MINING NON-TAXONOMIC CONCEPT PAIRS FROM UNSTRUCTURED TEXT

A Concept Correlation Search Framework

Mei Kuan Wong, Syed Sibte Raza Abidi
Faculty of Computer Science, Dalhousie University, 6050 University Avenue, Halifax, NS, Canada

Ian D. Jonsen
Department of Biology, Dalhousie University, 1459 Oxford Street, Halifax, NS, Canada

Keywords: Correlation Search, Non-taxonomic Relations, Association Rule Mining, Lift Interestingness Measure.

Abstract: Ontology consists of concepts, taxonomic relations and non-taxonomic relations. The majority of the ontology learning tools focus on discovering concepts and taxonomic relations. Very little effort has been put on discovering non-taxonomic relations. In this paper, we present a concept correlation search framework to discover non-taxonomic concept pairs from unstructured text. Our framework features the (a) extraction of correlated concepts beyond ordinary search window size of a single sentence; (b) use of lift as interestingness measure for association rule mining; (c) harness of 2- itemsets association rules from n-itemsets association rules where n>2; and (d) identification of non-taxonomic concept pairs based on existing domain ontology. The proposed framework has been tested with the Fisheries Oceanography journals, and the results demonstrate significant improvements over traditional association rule approach in search of non-taxonomic concept pairs.

1 INTRODUCTION

Ontologies serve as semantic representations of domain-specific knowledge, and are used for knowledge sharing, interoperability and reuse (Shamsfard & Barforoush 2003). An ontology consists of a set of concepts or classes, C, which is taxonomically related by the transitive, IS-A relation $H \in C \times C$ and non-taxonomically related by named object relation $R* \in C \times C \times String$.

Developing ontologies—i.e. ontology engineering—is a tedious process that demands a sound understanding of the domain and the ability to abstract and model the knowledge. In recent years, ontology engineering has been pursued by 'learning' the ontology from domain-specific documents. Ontology learning from text involves the application of natural language processing, text analysis and logical reasoning methods to capture knowledge—i.e. domain concepts, relationships between concepts, descriptions of concepts—from documents to serve as the building blocks of an ontology. This approach leads to a reduction in the time, effort and manpower required in the ontology engineering process.

Typically, the existing ontology learning tools focus on discovering concepts and their taxonomic relations from texts. However, the extraction of non-taxonomic relations, which are an integral aspect of an ontological description of a domain, is not well-researched (Sánchez & Moreno 2008). An example of a non-taxonomic relation is the relation *cure* between the concept pairs of *doctor* and *patient*. Current research on discovering non-taxonomic relations is pursued based on (a) statistical approach and (b) semantic analysis approach. Semantic analysis approaches rely on lexico-syntactic patterns to discover relations between a pair of co-occurring concepts. Statistical approaches involve studying the distributional properties of words in order to determine the salient concepts and then use correlation measures between concepts to establish potential non-taxonomic relations between them. Association rule mining is a popular statistical

Kuan Wong M., Sibte Raza Abidi S. and D. Jonsen I..
MINING NON-TAXONOMIC CONCEPT PAIRS FROM UNSTRUCTURED TEXT - A Concept Correlation Search Framework.
DOI: 10.5220/0003482707070716
In *Proceedings of the 7th International Conference on Web Information Systems and Technologies* (WTM-2011), pages 707-716
ISBN: 978-989-8425-51-5

method to extract non-taxonomic relations, and is used in ontology learning tools such as Text2Onto (Cimiano et al. 2005) and OntoLearn (Velardi et al. 2005). These ontology learning tools use association rule mining with traditional confidence measure to extract non-taxonomic relations. However, there are noted limitations about confidence measure is that it (a) is sensitive to the frequency of the concepts in the data set and may return pairs of concepts even if there is no association between them, and (b) suffers from rare itemset problem whereby even if an association rule representing an important relationship between concepts exists but since it is rare it is pruned altogether (Sheikh et al. 2005).

In this paper, we pursue the extraction of concept pairs, from unstructured text, that has a non-taxonomic relation between them. We present a concept correlation search framework that employs a statistical approach that is an extension to the traditional association rule mining approach used in ontology learning tools for non-taxonomic relation extraction. Our approach to search for correlated concepts has three distinct elements: (i) we investigate the use of the *lift measure* (Sheikh et al. 2005), as opposed to the traditional support and confidence measures, to establish the interestingness between correlated concepts. The key advantage of our use of the lift measure is that it determines how many times more often concept X and concept Y occurs together than expected if they were statistically independent. Lift does not suffer from the rare item problem (Sheikh et al. 2005); (ii) when searching for correlated concept pairs we look beyond the traditional one-sentence window to include multiple adjacent sentences. Our approach is based on the observation that quite often scientific authors discuss correlated concepts across multiple sentences, therefore we search correlated concepts across two adjoining sentences; (iii) we employ a domain ontology, as background knowledge, to filter out the correlated concepts that have a taxonomic relationship between them. This leaves us with a set of non-taxonomic concept pairs that serve as candidates for non-taxonomic relations during ontology learning. We apply our framework to search for non-taxonomic concept pairs for the domain of marine biology—we worked with 374 Fisheries Oceanography journal publications over a period of 10 years (1999-2008). We extracted 130 concept pairs out of which 108 non-taxonomic concept pairs were identified. The results were validated by domain experts.

2 LITERATURE REVIEW-RELATED WORK

Ontology learning involves Machine Learning (ML) and advance Natural Language Processing (NLP) technologies, starting from term extraction and concept definition to more complex tasks such as learning taxonomic and non-taxonomic relations. In this section, we review the state-of-the-art in ontology learning tools specific to non-taxonomic relation extraction.

From a statistical perspective, the pioneer research work in non-taxonomic relation extraction was performed by Maedche & Staab (2000) using association rule mining. Subsequently, ontology learning tools such as Text2Onto (Cimiano et al. 2005) and OntoLearn (Velardi et al. 2005) also approach the non-taxonomic relation extraction task from the statistical point of view using association rule mining with traditional confidence measure.

Hasti (Shamsfard & Barforoush 2004), another ontology learning tool, extracts non-taxonomic relations from the semantic analysis point of view. Hasti combines logical, linguistic-based, template driven and semantic analysis methods in their non-taxonomic relation extraction. A hybrid of both approaches is taken by RelExt (Schutz & Buitelaar 2005) in their non-taxonomic relation extraction where relevant terms and verbs are extracted from a given text collection. Then, a combination of both linguistic and statistical processing is used to compute relations between them. The problem with these methods is that they are dependent on sentence structure. Thus, the search window size for correlated concepts is short and constrained to a single sentence. Short search window size used often proves to be deficient in discovering relations (Chagnoux et al. 2008).

From the literature review, it is clear that ontology learning, especially the extraction of non-taxonomic relations from unstructured text is a challenging, yet much pursued area. Our work is an extension to the traditional association rule mining used in some of the abovementioned tools. We pursue to look beyond single-sentence window and use lift as the interestingness measure to yield interesting concept pairs that represent potential non-taxonomic relations in ontology learning context.

3 OUR CONCEPT CORRELATION SEARCH FRAMEWORK

In order to extract non-taxonomic concept pairs from unstructured text, we propose a concept correlation search framework, which consists of four phases: *text preprocessing*, *concept extractor*, correlated *concept search* and *concept pair classifier* (see Figure 1).

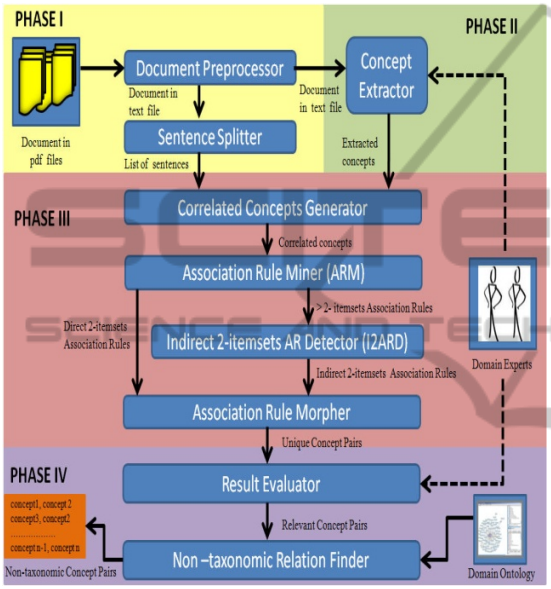

Figure 1: Functional design of our concept correlation search framework.

Phase I begins with processing the collection of text documents to extract the sentences within the documents. Phase II extracts domain concepts from the sentences. Phase III takes as input the sentences (in the order they appear in the document) and the extracted domain concepts to find correlated concept pairs. In Phase IV we measure the relevancy of the extracted correlated concept pairs to identify the concept pairs that are relevant to the domain, and then we use background knowledge (a domain ontology) to identify the non-taxonomically related concept pairs.

The distinct aspects of our approach are: (a) In Phase III our correlated concept generator searches for correlated concept pairs beyond the traditional one-sentence window. This allows the potential correlation of important concepts that are spread across two adjoining sentences thus yielding a larger set of correlated concept pairs; (b) In Phase III, we apply the lift interestingness measure to association rule mining to assess the degree to which the concept pairs are of interest within our context; (c) In Phase III, we make use of the association rules with more than 2 itemsets whereby we derive indirect 2-itemsets association rules. This is important as most of the previous work tends to ignore these rules while solving the non-taxonomic relations extraction problem; (d) In Phase IV, we engage domain experts to evaluate the relevancy of the concept pairs extracted; and (e) In Phase IV, we leverage a domain ontology to distinguish taxonomic concept pairs from non-taxonomic concept pairs.

In the next few sections, we explain the methods developed for each processing phase.

3.1 Phase I: Text Preprocessing

This phase involves the processing of the unstructured text document which is in the form of Portable Document Format (PDF). The PDF files are converted to text files using pdf2Text, an open-source software that converts PDF documents into text files. We also remove non-essential information from the text files such as the headers (journal title, author information, etc.) and footers (acknowledgements, references, etc.). The processed text files are then combined into a single text file. The resulting output is then (a) used in Phase II for concept extraction and (b) further processed using a sentence splitter developed in Perl to split the text into a list of sentences. A total of 74,280 sentences were produced from a total of 374 Fisheries Oceanography journals in PDF files.

3.2 Phase II: Concept Extractor

In this phase, our main objective is to identify key domain concepts from the processed text file in Phase I for the domain being investigated. We use KEA (Keyphrase Extraction Algorithm) to extract key phrases for the document. In KEA (Witten et al. 1999), the commonly used information retrieval method, the tf-idf weight (term frequency-inverse document frequency) is used to rank the key phrases. As not all key phrases generated by KEA are domain specific, we engaged domain experts in this phase to manually evaluate the generated key phrases. The key phrases produced by KEA were shown to the domain experts to determine their relevancy to the domain. The relevant key phrases are then used to represent key domain concepts in the domain ontology. A total of 102 domain concepts were selected from the top 200 key phrases generated by

KEA. These concepts are then used as candidates in Phase III in order to find correlations between them.

3.3 Phase III: Correlated Concept Search

In this phase, we pursue the search for concepts that are deemed to be correlated. These correlated concept pairs will further be candidates for non-taxonomic relations. We have developed four tools to search for correlations between concepts as follows: (i) correlated concept generator; (ii) association rule miner; (iii) indirect 2-itemsets association rule detector; and (iv) association rule filter (see Figure 1). We explain the tools developed for each task in the next few sections.

3.3.1 Task I: Correlated Concept Generator

Based on the list of sentences generated in Phase I and the extracted concepts generated in Phase II, the first task in correlation search is to search for all tightly correlated concepts. In our proposed framework, we differentiate our correlation search by extending the search window size to multiple adjacent sentences (see Figure 2). By doing so, firstly, we are able to generate more correlated concepts and secondly, we are able to minimize the number of missing concepts by combining all adjacent stand alone concepts.

In order to locate correlated concepts in multiple adjacent sentences, we devised a correlated concept generator. First, we run through all the single sentences to locate all salient concepts. Subsequently, we perform three iterations over the list of sentences to combine multiple adjacent sentences as follows:

ITERATION 1: Combine two consecutive sentences if each adjoining sentence consists of a single concept;

ITERATION 2: Merge two consecutive sentences if the first sentence consists of a single concept while the adjoining sentence consists of more than one concept. This is the look forward strategy.

ITERATION 3: Repeat ITERATION 2 but using the look backward strategy. In look backward strategy, we work from the end to the start of the text file.

An example of the execution of the correlated concept generator is shown in Figure 2. We noted that stand alone concepts such as *a, b, c* and *d* found in one-sentence window are captured in our extended window size approach. In addition, more correlated concept pairs are generated in our approach by combining the multiple adjacent sentences.

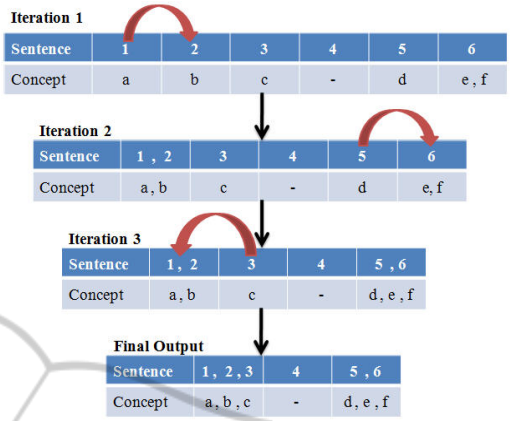

Figure 2: An example of our correlated concept generator execution.

We examine the distribution of domain concepts for the Fisheries Oceanography journal within a single sentence and multiple adjacent sentences in order to determine the salient concepts (see Table 1). In quest for correlated concept pairs, sentences with more than one salient concept are desirable. By using original approach of one-sentence window, we noted that only 36.3% of the sentences consist of more than one salient concept. This means that the search base for correlated concept pairs is constrained to one third of the whole document. In addition, concepts that are found solely in sentences with one salient concept will be lost in the process of extracting the correlated concept pairs.

Table 1: Number of sentences with salient concepts versus search window size.

Search Window Size	Number of sentences with 1 salient concept (%)	Number of sentences with > 1 salient concept (%)
Single Sentence (Original Approach)	23,725 (63.7%)	13,540 (36.3%)
Multiple Sentences(Our Approach)	11,615 (39.6%)	17,713 (60.4%)

Interestingly, in our proposed approach, the percentage of sentences with more than one salient concept has increased drastically from 36.3% to 60.4% (see Table 1). This indicates that our proposed method of finding correlated concepts within multiple adjacent sentences is capable of returning more correlated concept pairs. In addition,

the potential for losing adjacent stand alone concepts is reduced substantially.

3.3.2 Task II: Association Rule Miner

In the second task, we mine for correlated concept pairs through the use of association rule. Association rule mining is a data mining technique that identifies data or text elements that co-occur frequently within a dataset (Agrawal et al. 1993). An association rule describes the association among items in which when some items are purchased in a transaction, others are purchased too. The problem in association rule mining can be represented as follows:

> A transaction T supports an itemset X if X is contained in T. The support for an itemset X is defined as the ratio of the number of transactions that supports the itemset X to the total number of transactions. If the support for an itemset X satisfies the user specified minimum support threshold, then X is called frequent itemset.

In our context, items are concepts while transactions are sentences. It can be represented as $X \Rightarrow Y$, in which X is an antecedent and Y is a consequent of this rule, and X and Y are two itemsets. However, in our quest for correlated concept pairs, we treat rule $X \Rightarrow Y$, to be equivalent to rule $Y \Rightarrow X$.

Association rule mining typically results in large amounts of redundant rules. Due to the large amount of redundant rules; various measures have been developed to help in evaluating the interestingness of the association rules. Some of the existing ontology learning tools such as Text2Onto (Cimiano et al. 2005) and OntoLearn (Velardi et al. 2005) use traditional confidence measure in extracting non-taxonomic relations. The confidence of a rule $X \Rightarrow Y$ is defined as the ratio of the support for the itemsets $X \cup Y$ to the support for the itemset X. If itemset $Z = X \cup Y$ is a frequent itemset and the confidence of $X \Rightarrow Y$ is no less than the user-specified minimum confidence, then the rule $X \Rightarrow Y$, is an association rule. As mentioned by (Sheikh et al. 2005), support-confidence framework suffers from rare itemset problem. Yet, rare itemset in an association rule may represent an important relationship exist between concepts. It is therefore important, from an ontology learning standpoint, to recognize all these rare itemsets.

In our proposed framework, we therefore use lift as the interestingness measure for association rules. Lift allows to measure how many times more often X and Y occurs together than expected if they were statistically independent. Lift does not suffer from the rare itemset problem (Sheikh et al. 2005). The lift measure is defined over $[0,\infty]$ and can be interpreted as follows:

$$\text{Lift}(X, Y) \begin{cases} = 1, \text{ if X and Y are independent;} \\ > 1, \text{ if X and Y are positively correlated;} \\ < 1, \text{ if X and Y are negatively correlated.} \end{cases}$$

In our search for correlated concept pairs, lift values greater than 1 is desirable. Typically, the higher the lift value, the more likely that occurrence of X and Y together, is not just random occurrence, but, because of some relationships occur between them.

In our experiment, we use Weka, an open source data mining software to perform the association rule mining (Hall et al. 2009). Association rules generated in this task are further categorized into 2 groups based on the number of itemsets present in each association rule: (a) direct 2-itemsets association rules and (b) n-itemsets association rules where $n > 2$ (see Table 2). The group of n-itemsets association rules are then used as candidates in Task III to further derive more indirect 2-itemsets association rules.

Table 2: Number of association rules generated by Weka using support-lift framework.

Description	Single sentence	Multiple sentences
Total number of association rules	60	156
Number of direct 2-itemsets association rules	50	105
Number of n-itemsets association rules (n >2)	10	51

3.3.3 Task III: Indirect 2-itemsets Association Rule Detector (I2ARD)

The aim of this phase is to further extract indirect 2-itemsets association rules from n-itemsets association rules where n >2. In order to achieve this objective, we have developed an indirect 2-itemsets Association Rule Detector (I2ARD) using Perl. The main idea behind this detector is to employ the *anti-monotone constraint*, which means that, if an itemset *I* satisfies the constraint, so does any of its subset (Sheikh et al. 2005). In our I2ARD, we use the anti-monotone constraint to generate indirect 2-itemsets sub association rules from association rules with more than 2-itemsets. An example of our I2ARD is shown in Figure 3.

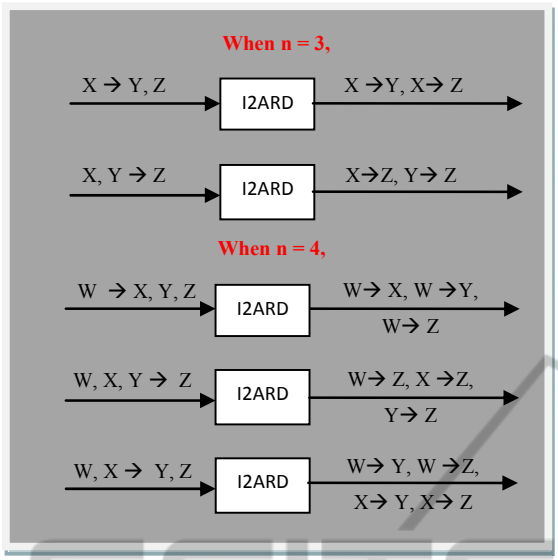

Figure 3: Examples of our Indirect 2-itemsets Association Rule Detector (I2ARD) when n =3 and n =4.

In Table 3, we can see that our I2ARD is capable of producing a substantial number of indirect 2-itemsets association rules from n-itemsets association rules where n > 2 for both original approach and our approach. We noted that the number of indirect 2-itemsets association rules generated for both approaches doubled the original number of n-itemsets association rules (n>2). This can be attributed to the nature of the text collection, in which the association rule miner returned association rules with maximum of 3-itemsets. For each 3-itemsets association rules, our I2ARD produced two indirect 2-itemsets association rules. Some of these indirect 2-itemsets association rules may have existed as direct 2-itemsets association rules discovered earlier by the association rule miner. These redundant rules indicate that the rules produced using our I2ARD are interesting and can be considered for potential concept pairs.

Table 3: Number of indirect 2-itemsets association rules generated by our I2ARD.

Description	Single sentence	Multiple sentences
Number of n-itemsets association rules (n>2)	10	51
Number of indirect 2-itemsets association rules	20	102

3.3.4 Task IV: Association Rule Filter

In this task, our objective is to aggregate all direct 2-itemsets association rules discovered in Task II with all indirect 2-itemsets association rules discovered in Task III. In the process of aggregation, we eliminate all redundant and symmetric rules. The rationale of having all symmetric rules removed is that, we treat rule X=> Y and rule Y=> X the same in our concept pair extraction. The resulting output is a list of unique concept pairs that are not redundant. We applied the association rule filter on both the original approach of one-sentence window as well as our approach of extending the search window size to multiple adjacent sentences. Table 4 shows the number of unique concept pairs produced in Phase III. We noted that our approach of combining multiple sentences is capable of generating more than double the number of unique concept pairs generated by the original approach of one-sentence window.

Table 4: Number of unique concept pairs.

Description	Single sentence	Multiple sentences
Number of direct 2-itemsets association rules	50	105
Number of indirect 2-itemsets association rules	20	102
Number of unique concept pairs	57	130

3.4 Phase IV: Concept Pair Classifier

The final phase of our concept correlation search framework is to make an assessment on the correlated concept pairs found in Phase III. Our two stages classifying strategy is to (a) classify the correlated concept pairs to *highly related*, *related* and *not related*; and (b) further classify the *highly related* and *related* concept pairs to taxonomic concept pairs and non-taxonomic concept pairs.

3.4.1 Stage I: Result Evaluator

Evaluation of ontology relationships learning systems against any gold standard is notoriously difficult as there are not many gold standards that are available for evaluation (Gulla et al. 2009). Nevertheless, there is always another option in which domain experts are engaged in performing the manual evaluation. In this stage, we have engaged 2 domain experts to rate the suggested relationships independently. We presented all unique concept pairs found in Phase III to the experts for their

review. Each expert was asked to rank the concept pairs as *highly related* (there is definitely a relationship between the two concepts with a numerical score of 1.0), *related* (there is probably a relationship between the two concepts, score of 0.5) or *not related* (these two concepts are not related, score of 0).
Based on domain experts' feedback for each concept pairs, we computed the average score for each concept pair and determined its relevance as shown in Table 5.

Table 5: Score range matrix.

Ranking	Score Range
Highly Related	> 0.5
Related	0.5
Not Related	< 0.5

For our purpose of non-taxonomic concept pair extraction, we are only interested in *highly related* and *related* concept pairs.

3.4.2 Stage II: Non-taxonomic Concept Pair Finder

We further classify the *highly related* and *related* concept pairs identified in Stage I into taxonomic concept pairs and non-taxonomic concept pairs. Our approach is based on the adoption of domain ontology and exploiting the knowledge in the taxonomy to isolate all taxonomic relations. If any of the *highly related* and *related* concept pairs are found in the domain ontology having super-class and sub-class relations, these concept pairs are then classified as taxonomic concept pairs. The remaining concept pairs would then be classified as non-taxonomic concept pairs.

4 EVALUATION AND DISCUSSION

In this section, we present the evaluation results for various methods developed for Phase III and Phase IV of our concept correlation search framework.

4.1 Evaluating Correlated Concept Generator

Table 6 presents the number of *highly related* and *related* concept pairs for different search window size used in the correlated concept generator as discussed in Section 3.3.1. The ranking of the concept pairs found is determined based on evaluation by domain experts. It is interesting to note that our approach of using multiple adjacent sentences as search window size offered an addition of 81.25% of *highly related* concept pairs and an addition of 153.85% of *related* concept pairs as compared to the traditional approach of using a single sentence as search window size. This vindicates our proposed approach of extending the search window size, and also confirms the assumption that short search window size of a single sentence is deficient in extracting non-taxonomic relations.

Table 6: Number of concept pairs versus different approaches.

Approach	Number of Highly Related Concept Pairs	Number of Related Concept Pairs
Single Sentence (Original Approach)	32	13
Multiple Sentences (Our Approach)	58	33
% Increase	81.25%	153.85%

4.2 Examining Support-lift Framework

To demonstrate the effectiveness of our choice of the interestingness measure for association rules, we compared the association rules generated by two different interestingness measures—i.e. (a) the support-confidence and (b) the support-lift. In the experiment, we use minimum support value of 0.01, minimum confidence value of 0.1 and minimum lift value of 1.01. Table 7 displays that our proposed approach of using support-lift framework produces higher percentage of both *highly related* and *related* concept pairs in comparison to the original approach. Thus, this technique is proven to be useful for relation extraction in which a significant lift value can be more important than a high confidence rule (Alvarez 2003).

Table 7: Support-confidence measure versus support-lift measure.

Interestingness Measure	% of Highly Related Concept Pairs	% of Related Concept Pairs
Support-Confidence (Original Approach)	44.85%	36.76%
Support-Lift (Our Approach)	49.23%	36.92%
% Change	4.38%	0.16%

We went on further to investigate the intersection of association rules produced by different frameworks. The rationale behind this investigation is to assess

the significance of rules produce by each framework. If there is a substantial number of overlapping rules found, this indicates that each framework is capable of generating significance rules. Figure 4 shows the intersection of association rules produced by our support-lift framework against the popular support-confidence framework.

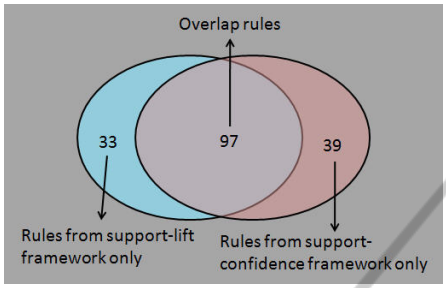

Figure 4: Rules produce by various frameworks.

Interestingly, we noted that 97 concepts pairs are found in both frameworks and 88.7% of them are ranked as either *highly related* or *related* (see Table 8). Besides, we also studied rules generated solely by each framework. The rules found in support-lift framework alone seem to be of higher relevance in comparison to rules found in support-confidence framework alone. This implies that correlations generated by the support-lift approach are of more relevance to the domain experts as compared to the correlations generated by the support-confidence approach.

Table 8: Relevancy of association rules produce by various frameworks.

Framework	% of Highly Related Concept Pairs	% of Highly Related Concept Pairs
Overlap Rules	51.55%	37.11%
Rules from Support-Lift Framework Only	42.42%	36.36%
Rules from Support-Confidence Framework Only	30.77%	33.33%

4.3 Evaluating Indirect 2-itemsets Association Rule Detector (I2ARD)

Evaluation of I2ARD involved examination of ranking on the indirect 2-itemsets association rules. Out of the 25 indirect 2-itemsets association rules detected, 21 of these rules are ranked as *highly related* or *related*. By employing our I2ARD, the number of *highly related* concept pairs increase by 10.34% whereas the number of *related* concept pairs increase by 45.45% (see Table 9). This signifies the importance of mining indirect 2-itemsets association rules from n-itemsets association rules where n>2.

Table 9: Association rule miner versus association rule miner + I2ARD.

Approach	Number of Highly Related Concept Pairs	Number of Related Concept Pairs
Association Rule Miner only	58	33
Association Rule Miner + I2ARD	64	48
% Increase	10.34%	45.45%

4.4 Evaluating the Concept Correlation Search Framework

In order to evaluate our proposed concept correlation search framework, we compare the outcome of our framework with the outcome of a baseline approach. The main difference between the baseline approach and our concept correlation search framework is in the components in Phase III of our framework (see Table 10).

Table 10: Baseline approach versus our approach.

Tasks	Baseline Approach	Our Approach
Search window size	Single sentence	Multiple sentence
Association Rule Miner	Yes	Yes
Indirect 2-itemsets Association Rule Detector (I2ARD)	No	Yes

Table 11 exhibits the number of *highly related* concept pairs and *related* concept pairs discovered in both approaches. The results indicate that our approach is capable of extracting a significantly larger number of *highly related* concept pairs and more than double the number of *relevant* concept pairs in comparison to the baseline approach.

Table 11: Comparing our concept correlation search framework against baseline.

Approach	Number of Highly Related Concept Pairs	Number of Related Concept Pairs
Baseline Approach	32	13
Our Approach	64	48
% Increase	100.00%	269.23%

4.5 Identifying the Non-taxonomic Relations

We use the taxonomy of existing domain ontology to distinguish taxonomic concept pairs from non-taxonomic concept pairs for all the *highly related* and *related* concept pairs found using our proposed framework. With the knowledge that taxonomic relation exists between a super class and a subclass represented in the domain ontology, if a concept pair is not found to be having an ancestral relation with each other, it can be regarded as non-taxonomic concept pair (given the knowledge captured within the domain ontology). From our experiment, we have identified 4 taxonomic concept pairs and 108 non-taxonomic concept pairs out of a total of 130 concept pairs extracted (see Table 12).

Table 12: Number of concept pairs versus types.

Type	Number of Highly Related Concept Pairs	Number of Related Concept Pairs
Non-taxonomic concept pairs	60	48
Taxonomic concept pairs	4	0

Table 13 shows some examples of the taxonomic concept pairs and non-taxonomic concept pairs for the domain being studied. Since the fisheries oceanography domain is a marriage between the fisheries domain and the oceanography domain, it is interesting to note that our proposed approach is capable of finding correlated concept pairs within each domain and across both domains. Concept pairs across both domains are highlighted in **bold** (see Table 13). These concept pairs are of special interest to the domain experts as they provide clues to the domain experts on the potential interactions between both domains.

Table 13: Example of concept pairs generated from the Fisheries Oceanography journals.

Taxonomic Concept Pairs	Non-taxonomic Concept Pairs
fish, capelin fish, salmon salmon, pacific salmon summer, seasons	fish, length **fish, temperature** **fish, depth** **summer, migration** **summer, production**
	salinity, depth
	temperature, depth

5 CONCLUDING REMARKS

In this paper, we presented a concept correlation search framework to extract non-taxonomic concept pairs from unstructured text, and applied it to the marine biology domain. The novel features of our framework are that: (a) we search for correlated concept pairs within multiple adjacent sentences. This is an extension to the traditional approach of using a single sentence in search of correlated concept pairs; (b) we apply the lift interestingness measure to association rule mining to assess the degree to which the concept pairs are of interest within our context; and (c) we derive new correlations between pairs of concepts from n-itemsets association rules where n>2. Our results show that these features generate more and better concept pairs in comparison to existing ontology learning tools that use traditional association rule mining to mine non-taxonomic relations. Our framework also distinguishes non-taxonomic concept pairs from taxonomic concept pairs using background knowledge existing in domain ontology. These non-taxonomic concept pairs will further be candidates for non-taxonomic relations extraction in ontology learning. Our framework is domain-independent and will bring us a step closer towards the semi-automation of non-taxonomic relation extraction in support of ontology learning.

As future line of research, we intend to work on another least tackled problem in ontology learning, which is the labelling of the non-taxonomic concept pairs, i.e. to employ linguistic structure approach to determine the most appropriate verb that connect the correlated concept pairs.

ACKNOWLEDGEMENTS

This research is supported by a R&D grant from CANARIE, Canada through the Network Enabled Platform program. We would also like to extend our gratitude to Dr. Isidora Katara for her valuable help in the evaluation of the proposed framework.

REFERENCES

Agrawal, R., Imieliński, T. & Swami, A. 1993, "Mining association rules between sets of items in large databases", ACM SIGMOD Record, vol. 22, no. 2, pp. 207-216.

Alvarez, S. A. 2003, "Chi-squared computation for association rules: Preliminary results",

Comput.Sci.Dept., Boston College, Chestnut Hill, MA, Tech.Rep.BC-CS-2003-01.

Chagnoux, M., Hernandez, N. & Aussenac-Gilles, N. 2008, "An interactive pattern based approach for extracting non-taxonomic relations from texts", *Workshop on Ontology Learning and Population* (associated to ECAI 2008)(OLP), University of Patras, Patras, Greece, pp. 1–6.

Cimiano, P., Pivk, A., Schmidt-Thieme, L. & Staab, S. 2005, "Learning taxonomic relations from heterogeneous sources of evidence", Ontology Learning from Text: Methods, evaluation and applications, pp. 59–73.

Gulla, J. A., Brasethvik, T. & Kvarv, G. S. 2009, "Association Rules and Cosine Similarities in Ontology Relationship Learning", Enterprise Information Systems, pp. 201-212.

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P. & Witten, I.H. 2009, "The WEKA data mining software: An update", ACM SIGKDD Explorations Newsletter, vol. 11, no. 1, pp. 10-18.

Maedche, A. & Staab, S. 2000, "Discovering conceptual relations from text", ECAICiteseer, pp. 321.

Sánchez, D. & Moreno, A. 2008, "Learning non-taxonomic relationships from web documents for domain ontology construction", *Data & Knowledge Engineering*, vol. 64, no. 3, pp. 600-623.

Schutz, A. & Buitelaar, P. 2005, RelExt: A Tool for Relation Extraction from Text in Ontology Extension.

Shamsfard, M. & Barforoush, A. A. 2003, "The state of the art in ontology learning: a framework for comparison", The Knowledge Engineering Review, vol. 18, no. 04, pp. 293-316.

Shamsfard, M. & Barforoush, A. A. 2004, "Learning ontologies from natural language texts", *International Journal of Human-Computer Studies*, vol. 60, no. 1, pp. 17-63.

Sheikh, L., Tanveer, B. & Hamdani, M. 2005, "Interesting measures for mining association rules", *Multitopic Conference*, 2004. Proceedings of INMIC 2004. 8th International IEEE, pp. 641.

Velardi, P., Navigli, R., Cucchiarelli, A., Neri, F., Buitelaar, P., Cimiano, P. & Magnini, B. 2005, "Evaluation of OntoLearn, a methodology for automatic learning of domain ontologies", Ontology Learning from Text: Methods, evaluation and applications, pp. 92–106.

Witten, I. H., Paynter, G. W., Frank, E., Gutwin, C. & Nevill-Manning, C. G. 1999, "KEA: Practical automatic keyphrase extraction", *Proceedings of the 4th ACM conference on Digital libraries ACM*, pp. 254.