Using Associations and Fuzzy Ontologies for Modeling Chemical Safety Information

Mika Timonen, Antti Pakonen and Teemu Tommila
Technical Research Centre of Finland, VTT, PO Box 1000, FI-02044 VTT, Espoo, Finland

Keywords: Association Modeling, Chemical Safety Information, Fuzzy Ontology, Knowledge Modeling, Query Expansion.

Abstract: In this paper we propose a novel approach for domain modeling that combines two different types of models: (1) fuzzy ontology that describes the concepts of the domain and their relations in a formal way, and (2) association model that presents the associations between the terms of the domain. We utilize the combined model for query expansion by finding both highly associative and related concepts for the query terms. To demonstrate the feasibility of the model and its utilization, we use the query expansion in a search engine of chemical safety cards.

1 INTRODUCTION

In this paper we introduce a novel approach for domain modeling that utilizes both fuzzy ontologies and associations for retrieving relevant information from document databases. We focus on a database consisting of approximately 2,000 chemical safety cards. As the search space is small, the relevance of the search results may be poor. Therefore, we use query expansion that utilizes the domain model to enhance the search results.

Ontologies are often used for defining the semantics of a domain terminology in a machine processable form. The idea is to use the ontology to enrich the information in the domain and model the relationships of the concepts in the domain. One problem with current ontology languages, such as OWL (W3C Recommendation, 2004), is that they lack sufficient means of addressing the uncertainty inherent in human communication (Carlsson et al., 2010). OWL ontologies are ”crisp” representations of a black-and-white world, whereas human communication is inexact, person-dependent, and often ambiguous. *Fuzzy ontologies* provide a way to represent the uncertainty by including weights into relationships between the concepts (Hirvonen et al., 2010; Parry, 2006; Sanchez and Yamanoi, 2006; Widyantoro and Yen, 2001).

We make a distinction between abstract concepts that in our thoughts refer to real-world things and the symbols for those concepts that we use to communicate our ideas to other people. In our case, the ontology is basically an abstraction of the domain concepts and the terms in the documents represent the natural language symbols for them. Consequently, there is the need to describe uncertainties in two places. Firstly, we must describe the fundamental imprecision of the concepts themselves, for example, in the case of overlapping geographical areas. Secondly, the terms used in natural languages are often ambiguous. For example, the word ”Jaguar” can symbolize the concept of an animal or a car. Moreover, natural language terms can be related to each other in various, imprecise ways.

To describe the conceptual uncertainties, we have developed a fuzzy ontology. And to model uncertainties in natural language terms, we apply the idea of an association network. As ontologies are often implemented in such a way that they contain mostly hierarchical *is-a* and *part-of* relationships between the concepts, they do not capture all the information of the domain. For example, ”car” is-a ”vehicle” represents a relationship that is often described by an ontology. If we want to model a relationship between ”car” and ”road”, or ”car” and ”traffic lights”, we would have to build a more complex ontology that would include relationships which can models these relationships.

An *association network* (Timonen et al., 2011) aims to address this issue by modeling associations between the terms of a domain. The associations are not limited to semantic or hierarchical relationships as they aim to model, for example, the co-occurrence of the terms in the domain. An association between two

Timonen M., Pakonen A. and Tommila T..
Using Associations and Fuzzy Ontologies for Modeling Chemical Safety Information.
DOI: 10.5220/0004536400260037
In *Proceedings of the International Conference on Knowledge Engineering and Ontology Development* (KEOD-2013), pages 26-37
ISBN: 978-989-8565-81-5

terms holds only a weight that depicts the strength of the association. It does not identify the type of the relationship. When compared with ontologies, the biggest benefit an association network provides comes from the fact that it can be trained from a set of documents using an unsupervised method (Timonen et al., 2011). That is, it requires very little manual labor.

In this paper, we focus on a prototype search engine that is used for searching chemical safety information. We utilize the information from the safety cards to build two models: a fuzzy ontology and an association network that aim to describe the terms and their relationships in the domain. The aim of the search engine is to support the information gathering of chemical safety experts who write the safety data sheets for their products. The purpose of the data sheets required by the regulations is to ensure that the hazards presented by chemicals are clearly communicated to workers and consumers. During the writing process the expert often needs information about related and similar chemicals. As a test material we use the International Safety Cards (ICSC) maintained by the International Labour Organization (ILO).

The challenge when focusing only on a small set of documents is that a search may often produce only a small set of results that may not be relevant to the original query, in particular if the user does not know the correct query terms. Therefore, we use the two models for query expansion where the aim is to include related concepts to the original query so that the result set consisting of chemical cards contains as much relevant information as possible. In this case, a fuzzy ontology provides a distinct benefit by not limiting the query expansion to crisp relationships.

This paper makes the following contributions: (1) a novel approach for domain modeling that utilizes both the uncertain domain knowledge in a fuzzy ontology and the associations of the terms, (2) a novel query expansion approach that uses the domain model, and (3) a case study where we present the use of the query expansion in search of chemical safety information.

This paper is organized as follows: in Section 2 we discuss the related work in the areas of ontologies, association modeling, and query expansion. In Section 3 we propose a novel method for domain modeling. In Section 4 we present the case study we conducted with chemical safety cards and propose a novel approach for query expansion. We conclude our work in Section 5.

2 RELATED WORK

In this section we describe the related work in the three areas relevant to our work: ontologies, association modeling, and query expansion.

2.1 Fuzzy Ontologies

Since the knowledge in the real world is often characterized by uncertainty and inconsistency, the ”crisp” logic of Semantic Web languages like OWL (W3C Recommendation, 2004) has been challenged.

To some extent, it seems that the response of the Semantic Web community is that addressing uncertainty in ontologies would result in solutions that do not scale (Thomas and Sheth, 2006). However, there has been a World Wide Web Incubator Group at W3C working on uncertainty reasoning for the WWW. In a report published by the group (Laskey and Laskey, 2008), it is stated that information in large networks is likely to be uncertain, incomplete, and often also incorrect. Uncertainty representation and reasoning is needed to deal with different levels of confidence and trust, and also to enable conceptually overlapping ontologies to interoperate.

The Incubator Group goes on to recommend that addressing uncertainty in ontologies would increase the usefulness of Web-based information, and a standard way of representing uncertainty should be developed (Laskey and Laskey, 2008). The representation should also support defining properties for different uncertainty formalisms. Possible formalisms examined by Laskey and Laskey (Laskey and Laskey, 2008) include probability theory, fuzzy logic, and belief functions.

2.1.1 Methods for Addressing Uncertainty

Ontologies based on probability theory employ a mathematical representation language for specifying degrees of belief over statements regarding domain knowledge. The approach is promising for systems where there are different sources that contain uncertain and imperfect knowledge, making it necessary to assess the likelihood of a statement to be true or false. There are many different ways to combine probability theory with ontologies. For a review of several such approaches, see the appendices of (Laskey and Laskey, 2008).

Fuzzy ontologies (Parry, 2006; Sanchez and Yamanoi, 2006; Widyantoro and Yen, 2001)), on the other hand, deal with vagueness and imprecision, and draw influence from both fuzzy logic and crisp ontology languages. A certain term in a fuzzy ontology can have many different meanings, each with an assigned membership value. A fuzzy mapping enables the task of finding knowledge from systems with inconsistent views on domain vocabulary (Thomas and Sheth, 2006). For example, according to Parry (Parry, 2006), the fuzzy ontology is based on the idea that each ontology concept is related to every other concept in the ontology, with a degree of membership assigned to that relationship based on fuzzy logic. As in Figure 1 we can then specify that the term ”Apple” can represent a type of both a fruit and a computer company. Note that this example does not make a distinction between concepts and their corresponding symbols in natural language.

2.1.2 Utilizing Uncertainty in Query Expansion

When using a text- or keyword-based search engine, the query must be precisely articulated, and it can be challenging to find the exact right query terms that will lead to the discovery of the most useful information. A solution to this problem is query expansion (or ”query refinement”) - responding to the initial query by suggesting a list of terms that are broader, narrower, or otherwise related (Widyantoro and Yen, 2001). Fuzzy ontologies, in particular, have been proposed as a mechanism for enabling the expansion of information queries (Bordogna and Pasi, 2000; Parry, 2006; Widyantoro and Yen, 2001).

In essence, query expansion is a matter of assessing the semantic similarity of query terms, comparing meanings rather than syntactic differences (which can easily be handled by generic search engines like Google). For measuring semantic similarity - or semantic distance - several methods have been proposed, based on, e.g., distance within an ontological structure or concept feature matching (Cross, 2004; Janowicz et al., 2012).

2.2 Association Modeling

Associations have been used previously in neural networks and for gene function mapping. For example, Mostafavi et al. (Mostafavi et al., 2008) use association network to represent a network of genes and proteins where they are linked with undirected edges that are weighted according co-functionality implied by a data source. The network is used for predicting annotated gene functions in blind tests (Peña-Castillo et al., 2008).

Timonen et al. (Timonen et al., 2011; Timonen, 2013) use the term ”association network” to describe a form of domain modeling that aims to identify strong links between keywords. The term should not be confused with associative neural networks (Tetko, 2002a; Tetko, 2002b). Ontologies aim to describe semantic relations, such as classification (hyponyms and hypernyms), composition (part-of) and various dependencies between domain concepts. The aim of association modeling is to include other types of relations to the domain model. Its intuition comes from human associative memory: what other concepts we tend to think when we think of a particular concept. These concepts may have semantic relations but they can also be words that occur often together. For example, when thinking of ”car”, in addition to specific brands of cars (”Ford”, ”Volkswagen”) we may also think concepts like ”road”, ”speed limit” and ”traffic jams”.

An association network consists of nodes (n) and edges (e). Nodes are the terms (i.e., words and noun phrases) of the domain. The nodes are linked together with edges that hold a weight which represents the strength of the association. The weights range between $0 < w \leq 1$ where a strong weight is depicted with 1.0. There is no link between edges with weight 0.0.

Timonen et al. (Timonen et al., 2011) utilize three components when assessing the association weight: *confidence* (i.e., co-occurrence), *distance*, and *age*. The main component of the weight is based on the confidence used in association rule mining (Agrawal et al., 1993). That is, the confidence of term c_2 given the term c_1, i.e., link from c_1 to c_2:

$$confidence(c_1 \rightarrow c_2) = \frac{frequency(c_1 \cap c_2)}{frequency(c_1)}, \quad (1)$$

where $frequency(c_1 \cap c_2)$ is the number of times c_1 occurs with c_2, and $frequency(c_1)$ is the number of times c_1 occurs in total.

The aim of the confidence value is to give high weights to term pairs that co-occur often. However, in addition to confidence, two other features can be included to the associations: distance between the terms (in the document), and age of the terms in the domain. The distance aims to measure the average distance between the keywords; if they occur often close to each other in the documents, they have a higher weight. The age component aims to mimic the deterioration of unused pathways; i.e., if the keyword is old it is not as interesting. Therefore, the association weights of newer keywords should be stronger than of the old ones. The way Timonen et al. (Timonen et al., 2011) assessed the distance and age components are domain specific; for more information about the components we refer the reader to the original publications (Timonen et al., 2011; Timonen, 2013).

The strength of the association from c_1 to c_2 is the

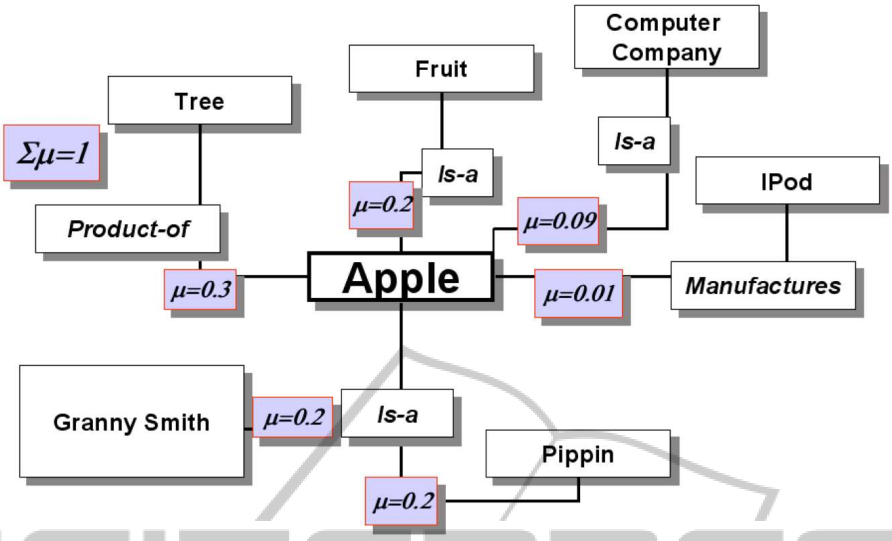


Figure 1: In a fuzzy ontology, the term ”apple” can mean many things (originally presented in (Parry, 2006)).

combination of these three components:

$$Strength(c_1 \rightarrow c_2) = confidence(c_1 \rightarrow c_2) \times \frac{age(c_1)}{distance(c_1, c_2)}. \quad (2)$$

The strength is normalized to fall between $[0,1]$ by dividing each weight with the maximum out-going weight of the node. The normalization is done to all of the node's out-going edges so that the strongest weight is scaled to 1.0.

This approach has similarities with TextRank (Mihalcea and Tarau, 2004) algorithm as it considers co-occurrence as the main component of the weight between the terms. However, as the weights are anti-symmetric (i.e., $Strength(c_1 \rightarrow c_2) \neq Strength(c_2 \rightarrow c_1)$) in Timonen et al. approach, we consider association networks more suitable for domain modeling.

2.3 Query Expansion and Reformulation

Query expansion is a process that aims to reformulate a query to improve the results of information retrieval. This is important especially when the original query is short or ambiguous and would therefore give only irrelevant results. By expanding the query with related terms the reformulated query may produce good results.

Carpineto and Romano (Carpineto and Romano, 2012) have surveyed query expansion techniques. According to them, the standard methods include: semantic expansion, word stemming and error correction, clustering, search log analysis and web data utilization.

In semantic expansion, the idea is to include semantically similar terms to the query. These words include synonyms and hyponyms. When using word stemming, the idea is to use a stemmed version of the word so that different types of spellings can be found (e.g., singular and plural). Term clustering is a way to find similar terms by using term co-occurrence. Search log analysis is another way of finding similar terms. In this case, the logs are analyzed to identify terms that often co-occur with the given query terms. Finally, web data utilization is an approach where an external data source (e.g., Wikipedia[1]) is used for query expansion. The idea here is to use hyperlinks in Wikipedia to find related topics for the query terms.

Bhogal et al. (Bhogal et al., 2007) also reviewed query expansion approaches. They mainly focus on three areas in their review: relevance feedback, corpus dependent knowledge models and corpus independent models. Relevance feedback is one of the oldest methods for expansion. It expands the query using terms from relevant documents. The documents are assessed as relevant if they are ranked highly in previous queries or identified as relevant in other ways (e.g., manually). Corpus dependent knowledge models take a set of documents from the domain and uses them to model the characteristics of the corpus. This includes the previously mentioned stemming and co-occurrence approaches. Corpus independent knowledge models includes semantic expansion and the web data utilization as it uses dictionaries such as WordNet[2] to include synonyms and hyponyms into the search. For more information, we re-

[1] http://www.wikipedia.org/

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

fer the reader to the original articles by Carpineto and Romano (Carpineto and Romano, 2012) and Bhogal et al. (Bhogal et al., 2007).

3 MODELING OF CHEMICAL SAFETY INFORMATION

A document space consists of all the documents that are stored to the database. The space also holds all the terms that are found from the documents and from the metadata of the documents. We use two approaches to model the document space: the fuzzy ontology and the association network. In addition to the document space and the search engine, we have implemented helper tools for the laborious tasks of ontology creation, association network generation and automatic annotation of the chemical cards. In this section we describe the ontology and the association modeling in more detail.

3.1 Fuzzy Domain Ontologies

We have been working on a weighted ontology, inspired by the languages of the Semantic Web, and different approaches at depicting uncertainty in conceptual models. Our basic need is to be able to address the inherent uncertainty and conceptual overlap in the properties of different chemicals. Specifically, we have been interested in specifying *weighted relationships* to support expanded queries based on domain *keywords*. Therefore, the idea of *relatedness* of keywords has been a guiding factor in the way the ontology is specified.

Our approach draws influence from fuzzy ontologies, but looks for structures that are simple to define and process. In a fuzzy ontology, a *fuzzy set* is defined by its membership function mapping each element of the domain to a *membership degree* value. A *fuzzy number*, such as a ”young person”, is a fuzzy set of numerical values like real numbers or integers. Our ontology only uses membership degrees between 0 and 1 (expressed with a qualitative value like ”minor” or ”significant”) to describe conceptual uncertainties. There are similar approaches available for other domains (Formica et al., 2008; Yang et al., 2005; Zhang et al., 2006).

Figure 2 illustrates the data model we use for describing the conceptual uncertainties in OWL. In general, knowledge repositories contain valuable pieces of knowledge called nuggets. In our example, chemical safety card is the only subclass of nugget. Each card instance is annotated with keywords picked from the ontology. The space of all domain concepts is first divided into a few orthogonal dimensions that we call keyword categories. In our case the categories are:

- **Material:** Characteristics of materials, e.g., chemical properties (acid, liquid, etc.).
- **Usage:** Typical area where a product is used, e.g., industrial sector.
- **Danger:** Hazards related to a chemical, e.g., toxicity.
- **Entity:** Entities possibly harmed by the danger, e.g., people or the environment.
- **Exposure:** Route of the harmful effect, e.g., inhalation.
- **Precaution:** Preventive and corrective measures, e.g., use of rubber gloves.

Within each category, the keyword instances that correspond to the domain concepts are organized with three fuzzy relationship types. The specialization is used to represent the classification (is-a) of concepts in a category. Similarly, the part-of relationship describes the decomposition of wholes to part. In addition to these common relationships present in most ontologies, we model all other link types with the generic dependency relationship. It can be used, for example, across category boundaries to tell that a chemical is typically used in a particular area of industry. All these relationships between keywords are modeled as relationship instances that associate a weight value with the relationship. For convenience, this value is selected from a predefined set of linguistic labels. A noteworthy feature of our model is that specializations and part-of relationships have two separate weights (one for inclusion and another for coverage) depending on the direction up or down in the classification or part-of hierarchy. This makes the links asymmetric which has an effect on the search algorithm and results.

As is usually the case in ontology development, the number of different concepts is large and the relationships between them even more numerous. Consequently, the costs of the initial ontology and its continuous maintenance can be very high. On the other hand, more tuning parameters and mathematics are needed. In our experience, the complexity and cost of a fuzzy ontology is one of its main weaknesses.

To alleviate this issue we have developed a few tools to automate the process. First, we use the European collection of standard phrases for chemical data sheets EuPhraC[3] as a starting point. In contains large number of relevant terms partly organized as fragmentary taxonomies. It was relatively easy to tag and

[3] http://www.esdscom.eu/euphrac.html

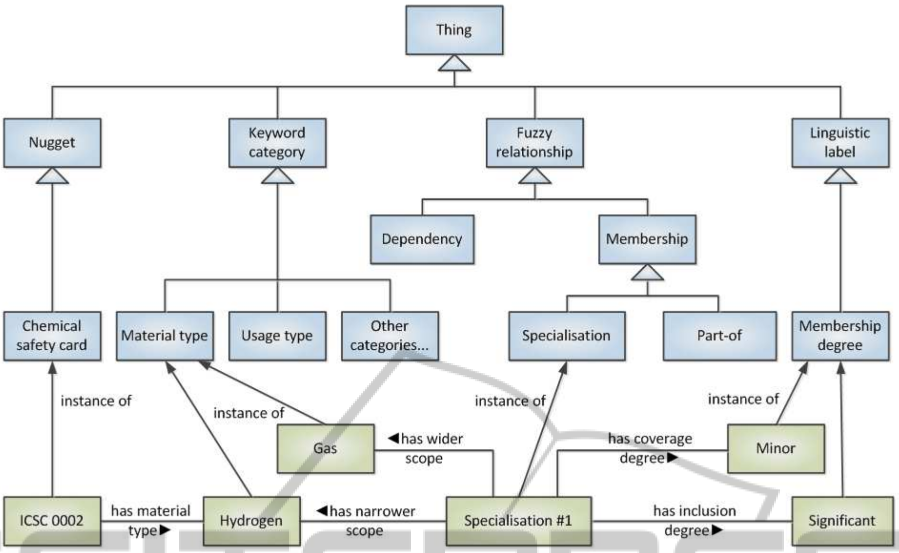


Figure 2: OWL representation of a fuzzy ontology.

automatically process the Excel files, and to generate the ontology in OWL format. However, manual additions and refinements were needed to create collection of about 650 keywords and 900 fuzzy relationships.

Another task was to annotate the nearly 1,700 chemical cards with suitable keywords in each keyword category. Fortunately, ILO's web server maintains the information in a fixed HTML format, mostly containing table elements. This made it possible to download the cards and to look selected table cells for terms found in the ontology.

3.2 Association Modeling

The idea behind association modeling is to represent the term relationships to mimic human associative memory. That is, when we think of a term (e.g., ”car”) the model presents what other terms may come to mind. These terms can be semantically related, for example, synonyms such as ”automobile” or hypernyms such as ”vehicle”, or they can be otherwise related terms such as ”tyre”, ”pavement” or ”Ferrari”.

The main component of the association network is the links between the terms. In order to capture the associations between the terms we need to weight the links. Timonen et al. (Timonen et al., 2011) used three components when weighting the associations: co-occurrence, age, and distance. In this work we use only co-occurrence. Age is not used as the age of the chemical safety cards is not relevant; i.e., older cards are as relevant as the new cards. Distance of the terms is also not used due to the type of documents we are using.

The confidence we use to model the weight of the association is calculated using Equation 1. The aim is to identify the co-occurring terms and give strong association when two terms co-occur often. For example, we aim to find links such as *Fire* - *Fumes* - *Toxic* as they co-occur often and may therefore be important in the search: query with ”Fire” could produce interesting documents if we include also ”Fumes”.

We create the association network using the keywords from the ontology. That is, we need the information from ontology creation as the documents we use should contain concepts (i.e., keywords) and their categories. For example, a document may hold: (**Danger** - *Toxic*, *Fire*, *Nausea*; **Material** - *Lead*, *Oxide*, *Crystals*), where Danger and Material are the categories. As the aim is to find co-occurring chemical attributes, we assess the co-occurrence among all categories. That is, we take all the attributes from all the categories and assess their confidence.

By using only the keywords from the ontology we get formatted terms that can easily be linked to the ontology. We also experimented with building the network from all the words found from the chemical safety cards but without proper keyword identification, this resulted a lot of noise in the network. The weight between the keywords is their co-occurrence. That is, if term A occurs 100 % of time with term B, the weight $w(A \rightarrow B) = 1.0$. If B occurs with A 50 % of time, the weight $w(B \rightarrow A) = 0.5$. Additional weighting components may be beneficial and will provide an interesting research topic for the future.

Table 1 presents a sample set of keywords and their association mappings. The last two (Sodium → Sodium Oxide, and Alcohol → Prenyl Alcohol)

Table 1: A sample of keyword associations.

Term (from)	Term (to)	Weight
Tetrahydrofuran	Diethylene Oxide	1.0
Prenyl Alcohol	Alcohol	1.0
Sodium Oxide	Oxide	1.0
Sodium Oxide	Sodium	1.0
Fumes	Fire	0.97
Vapour	Fire	0.80
Fire	Fumes	0.54
Sodium	Acid	0.51
Sodium	Powder	0.43
Sodium	Sodium Oxide	-
Alcohol	Prenyl Alcohol	-

demonstrate the anti-symmetric property of the network. That is, even though the link from A to B is strong, when B is too common, the link from B to A is too weak and therefore not included to the network.

We filtered the weakest associations (weight $<$ 0.35) from the network as they do not contribute to the query expansion. The resulting association network held approximately 150,000 associations.

4 CASE STUDY: ICSC SEARCH

In this section we describe the case study we performed for the search of International Chemical Safety Cards (ICSC). The idea is to implement a search engine that can fetch relevant safety cards for reference when new cards are being written.

4.1 Search Process

Search is usually initiated by providing the search terms; i.e., the query. These terms are usually properties of the chemicals, such as *flammable* or *colorless liquid*. In addition, the name of the chemical can be used if a specific chemical is needed.

The first step of the search process is to expand the query with additional and relevant query terms. The query expansion is performed as the search space is small and the original query terms often produce imperfect results. In our work, we combine the ontology and the association model, and use them for query expansion.

We use the models as follows: for each query term q_n in the query, the query is first expanded to the neighboring terms a_n from the association network. Then, for this expanded term list, query is further expanded by searching for related concepts o_n in the fuzzy ontology (Figure 3).

The query is expanded in the association network with a spreading activation technique (Crestani, 1997) using the best first search (Pearl, 1984). The best first search uses a function to select the top n nodes from the network by assessing the association between the query node q and the node k. Association between nodes q and k is the maximum value of the product of the weights in any of the paths from q to k. The function uses a threshold t_n to select the expanded nodes. If the weight from q to k is below the threshold, k is not used in the expansion. We use $t_n = 0.5$ which was selected after empirical evaluation of performance.

For example, consider Table 1 where the path from node *Vapour* to node *Fumes* is *Vapour – Fire – Fumes*. The weight between *Vapour – Fumes* in this case is $0.8 \times 0.54 = 0.43$. If $t_n = 0.5$, the node would not be added to the expanded list.

The query is reformulated to include all the q_n,a_n,o_n terms. This query is then used to search the database and produce the result set (Figure 4). Each resulting document is weighted based on the matching query terms. For example, if the expanded term is added to the query with the weight 0.5, it will contribute to the weight of the document with this weight. All the original search terms have the weight 1.0.

Figure 5 presents the query results page. The documents are printed in the result page in the descending order. The user can see which of the query terms produced the document as the result by highlighting the matching query terms. The expanded terms are also highlighted in the results.

We also implemented a search that can fetch similar chemical safety cards. That is, each card shown to the user has a ”search similar cards” link that is used when the user wants to find similar cards. We utilized the association model for the similarity assessment. The similarity between two documents can be assessed using *cosine similarity* that measures the cosine of the angle between them:

$$cos(d_n, d_m) = \frac{d_n \cdot d_m}{||d_n||||d_m||}. \quad (3)$$

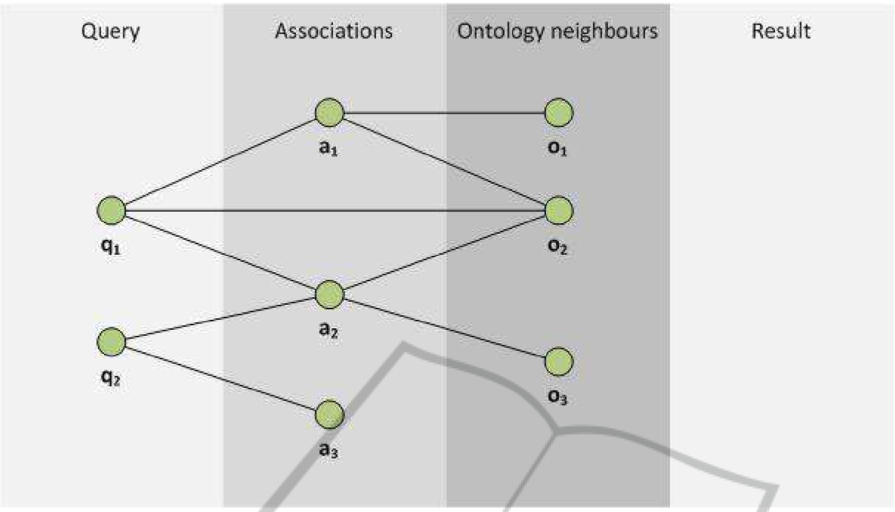


Figure 3: First, the query is expanded in the association network and then in the ontology.

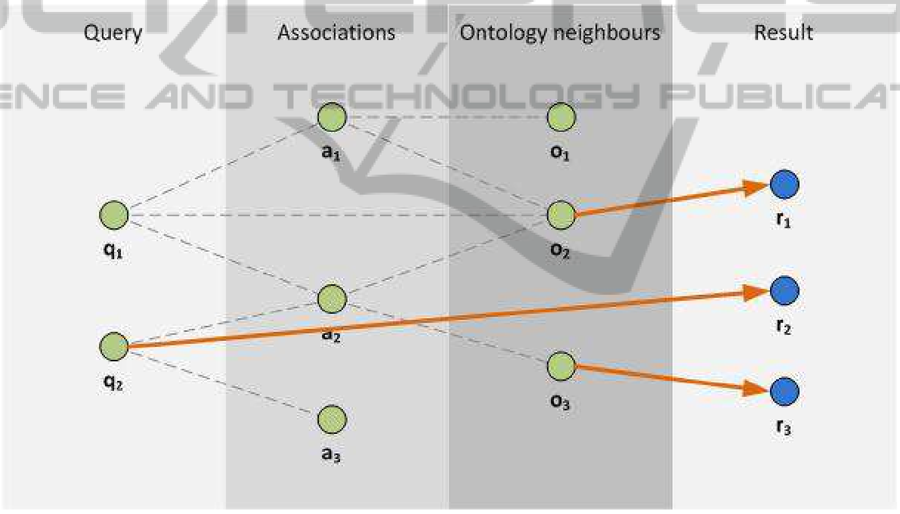


Figure 4: Query term q_2, and the expanded terms o_2 and o_3 produce the result set of r_1,r_2, and r_3.

Here, the dot product of d_n and d_m is the number of matching terms, and $\|d_n\|$ and $\|d_m\|$ is the length of the documents.

Instead of using the binary assessment of number of matching terms (where the term either matches or does not match), we use the association weight between the terms. That is, we assess the weight between the two terms in the network, as in query expansion. For example, the match between *Vapour* and *Fumes* is 0.43 instead of 0.0, when assessing the similarity between the documents. The match between the term n (in d_n) and the document d_m is the maximum association weight between n and all the terms in d_m.

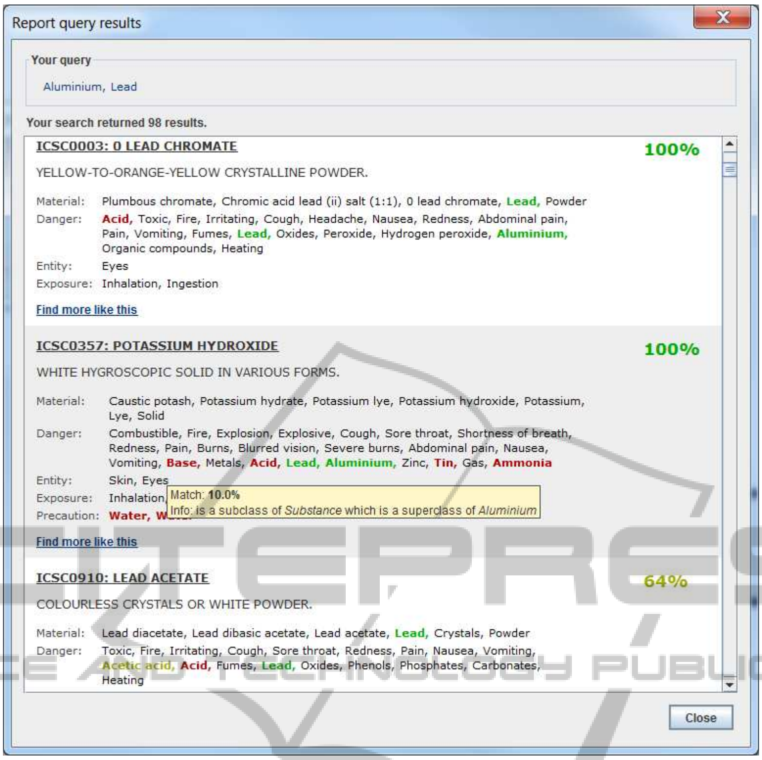
Figure 6 shows the result page for the similar chemical card search. The page shows the score of the document, which is percentage of matching terms weighted by the associations.

4.2 Search Results

The search produces a set of documents, i.e., chemical safety cards as the result. Each document is scored based on their relevance to the query. The results are shown in the descending order (Figure 5). The score for a document takes the number of the matching original query terms, and the weights of the matching keywords from query expansion:

$$w(d) = \frac{\sum_{q \in Q}\{q : q \in d\} + \sum_{e \in E}\{w(e) : e \in d\}}{|Q|}, \tag{4}$$

where Q is the set of original query terms, E is the set of query terms from query expansion, and $w(e)$ is the weight of the expansion term. If $w(d) > 1$, we use $w(d) = 1.0$.



Report query results

Your query

Aluminium, Lead

Your search returned 98 results.

ICSC0003: 0 LEAD CHROMATE — 100%

YELLOW-TO-ORANGE-YELLOW CRYSTALLINE POWDER.

Material: Plumbous chromate, Chromic acid lead (ii) salt (1:1), 0 lead chromate, **Lead**, Powder
Danger: **Acid**, Toxic, Fire, Irritating, Cough, Headache, Nausea, Redness, Abdominal pain, Pain, Vomiting, Fumes, **Lead**, Oxides, Peroxide, Hydrogen peroxide, **Aluminium**, Organic compounds, Heating
Entity: Eyes
Exposure: Inhalation, Ingestion

Find more like this

ICSC0357: POTASSIUM HYDROXIDE — 100%

WHITE HYGROSCOPIC SOLID IN VARIOUS FORMS.

Material: Caustic potash, Potassium hydrate, Potassium lye, Potassium hydroxide, Potassium, Lye, Solid
Danger: Combustible, Fire, Explosion, Explosive, Cough, Sore throat, Shortness of breath, Redness, Pain, Burns, Blurred vision, Severe burns, Abdominal pain, Nausea, Vomiting, **Base**, Metals, **Acid**, **Lead**, **Aluminium**, Zinc, **Tin**, Gas, **Ammonia**
Entity: Skin, Eyes
Exposure: Inhalation
Precaution: **Water**, W

Match: 10.0%
Info: is a subclass of *Substance* which is a superclass of *Aluminium*

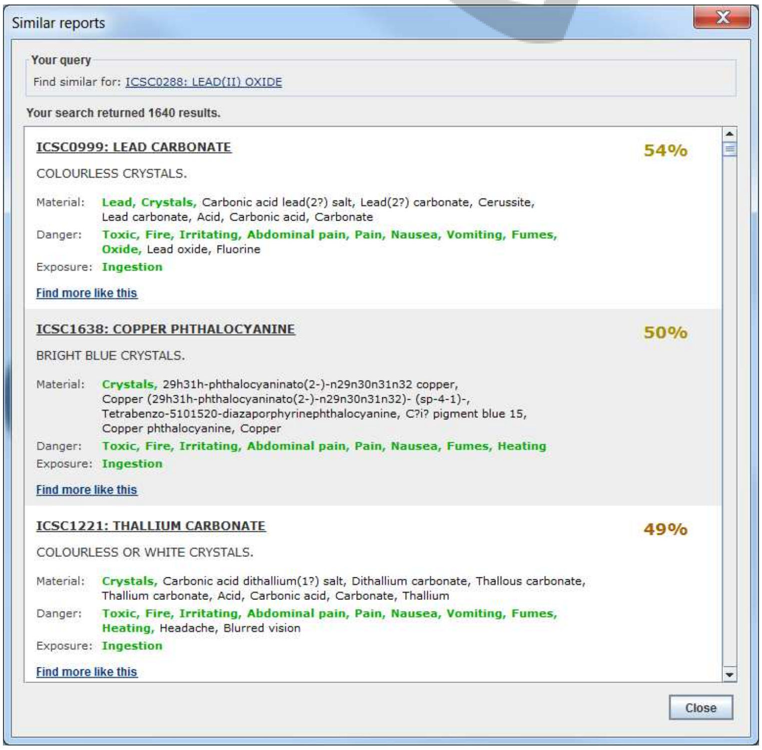
Find more like this

ICSC0910: LEAD ACETATE — 64%

COLOURLESS CRYSTALS OR WHITE POWDER.

Material: Lead diacetate, Lead dibasic acetate, Lead acetate, **Lead**, Crystals, Powder
Danger: Toxic, Fire, Irritating, Cough, Sore throat, Redness, Pain, Nausea, Vomiting, **Acetic acid**, **Acid**, Fumes, **Lead**, Oxides, Phenols, Phosphates, Carbonates, Heating

Figure 5: Search results for query ”Aluminium,Lead”.



Similar reports

Your query

Find similar for: ICSC0288: LEAD(II) OXIDE

Your search returned 1640 results.

ICSC0999: LEAD CARBONATE — 54%

COLOURLESS CRYSTALS.

Material: **Lead**, **Crystals**, Carbonic acid lead(2?) salt, Lead(2?) carbonate, Cerussite, Lead carbonate, Acid, Carbonic acid, Carbonate
Danger: **Toxic**, **Fire**, **Irritating**, **Abdominal pain**, **Pain**, **Nausea**, **Vomiting**, **Fumes**, **Oxide**, Lead oxide, Fluorine
Exposure: **Ingestion**

Find more like this

ICSC1638: COPPER PHTHALOCYANINE — 50%

BRIGHT BLUE CRYSTALS.

Material: **Crystals**, 29h31h-phthalocyaninato(2-)-n29n30n31n32 copper, Copper (29h31h-phthalocyaninato(2-)-n29n30n31n32)- (sp-4-1)-, Tetrabenzo-5101520-diazaporphyrinephthalocyanine, C?i? pigment blue 15, Copper phthalocyanine, Copper
Danger: **Toxic**, **Fire**, **Irritating**, **Abdominal pain**, **Pain**, **Nausea**, **Fumes**, **Heating**
Exposure: **Ingestion**

Find more like this

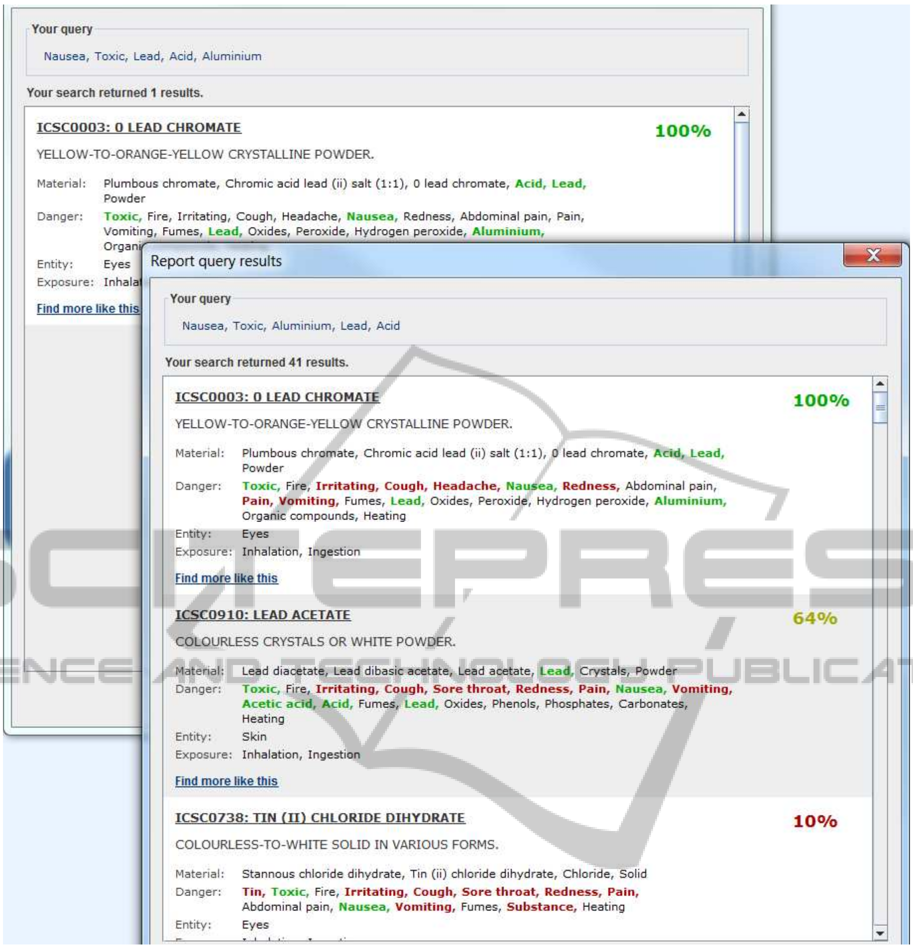
ICSC1221: THALLIUM CARBONATE — 49%

COLOURLESS OR WHITE CRYSTALS.

Material: **Crystals**, Carbonic acid dithallium(1?) salt, Dithallium carbonate, Thallous carbonate, Thallium carbonate, Acid, Carbonic acid, Carbonate, Thallium
Danger: **Toxic**, **Fire**, **Irritating**, **Abdominal pain**, **Pain**, **Nausea**, **Vomiting**, **Fumes**, **Heating**, Headache, Blurred vision
Exposure: **Ingestion**

Find more like this

Figure 6: Search results when searching for similar chemicals with ”Lead(II) Oxide”.



Your query

Nausea, Toxic, Lead, Acid, Aluminium

Your search returned 1 results.

ICSC0003: 0 LEAD CHROMATE 100%

YELLOW-TO-ORANGE-YELLOW CRYSTALLINE POWDER.

Material: Plumbous chromate, Chromic acid lead (ii) salt (1:1), 0 lead chromate, Acid, Lead, Powder

Danger: Toxic, Fire, Irritating, Cough, Headache, Nausea, Redness, Abdominal pain, Pain, Vomiting, Fumes, Lead, Oxides, Peroxide, Hydrogen peroxide, Aluminium, Organi

Entity: Eyes

Exposure: Inhalat

Find more like this

Report query results

Your query

Nausea, Toxic, Aluminium, Lead, Acid

Your search returned 41 results.

ICSC0003: 0 LEAD CHROMATE 100%

YELLOW-TO-ORANGE-YELLOW CRYSTALLINE POWDER.

Material: Plumbous chromate, Chromic acid lead (ii) salt (1:1), 0 lead chromate, Acid, Lead, Powder

Danger: Toxic, Fire, Irritating, Cough, Headache, Nausea, Redness, Abdominal pain, Pain, Vomiting, Fumes, Lead, Oxides, Peroxide, Hydrogen peroxide, Aluminium, Organic compounds, Heating

Entity: Eyes

Exposure: Inhalation, Ingestion

Find more like this

ICSC0910: LEAD ACETATE 64%

COLOURLESS CRYSTALS OR WHITE POWDER.

Material: Lead diacetate, Lead dibasic acetate, Lead acetate, Lead, Crystals, Powder

Danger: Toxic, Fire, Irritating, Cough, Sore throat, Redness, Pain, Nausea, Vomiting, Acetic acid, Acid, Fumes, Lead, Oxides, Phenols, Phosphates, Carbonates, Heating

Entity: Skin

Exposure: Inhalation, Ingestion

Find more like this

ICSC0738: TIN (II) CHLORIDE DIHYDRATE 10%

COLOURLESS-TO-WHITE SOLID IN VARIOUS FORMS.

Material: Stannous chloride dihydrate, Tin (ii) chloride dihydrate, Chloride, Solid

Danger: Tin, Toxic, Fire, Irritating, Cough, Sore throat, Redness, Pain, Abdominal pain, Nausea, Vomiting, Fumes, Substance, Heating

Entity: Eyes

Figure 7: Search results with and without the query expansion for the search ”Nausea,Toxic,Lead,Aluminium,Acid”.

We evaluated the models and their use for query expansion by manually assessing the search results. We compare the search results received both with and without the query expansion. We use a set of manually selected query terms. In addition, we evaluated the results of similar chemical safety card search.

We demonstrate an actual use of the system, consider a query with the following chemical attributes: *Nausea*, *Toxic*, *Lead*, *Aluminium*, *Acid*. The idea is that the chemical safety experts need information about chemicals that have these attributes in common with the new chemical.

When using these five query terms, we get only 1 chemical safety card when the query is not expanded. With expansion, there are 41 chemical safety cards in the result set. In this demonstration the query expansion has shown clear benefits as the search produces more information. Figure 7 presents this search case. The highest scoring document is the same in both cases but the expansion adds several other potentially relevant cards to the result set. The highest scoring expanded card has the score of 64 %. As this card (*Lead Acetate*) is quite similar with the highest scoring card (*0 Lead Chromate*) we consider this result very good.

In an experiment of 20 different searches, the top scoring cards in the result set are the same for both expanded and non-expanded queries. However, the expansion includes documents to the result set that otherwise would not be there. In addition, when the non-expanded search would produce no documents, the expansion almost always produces some results.

When assessing the feasibility of the documents received with expansion, the score for the document indicate the relevancy reasonably well. That is, if the score for the document is high, it is more likely a relevant document. These results are promising and indicate that the proposed query expansion technique and the domain models address the issues faced with the small search space. We leave the further assessment of the feasibility of the included documents and the comparison against competing approaches out of scope of this paper.

To demonstrate the similarity search we use two chemical safety cards: *Lead Arsenite*, and *Graphite (Natural)*. We picked these two chemicals as we have some background knowledge about them. The most similar chemical card for *Lead Arsenite* was *Lead (II) Arsenite*, and for *Graphite (Natural)* was *Carbon*. Even without expertise in the area, we can see that these two chemicals are relevant considering the original chemical safety cards. To experiment the similarity search, we conducted the search for 20 different chemicals. In the experiment, 90 % of the highest scoring documents were considered very relevant. Even though the number of queries performed is small, we get a good indication of the feasibility of the approach.

5 CONCLUSIONS

In this paper we have proposed a novel approach for domain modeling that combines a fuzzy ontology and an association network. We use the fuzzy ontology to describe the hierarchical relationship of chemicals and their attributes whereas the association network describes the associations between the attributes of the chemicals. The model is used for query expansion in the chemical domain.

We also use the models to find similar chemical safety cards. Here, the aim is to use the associations to weight the cosine similarity assessment and find which chemicals have similar attributes.

The proposed models produced promising results. The query expansion performed as expected as it was able to produce more information than a search where the expansion was not used. In addition, the association weighted similarity assessment was able to find chemical safety cards that we consider relevant. Overall, the search engine performed better when the query expansion and the similarity assessment are used. This is crucial when considering the real world applications.

The main challenge for the future is to lessen the burden of the ontology creation without impacting its capability to capture the concepts of the given domain. Learning the association network is an easier task but identifying the keywords and terms to be used in the network also requires some work. In addition, the maintenance of both models is time consuming. In the future, it may be beneficial to utilize the association network in ontology creation as the network may hold valuable information about the domain and the relationships of the terms. In order to achieve this we would need to utilize an approach for keyword identification from documents. In conclusion, we believe that by combining the association network with the ontology we are able to create a richer model of the domain that can be better utilized in different types of applications.

ACKNOWLEDGEMENTS

Authors wish to thank the Finnish Funding Agency for technology and Innovation (TEKES) for funding this research and the anonymous reviewers for the valuable feedback.

REFERENCES

Agrawal, R., Imielinski, T., and Swami, A. N. (1993). Mining association rules between sets of items in large databases. In *Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data (SIGMOD'93), USA*, pages 207–216.

Bhogal, J., Macfarlane, A., and Smith, P. (2007). A review of ontology based query expansion. *Information Processing & Management*, 43(4):866 – 886.

Bordogna, G. and Pasi, G. (2000). Application of fuzzy set theory to extend boolean information retrieval. *Studies in Fuzziness and Soft Computing*, 50:21–47.

Carlsson, C., Brunelli, M., and Mezei, J. (2010). Fuzzy ontology and information granulation: an approach to knowledge mobilisation. *Information Processing and Management of Uncertainty in Knowledge-Based Systems. Applications*, pages 420–429.

Carpineto, C. and Romano, G. (2012). A survey of automatic query expansion in information retrieval. *ACM Computing Surveys*, 44(1):1:1–1:50.

Crestani, F. (1997). Application of spreading activation techniques in information retrieval. *Artificial Intelligence Review*, 11(6):453–482.

Cross, V. (2004). Fuzzy semantic distance measures between ontological concepts. In *Fuzzy Information, 2004. Processing NAFIPS'04. IEEE Annual Meeting of the*, volume 2, pages 635–640. IEEE.

Formica, A., Missikoff, M., Pourabbas, E., and Taglino, F. (2008). Weighted ontology for semantic search. *On the Move to Meaningful Internet Systems: OTM 2008*, pages 1289–1303.

Hirvonen, J., Tommila, T., Pakonen, A., Carlsson, C., Fedrizzi, M., and Fullér, R. (2010). Fuzzy keyword ontology for annotating and searching event reports. In *KEOD 2010 - Proceedings of the International Conference on Knowledge Engineering and Ontology Development, Valencia, Spain, October 25-28, 2010*, pages 251–256.

Janowicz, K., Raubal, M., and Kuhn, W. (2012). The semantics of similarity in geographic information retrieval. *Journal of Spatial Information Science*, (2):29–57.

Laskey, K. J. and Laskey, K. B. (2008). Uncertainty reasoning for the world wide web: Report on the URW3-XG incubator group. *URW3-XG, W3C*.

Mihalcea, R. and Tarau, P. (2004). Textrank: Bringing order into text. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing, (EMNLP'04), A meeting of SIGDAT, a Special Interest Group of the ACL, held in conjunction with ACL 2004, 25-26 July 2004, Spain*, pages 404–411.

Mostafavi, S., Ray, D., Warde-Farley, D., Grouios, C., Morris, Q., et al. (2008). Genemania: a real-time multiple association network integration algorithm for predicting gene function. *Genome Biol*, 9(Suppl 1):S4.

Parry, D. (2006). Fuzzy ontologies for information retrieval on the www. *Capturing Intelligence*, 1:21–48.

Pearl, J. (1984). *Heuristics - intelligent search strategies for computer problem solving*. Addison-Wesley series in artificial intelligence. Addison-Wesley.

Peña-Castillo, L., Tasan, M., Myers, C. L., Lee, H., Joshi, T., Zhang, C., Guan, Y., Leone, M., Pagnani, A., Kim, W. K., et al. (2008). A critical assessment of mus musculus gene function prediction using integrated genomic evidence. *Genome Biol*, 9(Suppl 1):S2.

Sanchez, E. and Yamanoi, T. (2006). Fuzzy ontologies for the semantic web. *Flexible Query Answering Systems*, pages 691–699.

Tetko, I. V. (2002a). Associative neural network. *Neural Processing Letters*, 16(2):187–199.

Tetko, I. V. (2002b). Neural network studies. 4. introduction to associative neural networks. *Journal of chemical information and computer sciences*, 42(3):717–728.

Thomas, C. and Sheth, A. (2006). On the expressiveness of the languages for the semantic webmaking a case for a little more. *Capturing Intelligence*, 1:3–20.

Timonen, M. (2013). *Term Weighting in Short Documents for Document Categorization, Keyword Extraction and Query Expansion*. PhD thesis, University of Helsinki, Faculty of Science, Department of Computer Science.

Timonen, M., Silvonen, P., and Kasari, M. (2011). *Modelling a Query Space Using Associations*, volume 255 of *Frontiers in Artificial Intelligence and Applications: Information Modelling and Knowledge Bases XXII*. IOS Press.

W3C Recommendation (2004). OWL Web Ontology Language. http://www.w3.org/TR/owl-features/.

Widyantoro, D. H. and Yen, J. (2001). A fuzzy ontology-based abstract search engine and its user studies. In *Fuzzy Systems, 2001. The 10th IEEE International Conference on*, volume 3, pages 1291–1294. IEEE.

Yang, L., Ball, M., Bhavsar, V., and Boley, H. (2005). Weighted partonomy-taxonomy trees with local similarity measures for semantic buyer-seller matchmaking.

Zhang, K., Tang, J., Hong, M., Li, J., and Wei, W. (2006). Weighted ontology-based search exploiting semantic similarity. *Frontiers of WWW Research and Development-APWeb 2006*, pages 498–510.