM, 1992.

H. Gürkök, M. Karamuftuoglu, and M. Schaal. A graph based approach to estimating lexical cohesion. In *Proceedings of the 2nd International Symposium on Information Interaction in Context (IIiX 2008)*, pages 35–43, London, UK, 2008. ACM.

S. Hattori and K. Tanaka. Extracting concept hierarchy knowledge from the web based on property inheritance and aggregation. In *Proceedings of the 2008 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT 2008)*, volume 1, pages 432–437. IEEE Computer Society, 2008.

M. Hearst. Automatic acquisition of hyponyms from large text corpora. In *Proceedings of the 14th International Conference on Computational Linguistics (COLING 1992)*, volume 2, pages 539 – 545. ACL, 1992.

L. Hollink, G. Schreiber, and B. Wielinga. Patterns of semantic relations to improve image content search. *Web Semantics: Science, Services and Agents on the World Wide Web*, 5(3):195–203, 2007.

E. Hovy, Z. Kozareva, and E. Riloff. Toward completeness in concept extraction and classification. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing (EMNLP 2009)*, volume 2, pages 948–957. ACL, 2009.

M. Hsu, M. Tsai, and H. Chen. Query expansion with conceptnet and wordnet: An intrinsic comparison. In *Proceedings of the 3rd Asia Information Retrieval Symposium (AIRS 2006)*, pages 1–13. Springer-Verlag, 2006.

B. Katz and J. Lin. REXTOR: a system for generating relations from natural language. In *Proceedings of the ACL Workshop on Recent Advances in Natural Language Processing and Information Retrieval*, pages 67–77. ACL, 2000.

Z. Kozareva and E. 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*, pages 1110–1118. Association for Computational Linguistics, 2010.

U. Kruschwitz. An adaptable search system for collections of partially structured documents. *IEEE Intelligent Systems*, 18(4):44–52, 2003.

U. Kruschwitz and H. Al-Bakour. Users want more sophisticated search assistants: Results of a task-based evaluation. *Journal of the American Society for Information Science and Technology (JASIST)*, 56(13):1377–1393, 2005.

U. Kruschwitz, M-D. Albakour, J. Niu, J. Leveling, N. Nanas, Y. Kim, D. Song, M. Fasli, and A. De Roeck. Moving towards Adaptive Search in Digital Libraries. In *Advanced Language Technologies for Digital Libraries*, volume 6699 of *Lecture Notes in Computer Science*, pages 41–60. Springer Berlin, Heidelberg, 2011. Forthcoming.

R. Lau, J. Hao, M. Tang, and X. Zhou. Towards context-sensitive domain ontology extraction. In *Proceedings of the 40th Annual Hawaii International Conference on System Sciences (HICSS 2007)*, page 60. IEEE Computer Society, 2007.

R. Lau, P. Bruza, and D. Song. Towards a belief-revision-based adaptive and context-sensitive information retrieval system. *ACM Transactions on Information Systems (TOIS)*, 26(2):1–38, 2008.

R. Lau, D. Song, Y. Li, T. Cheung, and J. Hao. Toward a fuzzy domain ontology extraction method for adaptive e-learning. *IEEE Transactions on Knowledge and Data Engineering*, 21(6):800–813, 2009.

D. Lawrie and W. Croft. Discovering and comparing topic hierarchies. In *Proceedings of 6th International Conference on Computer-Assisted Information Retrieval (RIAO 2000)*, pages 314–330, 2000.

D. Lenat, M. Prakash, and M. Shepherd. Cyc: Using common sense knowledge to overcome brittleness and knowledge acquisition bottlenecks. *AI magazine*, 6(4):65–85, 1985.

C. Lin and E. Hovy. The automated acquisition of topic signatures for text summarization. In *Proceedings of the 18th International Conference on Computational Linguistics (COLING 2000)*, volume 1, pages 495–501. ACL, 2000.

H. Liu and P. Singh. ConceptNet: a practical commonsense reasoning toolkit. *BT Technology Journal*, 22(4):211–226, 2004.

D. Lungley and U. Kruschwitz. Automatically maintained domain knowledge: Initial findings. In *Proceedings of the 31st European Conference on IR Research (ECIR 2009)*, volume 5478 of *Lecture Notes In Computer Science*, pages 739–743. Springer Verlag, 2009.

A. Maedche and S. Staab. Measuring similarity between ontologies. *Knowledge Engineering and Knowledge Management: Ontologies and the Semantic Web*, pages 15–21, 2002.

A. Maedche, B. Motik, L. Stojanovic, R. Studer, and R. Volz. An infrastructure for searching, reusing and evolving distributed ontologies. In *Proceedings of the 12th International Conference on World Wide Web (WWW 2003)*, pages 439–448. ACM, 2003.

G. Marchionini and R.W. White. Information-Seeking Support Systems [Guest Editors' Introduction]. *Computer*, 42(3):30–32, 2009.

O. Medelyan and C. Legg. Integrating Cyc and wikipedia: Folksonomy meets rigorously defined common-sense. In *Proceedings of the Workshop on Wikipedia and Artificial Intellignece: An Evolving Synergy (WIKI-AI 2008)*, volume 8, pages 13–18, 2008.

M. Mintz, S. Bills, R. Snow, and D. Jurafsky. Distant supervision for relation extraction without labeled data. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the Association for Computational Linguistics and the 4th International Joint Conference on Natural Language Processing of the AFNLP (ACL-IJCNLP 2009)*, volume 2, pages 1003–1011, Morristown, NJ, 2009. ACL.

P. Monachesi, T. Markus, and E. Mossel. Ontology enrichment with social tags for elearning. In *Learning in the Synergy of Multiple Disciplines*, volume 5794 of *Lecture Notes in Computer Science*, pages 385–390. Springer, 2009.

E. Morin and C. Jacquemin. Automatic acquisition and expansion of hypernym links. *Computers and the Humanities*, 38(4):363–396, 2004.

N. Nanas. *Towards Nootropia: A Non-Linear Approach to Adaptive Document Filtering*. PhD thesis, Knowledge Media Institute, The Open University, UK, 2003.

N. Nanas and A. de Roeck. Autopoiesis, the immune system, and adaptive information filtering. *Natural Computing*, 8(2):387–427, 2009.

N. Nanas, V. Uren, and A. De Roeck. A review of evolutionary and immune-inspired information filtering. *Natural Computing*, 9:1–29, 2010.

V. Nastase, M. Strube, B. Boerschinger, C. Zirn, and A. Elghafari. Wikinet: A very large scale multi-lingual concept network. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC 2010)*, pages 1015–1022, Valletta, Malta, May 2010. European Language Resources Association (ELRA).

R. Navigli. Semi-automatic extension of large-scale linguistic knowledge bases. In *Proceedings of the 18th International Florida Artificial Intelligence Research Society Conference (FLAIRS 2005)*, pages 548–553, 2005.

R. Navigli. Using cycles and quasi-cycles to disambiguate dictionary glosses. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2009)*, pages 594–602, Athens, Greece, 2009a. ACL.

R. Navigli. Word sense disambiguation: a survey. *ACM Computing Surveys*, 41(2):1–69, 2009b.

R. Navigli and M. Lapata. An experimental study of graph connectivity for unsupervised word sense disambiguation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(4):678–692, 2010.

R. Navigli and S. Ponzetto. Babelnet: Building a very large multilingual semantic network. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics (ACL 2010)*, pages 216–225, Uppsala, Sweden, July 2010.

R. Navigli and Velardi. From glossaries to ontologies: Extracting semantic structure from textual definitions. In *Proceedings of the 2008 Conference on Ontology Learning and Population: Bridging the Gap between Text and Knowledge*, pages 71–87, Amsterdam, The Netherlands, 2008. IOS Press.

R. Navigli and P. Velardi. An analysis of ontology-based query expansion strategies. In *Proceedings of the ECML Workshop on Adaptive Text Extraction and Mining (ATEM2003)*, pages 42–49, Cavtat Dubrovnik, Croatia, 2003.

R. Navigli and P. Velardi. Ontology enrichment through automatic semantic annotation of on-line glossaries. *Managing Knowledge in a World of Networks*, pages 126–140, 2006.

R. Navigli, P. Velardi, A. Cucchiarelli, and F. Neri. Quantitative and qualitative evaluation of the OntoLearn ontology learning system. In *Proceedings of the 20th international Conference on Computational Linguistics (COLING '04)*, pages 1043–1050, Morristown, NJ, 2004. ACL.

M. Paşca and E. Alfonseca. Web-derived resources for web information retrieval: from conceptual hierarchies to attribute hierarchies. In *Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2009)*, pages 596–603. ACM, 2009.

P. Pantel and D. Lin. Discovering word senses from text. In *Proceedings of the 8th ACM International Conference on Knowledge Discovery and Data Mining (SIGKDD 2002)*, pages 613–619. ACM, 2002.

P. Pantel and M. Pennacchiotti. Espresso: Leveraging generic patterns for automatically harvesting semantic relations. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th Annual Meeting of the Association for Computational Linguistics (COLING-ACL 2006)*, pages 113–120. ACL, 2006.

A. Passant. Using ontologies to strengthen folksonomies and enrich information retrieval in weblogs. In *Proceedings of the International Conference on Weblogs and Social Media (ICWSM 2007)*, 2007.

M. Pennacchiotti and P. Pantel. Ontologizing semantic relations. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th Annual Meeting of the Association for Computational Linguistics (COLING-ACL 2006)*, pages 793–800. ACL, 2006.

M. Phillips. *Aspects of Text Structure: an investigation of the lexical organization of text*, volume 52 of *North-Holland Linguistic Series*. Elsevier, 1985.

M. Poesio and A. Almuhareb. Extracting concept descriptions from the web: the importance of attributes and values. In *Proceedings of the 2008 Conference on Ontology Learning and Population: Bridging the Gap Between Text and Knowledge*, pages 29–44. IOS Press, 2008.

S. Ponzetto and R. Navigli. Large-scale taxonomy mapping for restructuring and integrating wikipedia. In *Proceedings of the 21st International Joint Conference on Artificial Intelligence (IJCAI 2009)*, pages 2083–2088, 2009.

S. Ponzetto and M. Strube. Deriving a large scale taxonomy from wikipedia. In *Proceedings of the 22nd National Conference on Artificial Intelligence (AAAI 2007)*, volume 2, pages 1440–1445. AAAI Press, 2007.

M.R. Quillian. Word concepts: A theory and simulation of some basic semantic capabilities. *Behavioral Science*, 12:410–430, 1967.

F. Radlinski and T. Joachims. Query chains: learning to rank from implicit feedback. In *Proceedings of the 11th ACM International Conference on*

Knowledge Discovery and Data Mining (SIGKDD 2005), pages 239–248. ACM, 2005.

F. Reichartz, H. Korte, and G. Paass. Composite kernels for relation extraction. 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 (ACL-IJCNLP 2009)*, pages 365–368, 2009.

S.D. Richardson, W.B. Dolan, and L. Vanderwende. MindNet: acquiring and structuring semantic information from text. In *Proceedings of the 17th International Conference on Computational Linguistics and the 36th Annual Meeting of the Association for Computational Linguistics (COLING-ACL 1998)*, volume 2, pages 1098–1102. ACL, 1998.

M. Sanderson and B. Croft. Deriving concept hierarchies from text. In *Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 1999)*, pages 206–213. ACM, 1999.

H. Schütze. Automatic word sense discrimination. *Computational Linguistics - Special Issue on Word Sense Disambiguation*, 24(1):97–123, 1998.

K. Shinzato and K. Torisawa. Acquiring hyponymy relations from web documents. In *Proceedings of Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics (HLT-NAACL 2004)*, pages 73–80. ACL, 2004.

P. Singh, T. Lin, E. Mueller, G. Lim, T. Perkins, and W. Li Zhu. Open mind common sense: Knowledge acquisition from the general public. In *On the Move to Meaningful Internet Systems 2002: CoopIS, DOA, and ODBASE*, volume 2519 of *Lecture Notes in Computer Science*, pages 1223–1237. Springer, 2002.

R. Snow, D. Jurafsky, and A. Ng. Learning syntactic patterns for automatic hypernym discovery. *Advances in Neural Information Processing Systems*, 17:1297–1304, 2005.

R. Snow, D. Jurafsky, and A. 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 (COLING-ACL 2006), pages 801–808. ACL, 2006.

J. F. Sowa. Conceptual graphs. In *Handbook of Knowledge Representation*, volume 3 of *Foundations of Artificial Intelligence*, chapter 5, pages 213–237. Elsevier, 2008.

J.F. Sowa. *Conceptual Structures*. Addison-Wesley., Reading, M.A., 1984.

R. Speer, C. Havasi, and H. Lieberman. AnalogySpace: reducing the dimensionality of common sense knowledge. In *Proceedings of the 23rd National Conference on Artificial Intelligence (AAAI 2008)*, volume 1, pages 548–553. AAAI Press, 2008.

F. Suchanek, G. Kasneci, and G. Weikum. Yago: a core of semantic knowledge. In *Proceedings of the 16th International Conference on World Wide Web (WWW 2007)*, pages 697–706, 2007.

Y. Y. Tang, C. D. Yan, and C. Y. Suen. Document processing for automatic knowledge acquisition. *IEEE Transactions on Knowledge and Data Engineering*, 6:3–21, 1994.

M. Thelen and E. Riloff. A bootstrapping method for learning semantic lexicons using extraction pattern contexts. In *Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing (EMNLP 2002)*, volume 10, pages 214–221. ACL, 2002.

A. Toumouth, A. Lehireche, D. Widdows, and M. Malki. Adapting WordNet to the medical domain using lexicosyntactic patterns in the ohsumed corpus. In *Proceedings of the 2006 IEEE/ACS International Conference on Computer Systems and Applications (AICCSA 2006)*, pages 1029–1036. ACL, 2006.

A. Valarakos, G. Paliouras, V. Karkaletsis, G. Vouros, E. Motta, N. Shadbolt, A. Stutt, and N. Gibbins. Enhancing ontological knowledge through ontology population and enrichment. In *Proceedings of the 14th International Conference on Engineering Knowledge in the Age of the Semantic Web (EKAW 2004)*, pages 144–156, 2004.

D. Vallet, M. Fernández, and P. Castells. An ontology-based information retrieval model. *The Semantic Web: Research and Applications*, pages 455–470, 2005.

L. van der Plas and J. Tiedemann. Using lexico-semantic information for query expansion in passage retrieval for question answering. In *Proceedings of the COLING Workshop on Information Retrieval for Question Answering (IRQA 2008)*, pages 50–57. ACL, 2008.

C. van Rijsbergen. *Information Retrieval*. Butterworths, London, U.K., 2nd edition, 1979.

C. Wagner. Breaking the knowledge acquisition bottleneck through conversational knowledge management. *Information Resources Management Journal*, 19(1):70–83, 2006.

D. Widdows. *Geometry and Meaning*. CSLI Lecture Notes, 2004.

D. Widdows and B. Dorow. A graph model for unsupervised lexical acquisition. In *Proceedings of the 19th International Conference on Computational Linguistics (COLING 2002)*, volume 1, pages 1–7. ACL, 2002.

D. Widdows, S. Cederberg, and B. Dorow. Visualisation techniques for analysing meaning. In *Proceedings of the 5th International Conference on Text, Speech, and Dialogue (TSD 2002)*, pages 107–114, 2002.

Y. Wilks and C. Brewster. *Natural Language Processing as a Foundation of the Semantic Web*. Foundations and Trends in Information Retrieval. Now Publishers, 2006.

T.D. Wilson. Models in information behaviour research. *Journal of Documentation*, 55(3):249–270, 1999.

W. A. Woods. What's in a Link: Foundations for Semantic Networks. In *Representation and Understanding: Studies in Cognitive Science*, pages 35–82. Academic Press, 1975.

O. Zamir and O. Etzioni. Web document clustering: A feasibility demonstration. In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 1998)*, pages 46–54. ACM, 1998.

C. Zhang and D. Wu. Concept extraction and clustering for topic digital library construction. In *Proceedings of the 2008 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT 2008)*, volume 3, pages 299–302. IEEE Computer Society, 2008.