A Review on Ontology Learning Approaches of Creating a Topic Map of Cybercrime Research

Kijung lee

University of Cincinnati, School of IT, Cincinnati OH, 45221, USA

kijung.lee@uc.edu

ABSTRACT

Conducting an academic research requires getting a firm grasp of ongoing research issues as well as locating research materials effectively. Often research in different fields on a similar topic can assume diverse approaches due to different objectives and research goals in their own fields. Especially in an interdisciplinary research field like cybercrime, many research topics overlap with those of other research fields. Researchers in such a field, therefore, can benefit from understanding the related domains of one's own research. Topic maps provide methods for understanding research domain and managing relevant information resources at the same time. In this paper, we review a topic map solution to acquire knowledge structure and to locate information resources effectively. We address current problems of cybercrime research, review previous studies that use automated methods for topic map creation, and examine existing sets of methods for automatically extracting topic map components. Especially, the methods we discuss here are text mining techniques for extracting ontology components, denoted as ontology learning.

Keywords: Cybercrime, Ontology learning, Topic Maps, Text mining

1. INTRODUCTION

Many human activities in modern society involve use of computer systems. As a consequence, computer related social phenomena such as cybercrime have increased in the number of cases and diversified in the form of crimes. Research in relation to cybercrime has burgeoned in various issues such as hacking, cracking, spam, computer security, privacy invasion, viruses, and piracy, to list a few.

Well-structured and value-added cybercrime research requires getting a firm grasp of primary research topics and domain concepts of the cybercrime. In the current practice of cybercrime research, however, the concepts of cybercrime are fuzzy, and research topics are not well-organized. First, terms are interchangeably used. Cybercrime, computer crime, internet crime, digital crime, and many other similar terms are used to describe similar concepts. Moreover, one term, like cybercrime, can be used to mean different concepts. It may mean crime within cyber space in some context, while crime in a real world facilitated through cyber tools in another. Second, cybercrime research is driven by industry or social issues in relation to cybercrime that emerges every day. Therefore, most research results in case by case solutions rather than a grand theory which may account for classes of phenomena. Largely, the problems of cybercrime research can be summarized to 1) lack of grand knowledge scheme, 2) concept incoherence among different research bodies, and 3) difficulty of identifying research topic distribution and locating relevant information resources. In order to tackle such issues, we consider a topic map approach of knowledge structure representation and information resource organization.

Topic maps organize large and heterogeneous sets of information resources and build a network of structured semantic link over the resources [1, 2]. In knowledge management perspective, topic maps provide ways to represent knowledge structure, enabling researchers to understand how the research topic is formed and to navigate among the research topics for proper information resources [e.g., 3, 4-6]. In information management perspective, topic maps can offer methods to represent the documents properly for effective retrieval of information resources for any given research topic [e.g., 7, 8, 9]. Other research in topic map concerns education [10, 11], topic map query [12], visualization [13], and semantic web [14, 15]. Topic maps provide, in summary, methods for understanding research domain and managing relevant information resources at the same time.

In constructing a topic map for a set of information resources, however, human intervention is unavoidable. Manual tasks are needed in selecting topics, identifying their occurrences and associations. Such need may be acceptable when the topic maps are used merely for navigation purpose over a relatively small collection of information resources. However, 'navigation' is not the only primary function of a topic map. Moreover, the volume of information resource in cybercrime research is generally large enough to prevent manual construction of topic maps. The applicability of topic maps can be expanded through some kind of automatic process during the construction. In this project, we review methods of knowledge discovery from text in a semi-automatic fashion. Denoted as ontology learning, the methods can suggest candidate components of a cybercrime topic map, e.g., topics, topic types, topic occurrences, and associations.

Ontology learning is a series of research activities that encompass development of methods, methodologies, tools, and algorithms to acquire and learn ontologies in a semi-automatic fashion. Especially, ontology learning from text concerns automated ontology engineering using the unstructured textual input data from the support of linguistic analysis, statistical analysis, and machine learning. A typical approach in ontology learning from text first involves term extraction from a domain specific corpus through a statistical process that determines their relevance for the domain corpus at hand. These are then clustered into groups with the purpose of identifying taxonomy of potential classes. Subsequently, also relations can be identified by computing a statistical measure of 'connectedness' between identified clusters. In many cases, part-of-speech tagger is run against a domain corpus and patterns are identified by domain experts, or linguistic analysis can be applied in other cases.

The objective of this paper is to address current problems of cybercrime research, to review previous studies that use automated methods for topic map creation, and to examine existing sets of methods for automatically extracting topic map components. Especially, the methods we discuss here are text mining techniques for extracting ontology components, denoted as ontology learning. It is important to discuss a method of automatic topic map extraction since, in current days, a major portion of research materials are published both physically and electronically, and, in some cases, only electronically. Managing affluence of digital documents requires some type of automated methods. By reviewing ontology learning techniques and identifying mapping information between domain ontology and domain topic map, this literature review will serve as a guideline for further research in building a topic map using ontology learning techniques and tools.

This paper is organized as follows; Section 2 presents problems of cybercrime research and, as a potential solution, topic map is suggested to overcome the problems. We, then, overview the topic map data model and review how the topic map data model can help improve cybercrime research. Section 3 reviews automated methods for creating a topic map. Section 4 reviews ontology learning methods as a way to create a topic map in a semi-automatic fashion. Section 5 covers immanent problems and questions in regards to ontology learning methods for topic map creation, discusses future works and directions, and summarizes the paper.

2. CYBERCRIME RESEARCH

Locating relevant cybercrime research materials is difficult without a precise understanding of cybercrime related knowledge. In the course of cybercrime research, one may find that idea of term usage and topic distribution of the subject of their interest is the key to finding relevant materials from research databases and/or search engines. . The primary matter of discussion in this section is two-fold; identifying problems of cybercrime research and suggesting a potential solution to tackle the problems. First, we address the problems in conducting cybercrime research using research databases, difficulty in identifying core concepts of domain and research topic distribution, and locating relevant information resources. Then, we review a topic map solution in terms of topic map data model and how it can contribute to the solution of addressed problems.

2.1. Problems in current cybercrime research environment

Problems in conducting cybercrime research can be addressed in several aspects; 1) domain concepts are not clearly defined, 2) research has been conducted based on industry issues rather than organized theme of research topics, and, as a consequence, 3) finding research materials is difficult without precise conceptualization and understanding of research subject at hand.

First, domain concepts are not well defined as a foundation of research. Although the term "cybercrime" is widely recognized as a concept with negative connotation, many researchers have hard time defining it with generally accepted, precise expressions. Gordon and Ford [16], in their endeavor to define and classify cybercrime, find that cybercrime significantly lacks in the clarity in its common usage. They present a framework of cybercrime definition and categorization on the linear continuum where one end represents more people aspects of crimes, whereas the other end represents more technology aspects of crimes, According to their argument, cyberstalking and cyberterrorism fall under people aspects of crimes, while phishing and denial of service attack are technology aspects of crimes. Given this framework, however, can we really define what cybercrime is? One may ask questions like; "What is phishing?" or "What is cyberstalking?" Levels of required concepts for proper definition of cybercrimes can be enormous. In other words, describing cybercrime in a limited scope of research domains is not easy since issues of cybercrime are widely related to variety of research domain such as crimes and criminals, business impacts, law enforcement related issues, information technology related issues, and, of course, various legal issues.

Second, research topics in cybercrime are not well structured and represented. Research topics of cybercrime, in conjunction with rapidly emerging core concepts in relation to cybercrime, have evolved and accumulated. Lu, Jen, and Chang [17] show that the trend of cybercrime research has chronologically evolved based on the paradigm change. In their bibliometric analysis, cybercrime research from 1961 to 1974 focuses on the implementation issues. From 1975 to 1993, research issues are about institutionalization which can be dissected into innate aspects of computer crimes in the beginning and legislation issues in the later. From 1991 to 2000, research issues are in relation to legislation and security. Finally, from 2000 to 2004, issues are specialized and diverse, such as law enforcement, software piracy, hacker behavior, victims of computer crime, the fight against money laundering, computer crime forensics, and information security. The study of Lu, Jen, and Chang may not be explaining the full aspects of cybercrime research trends in various research domains since their data is from one source of academic database (i.e., web of science) using a non-exhaustive set of keywords search (i.e., 'cybercrime', 'cyber crime', 'computercrime', and 'computer crime'). However, it can be inferred that cybercrime research has been evolved to cover more complex issues over time. Keeping track of such trend of research topics and analyze them for future research will enable academic researchers to be able to provide industry standards for future trend instead of conducting industry issue based cybercrime research.

Third, finding relevant research materials in cybercrime is difficult. Without a standardized method of search term recommendation and research topic guidance, one may get search results with too many false positives or true negatives. In database searches using a keyword "cybercrime", the ISI Web of Science results in 93 outcomes, the ACM Digital library results in 92 records, and the Google Scholar results in 5,300 records. However, when a keyword "phishing" is used, the ISI Web of Science finds 44 records, the ACM digital library finds 263 records, and the Google Scholar results in 5,180 records. Intuitively, concept of 'phishing' should be categorized under the concept of 'cybercrime' when we only compare those two concepts. Therefore, phishing papers should appear in the search results of cybercrime papers. The simple search tasks indicate that research materials are not organized according to the hierarchical relationship of cybercrime – phishing. Although we still need to verify the contents of searched documents, only based on the numeric values of the search results, it can be highly suspected that information resources in research databases are disproportionately represented and organized.

The reviewed factors above about cybercrime research may not be well-practiced in the current environment. We argue, however, that it is a fundamental course of cybercrime research to defining cybercrime related concepts, to track research topics for such concepts, and to organize information resources based on the

identified research topics. Moreover, such activities may lead the cybercrime activities to more solid ground of science based on a few aspects. First, they will provide researchers with a common language, necessary for sound collaboration and meaningful discussion. Second, they will help determine the scope of the problem to be addressed, and are necessary for clear communication about a subject. Therefore, such support for the study of cybercrime is needed, for effective communication in the short term, and, for societal needs in cybercrime prevention in the longer term.

2.2. How a topic map can benefit cybercrime research

What we describe in the previous section about cybercrime research may be expressed in a simple term, 'information overload'. In the information science stance, a way to reduce the problem of information overload is to organize domain knowledge coherently, at the same time providing effective and efficient way of locating relevant research materials. As a potential solution to such challenge, a topic map data model can be used. Topic maps organize large and heterogeneous sets of information resources, and build a structured semantic link network over the resources [1, 2, 9]. Therefore, it will be able to provide a method to build a nomenclature for knowledge share and information management. The main use of such system framework would benefit cybercrime researchers in many different fields, e.g., legal researchers, law enforcement professionals, corporate IT counselors, and also, general public.

2.2.1. Topic map Data Model

A topic map data model is a representation of structured knowledge network over a set of information resources. In general, the structural information conveyed by topic maps includes topics, occurrences, and associations [1]. Topics represent the subjects of stories that a given information resource bear. Occurrences are groupings of addressable information resources around topics and associations. Associations are the relationship between topics. A topic map defines a multidimensional topic space; a *topic* has one or more *names* within a *scope*; it can also have *occurrences* and may play a *role* as a member of zero or more *associations*. Topics, associations and occurrence may have a *type* which also is a topic. Figure 1 describes conceptualization of a topic map data model.

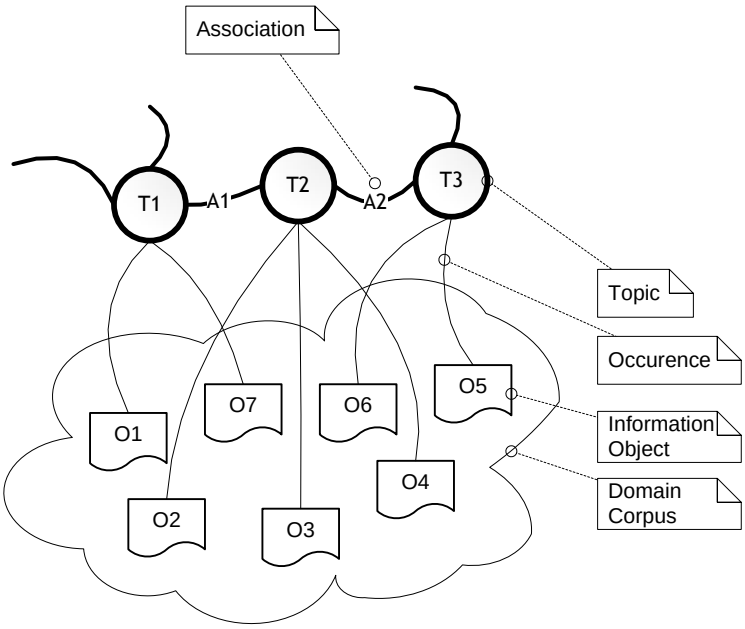

Figure 1 Conceptualization of a Topic Map Data Model

A topic map is an SGML (or XML) document in which different element types, derived from the basic set of architectural forms, are used to represent topics, occurrences of topics, and relationships between topics.

Figure 2 below shows an example XML syntax of a topic map declaring topic map element, topic element, association element and occurrence element. The topic map components are extracted manually from the given abstract of a paper by Liu, Deng, Huang, and Fu [18];

The authors' proposed antiphishing strategy uses visual characteristics to *identify* potential phishing sites and *measure* suspicious pages' similarity to actual sites registered with the system.

The structure basically identifies topics, 'Visual Characteristics' and 'Page Similarity', occurrence, a link to paper information page under the ISI web of science database, association, 'measure', and the association roles as in 'method' and 'data pattern'.

```
<!-- A topic representing the visual characteristics.-->

<topic id="vis_characteristics">
<instanceOf><topicRef xlink:href="#antiphishing strategy"/></instanceOf>
  <baseName>
    <baseNameString>Visual Characteristics</baseNameString>
</baseName>

<!-- An occurrence given by ISI Web of Science information page of the paper -->

  <occurrence>
    <instanceOf>
      <topicRef xlink:href="#html-format"/>
    </instanceOf>
    <resourceRef
xlink:href="http://apps.isiknowledge.com/full_record.do?product=WOS&search_mode=GeneralSearc
h&qid=1&SID=3APpemOJFFbG82BOg4M&page=1&doc=1"/>
  </occurrence>

</topic>

<!-- A topic representing "Page Similarity" -->

<topic id="p_similarity">

  <baseName>
    <baseNameString>Page Similarity</baseNameString>
  </baseName>

</topic>

<!-- An association representing an measurement relationship -->

<topic id="measure">
  <baseName>
    <baseNameString>measure</baseNameString>
  </baseName>
</topic>

<!-- Used here to associate Visual Characteristics and Page Similarity -->

<association>
  <instanceOf><topicRef xlink:href="#measures"/></instanceOf>
  <member>
    <roleSpec><topicRef xlink:href="#method"/></roleSpec>
    <topicRef xlink:href="# vis_characteristics "/>
  </member>
  <member>
    <roleSpec><topicRef xlink:href="#data pattern"/></roleSpec>
    <topicRef xlink:href="#p_similarity"/>
  </member>
</association>
```

Figure 2 XML Syntax of a Topic Map Data Model (Syntax structure was borrowed from[19])

Conceptually, the topic map data model above can be represented in a diagram, following the same notation in Figure 1, as indicated in Figure 3 below.

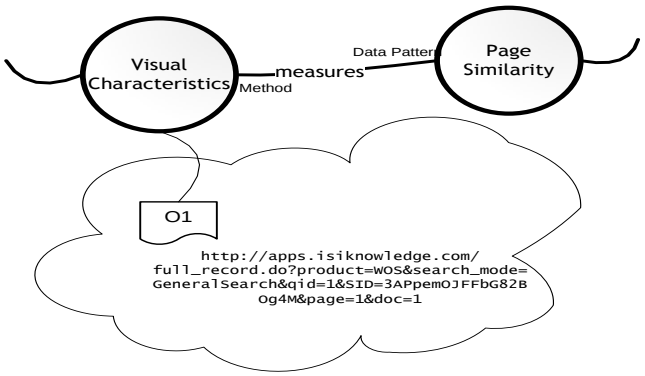

Figure 3 Conceptualization of a Phishing Research Topic Map Data Model

2.2.2. Applications of a topic map

Topic maps enable multiple, concurrent views of sets of information objects. The structural nature of these views is unconstrained; they may reflect an object oriented approach, or they may be relational, hierarchical, ordered, unordered, or any combination of them. Moreover, an unlimited number of topic maps may be overlaid on a given set of information resources [1]. Therefore, topic maps can be used in various situations of knowledge representation and information management. Among many applications, some primary ones are [1];

- Navigation among information objects,
- Navigation among concepts and topics,
- Information filtering based on creation of views, and
- Organization of unstructured information resources.

In the cybercrime research context exemplified above, a topic map will be able to provide researchers with methods to navigate from a paper about visual characteristics to a paper about page similarity by linking documents with domain concepts. Researchers also can navigate from a phishing research topic to a visual characteristics research. Information resources about page similarity can be filtered based on a facet, 'kinds of data' to get documents only in relation to web page similarity. As a consequence, seemingly unstructured information resources are organized for better visibility and findability. In sum, a topic map can be a decent candidate solution for supporting cybercrime research.

3. AUTOMATED METHODS OF TOPIC MAP GENERATION

Generation of a topic map usually requires professional skills and abilities. Expensive manual tasks are needed in topic selection, occurrence identification, and association identification. However, the volume of information resources in cybercrime research is generally so large that manual constructions of topic maps require assistance from computing powers. In this section, we discuss various approaches of automated topic map generation and explore possibilities of using text mining techniques for automated topic map generation.

3.1. Related studies: (Semi) automated methods of topic map generation

Human intervention is a necessary component in creating topic maps. Largely, semi-automatic topic map creation methods can be categorized into heuristics based methods and mathematic and/or procedural algorithm based methods which require different degrees of human intervention. We organize existing studies based on the expert intervention which can be categorized into heuristic based vs statistical algorithm based methods.

3.1.1. Methods with heuristics and rules

Rath [19] presents a framework for automatic generation of topic maps based on a 'topic map template' and a set of generation rules. The input information for the automatic generation is the topic map template, the address for information resources in the repository, metadata about information resources, and structure of information resources. The framework consists of information resource repository, topic map template, and a set of generation rules. According to their method, first, topics and occurrences are created based on a set of rules (Section 4.2);

If resource fulfills metadata <condition> and/or

contains structure <element> in <context> containing <content>,

then create a topic of <type> with name derived from metadata <field> or

name derived from <element> in <context> and

create occurrence to resource with <role>

Then, associations are created. The automatic process interprets the constraints and extracts the information regarding what topic types play what association role in what association. With these facts, the process is able to generate the lists containing the association types, the association role types, and the candidates for referenced topics. Although the author does not demonstrate the algorithm in the paper, he addresses that association algorithms should be specific to information resources in design. This method should demonstrate simpler way of automatic topic map generation. However, it requires great deal of human intervention since the rules and the template are to be constructed explicitly and manually.

Roberson and Dicheva [20] propose a set of heuristics that can be used for extracting semantic information from the HTML markup of the web pages. Heuristics 1 and 2 concern extraction of page topics; each web page has a unique topic and parsed topics within the webpage are sub-topics of the webpage topic. Heuristic 3 concerns use of anchor tag for naming of the page topics; the user-defined name in the anchor tag is the topic name for the target page. Heuristics 4, 5, and 6 concern use of heading tags; level information of headings can define topic – sub-topic hierarchy. Heuristic 7 concerns list element tags; list tags and lst element tags can form topic – sub-topic hierarchy. Heuristic 8 and 9 concern table tags; topics extracted from a same row in a table are related while topics extracted from a same column are in topic – sub-topic hierarchy. Finally, heuristic 10 instructs that topics from same HTML elements are related. Although their experiment shows an interesting result, their algorithm is based on the assumption that HTML pages are coherently organized. Moreover, the algorithm may be useless for web pages constructed through server side techniques.

Although the heuristic based approach provides straightforward method of identifying topic map components, it requires great amount of preprocessing on the input data if they do not conform to the set of predefined rules. Automatic generation of a topic map with unstructured input data is possible when such knowledge may be discovered from the underlying set of information resources through an automated process, which is generally known as *knowledge discovery from texts* or *text mining*. Moore (2000), while emphasizing on the value of human intervention, also argues that the automatic generation of topic maps is a useful first step in the construction of a topic map.

3.1.2. Methods with text mining algorithms

Hofmann [21] presents a probabilistic approach which combines a statistical model-based analysis with a topological visualization principle. His method can be used to derive topic maps which represent topical information by characteristic keyword distributions arranged in a two-dimensional spatial layout. The proposed method in the paper largely has two parts; latent semantic analysis and topology preservation. A latent semantic analysis technique for text collections models context-dependent word occurrences to map

documents and words to a more suitable representation in a probabilistic latent semantic space. A principle of topology preservation allows visualizing the extracted information. Experiments were conducted using the TDT1 collection (i.e., Topic Detection and Tracking, distributed by the Linguistic Data Consortium with 49,225 transcribed broadcast news stories) and CLUSTER (i.e., a collection of 1,568 abstract of research papers on clustering). A pyramidal visualization of the CLUSTER collection based on a 256 factor is generated. One can see that meaningfully coarsened maps can be obtained from the 16 x16 map, different areas like astronomy, physics, databases, and pattern recognition. A similar map hierarchy for the TDT1 collection is generated. Different topics and events can be identified from the word distributions. Sub-topics like the ones dealing with different events of international politics are mapped to neighboring positions on the lattice.

In Böhm and collegues [22], text corpus analysis with linguistic and statistical analysis algorithms, an infrastructure for text mining is described which uses collocation analysis as a central tool. In this paper, the term collocation is used for two or more words with the next neighbors and sentences. Their strategies are topic map bootstrapping by corpus comparison and topic map optimization. In the process of topic amp bootstrapping, topic candidates are generated by running a comparative analysis of a domain-specific corpus against a reference corpus. Significant concepts are filtered and used as starting points for topic map generation. A given topic map is enriched by selecting relevant collocations for the topics already in the map, thus enlarging the map.

Yang and Lee [23], in their work, design two feature maps for knowledge acquisition and representation; document cluster map and word cluster map. A document cluster map is clustering information of documents in the corpus. In the document cluster map, each neuron represents a document cluster that contains several similar documents with high word co-occurrence. In contrast, word cluster map represents cluster information of words in the domain corpus. Each neuron in this map represents a cluster of words that reveal the general concept of the corresponding document cluster associated with the same neuron in the document cluster map. Based on the two maps, the procedure of their algorithm is (p. 310);

Step 1. Find the neuron with the largest supporting cluster similarity. Select this neuron as dominating neuron.

Step 2. Eliminate its neighbor neurons so that they will not be considered as dominating neurons.

Step 3. If there is no neuron left or the number of dominating neurons exceeds a predetermined value, stop. Otherwise go to Step 1.

As a result, all identified topics in every layer of the hierarchy can be used as **topics**. Parent topic can be used as the **type** of its child topics. Since a topic is a word labeled to a neuron in the word cluster map, its **occurrences** can be assigned as the documents labeled to the same neuron in the document cluster map. A topic is **associated** with the other if there existed a path between them in the projected hierarchy. A topic in a neuron can be **associated** with topics in neighbor neurons.

Abramowicz and collegues [24] present a method consisting of Term Crawling (used for association identification) and Clustering Hierarchy Projection (for topic hierarchy identification), which are applied to build a topic map based on free text documents from local repositories and those downloaded from the Internet. Based on the technical set, they perform topic map generation in a sequence, i.e., data collection → base dictionary construction → term crawling → hierarchy projection → affinity grouping → RDF modeling. After the base dictionary has been created from a domain corpus, term crawling mechanism is used to identify and store relations between terms. Each term is taken from the dictionary, and related terms are searched. Using clustering, large document collections are divided into smaller, meaningful parts. A relationship is represented by RDF triple consisting of subject, property, and value.

Methods described in this section are highly related to text mining techniques. Although rule based approach may offer more precise interpretation and discovery of underlying knowledge structure, our approach should

be based on the methods dealing with unstructured data, since there are no guarantees that cybercrime related publications are properly structured for data processing.

4. ONTOLOGY LEARNING APPROACHES OF GENERATING A CYBERCRIME TOPIC MAP

In a semiotic perspective, words are the form of symbolic representation that contains meaning in the combination of objects and concepts [25]. Although the symbols may not capture the perfect nuance of the object or the concept, there is correspondence among them and, thus, research in textual analyses has been borrowing this theoretical framework. Knowledge discovery from text, in this sense, is a way to model knowledge based on written resource of information. In technical terms, natural language texts exhibit morphological, syntactic, semantic, pragmatic and conceptual constraints that interact in order to convey a particular meaning to the reader. The textual data is associated with conceptual structures of the text creator and the reader learns the creator's conceptual structures from the interacting constraints given through language. In this section, first, ontology learning techniques are reviewed and categorized by relevant text mining techniques and described in relation to automatic topic map creation. Then, ontology learning tools and systems are reviewed.

4.1. General framework of ontology learning

In most cases, an ontology learning framework involves a series of text mining techniques, i.e., term extraction, concept extraction, taxonomy extraction, and relation extraction. We present a broad idea of an ontology learning framework for creating a topic map in Figure 4. In Figure 4, ontology learning tasks identified are collections of existing approaches. It is not intended to illustrate that all tasks should be used for proper generation of a topic map but they list a set of techniques that are available for each phase in generating a topic map. One assumption in the framework is that concept (i.e, class) in an ontology representation is analogous to topic in a topic map representation. This can be supported by the fact that topic map, in practice, can be interchangeably used with semantic network, concept map, mind map, and other types of knowledge representation schema that has basic structure of concept – relation. Based on the framework, we review relevant approaches and existing studies for each step in the framework.

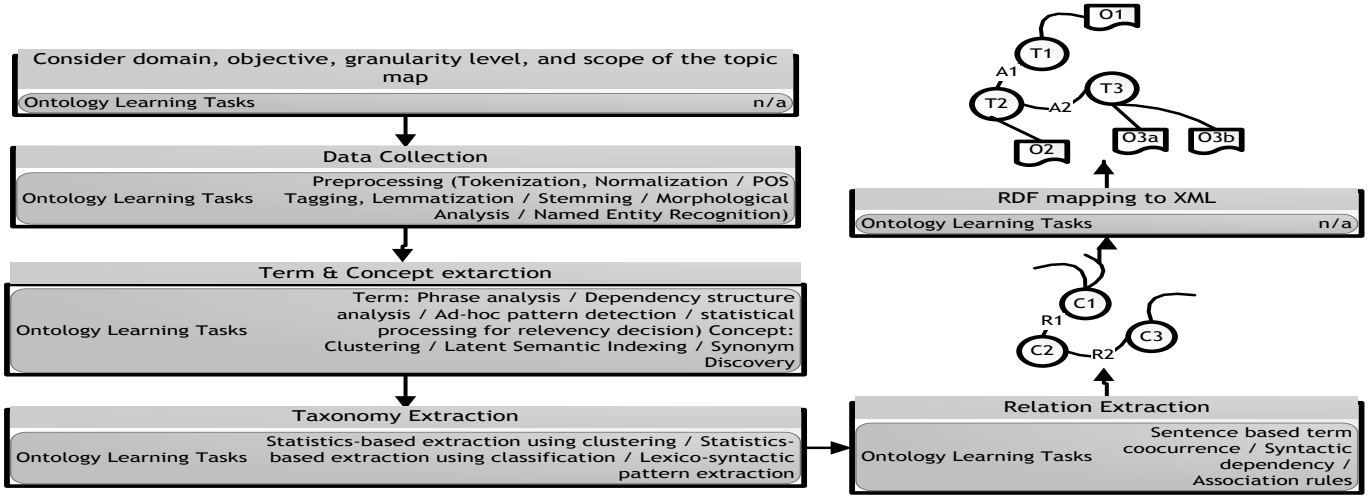

Figure 4 An Ontology Learning Framework of Topic Map Creation

4.1.1. Term extraction

Terms, in many times, are connected to the notion of concepts as we briefly mentioned in the beginning of this section. Consequently, the result of term extraction is closely related to the identification of concepts. In this process, natural language documents are used as input data and a set of terms that are relevant to the

domain corpus are produced as output. Most methods of term extraction are based on information retrieval methods for term indexing and terminology and NLP research. Phrase analysis is used for indentifying complex noun terms while dependency structure analysis for identifying internal structure. State-of –the-art is mostly run a POS tagger over a domain corpus and identifies possible terms by constructing ad-hoc patterns.

4.1.1.1. Term extraction with relevance measure

One approach is based on the assumption that frequent terms in a specific domain corpus are an indication of domain related terms. Term weighting metric is borrowed from information retrieval approach, i.e., TFIDF. The *tfidf* weight coefficient of a term *t* in a document *d* can be calculated as;

$$tfidf_{t,d} = tf_{t,d} \cdot \log\left(\frac{|D|}{df_t}\right)$$

when *D* denotes the total number of documents in a corpus. TFIDF weighs terms in a way that too rare or too frequent terms are ranked lower than the terms that are proportionate across documents.

In some cases, researchers design weighting measures for their own purpose. For example, In Cimiano and Volker [26], Relative Term Frequency (RTF), TFIDF, Entropy and the C-value/NC-value method are introduced as ways to assess relevance of terms to the domain corpus in their Text2Onto ontology learning system.

Xu and colleagues [27], in order to discover domain relevant terms, apply KFIDF measure, similar to TFIDF used in typical information retrieval. In their approach, a term is regarded as relevant if it occurs more frequently than other words in a certain category, but occasionally elsewhere. Extracted terms can be manually constructed or linguistically aligned for next steps.

The strength of term extraction using relevance measure is that it can be used without domain specific knowledge. In other words, given a set of domain documents, terms are extracted based on structural information rather than contextual information which may require expert intervention. However, such relevance measures can assume drawbacks in several respects; 1) term position in the document is disregarded, 2) context of the term is not considered, and 3) subtle nuance may be ignored. In addition, term co-reference should be resolved in order to measure the frequency accurately.

4.1.1.2 Term extraction with dictionary

Another approach of term extraction is related to the study of terminology, especially, using existing dictionaries. Important factors in this task are related to identifying proper sense of the term for proper synset extraction. The synsets can be extracted from Wordnet or clustering analysis. For the Wordnet related extraction, word sense disambiguation algorithms are frequently discussed while Latent Semantic Indexing algorithms are discussed for clustering related extraction.

Gupta and Oates [28] propose a method to enrich an existing ontology, the OntoSem ontology, consisting of concepts and taxonomic relations between concepts. To enrich existing concepts in the ontology with new words, they find the concept that best approximates the meaning of a new word in the concept hierarchy. This is accomplished by using the notion of word similarity, for which they use Latent Semantic Analysis.

In Nichols, Bond, Tanaka, Fujita, and Flickinger [29], the authors outline the development of a system that provides automatic acquisition of ontologies by extracting knowledge from dictionary definition sentences. They use the combination of deep and shallow parsing resource to extract ontological relations in greater quantity and quality. Using this method, they construct ontologies from two different Japanese lexicons and one English lexicon by linking them to existing, handcrafted ontologies, by aligning them at the word-sense level. This alignment provides a representative evaluation of the quality of the relations being extracted.

Agirre and colleagues [30] approach this problem to overcome shortcomings of Wordnet, i.e., the lack of topical links among concepts, and the proliferation of senses. Their methodology aims to enrich the concepts in Wordnet using text retrieved from the World Wide Web. Firstly, for each concept sense, web documents are retrieved using queries formulated with synonyms. Then the documents retrieved are organized in collections, one per word sense. Finally, for each collection, the words and their frequencies are extracted and compared with the data in other collections that cover other senses of the same concept. The words that have a distinctive frequency for one of the collections are grouped in a list, which then constitutes the topic signature for each concept sense. Given a word, the concepts that lexicalize its word sense are hierarchically clustered.

Alfonseca and Manandhar [31] describe a top-down classification algorithm to extend existing ontologies such as WordNet with new concepts. Using the topic signatures, each concept would be represented by the set of words that co-occur with it, and the frequencies with which they appear. Several similarity metrics, such as TFIDF or chi-square, are used to measure the distance between the different concepts.

The strength of term extraction using existing dictionaries is that the terms can be expanded and validated based on existing structure. Especially, when a domain dictionary is available, term extraction results can be highly accurate. However, domain dictionaries may not be available in all cases. In addition, using general thesaurus like Wordnet may result in extracting terms that do not exactly reflect domain semantics.

4.1.1.3. Other term extraction approaches

Other approaches facilitate methods requiring more expert intervention. Lame [32] proposes a methodology to build a legal ontology from the legal texts published every day in the Journal Officiel de la Republique Francaise, the French official publication for legal texts. After the terms are extracted using Lexter and Sylex, they are tested for relevance against predefined rules. Rules include exclusion of less useful term expressions, e.g., standard time patterns and cross reference between documents, and inclusion of useful term expressions, e.g., terms that match predefined legal terms. Then, extracted terms are explored for lexical relations in the first level knowledge acquisition and investigated in terms of hierarchical, structural, and functional models in the second level knowledge acquisition. This method utilizes predefined schema rather than discovering knowledge structure from the given data. However, it seems reasonable since legal domain knowledge includes significant amount of everyday concepts than other research fields. Another domain that may involve some degree of expert intervention is medicine. LeMoigno and colleagues [33] describe the building of an ontology in surgical intensive care using textual reports (CRH). The corpus is tagged in textual units to indicate unit, number of CRH, and what the paragraph is about. Then, using SYNTEX, they yield a dependency network of words and syntagms. Each syntagm in the network is connected to its head (H) and to its expansion(s) (E), the link being labeled by the name of the dependency relation (R).

By using predefined schema for term extraction, one may get a list of terms that are highly relevant to the purpose of the term extraction. Highly structured research area like medicine or areas with wide-ranging domain subjects like law may be appropriate for this type of method. However, it requires a significant amount of human intervention as human experts have to go through extracted terms and examine if they are appropriate for the given schema.

4.1.2. Concept extraction

Most research in concept extraction addresses the question from a linguistic or textual perspective, regarding concepts as clusters of related terms. Therefore, this approach overlaps significantly with that of term extraction. However, another approach defines a concept as a pair with lexicon consisting of the intension and extension of the concept, in some form of linguistic realization. Output of this process may contain lists of terms connected to domain concepts or term clusters which can indicate formulation of concepts in the domain.

One approach utilizes clustering techniques. Kahn and Luo [34] construct a hierarchy using clustering techniques. Similar documents in terms of content are associated with the same concept in the ontology. Then, a concept for each cluster of documents relative to the same topic in the hierarchy is assigned using a bottom up concept assignment mechanism. To achieve this goal, a topic tracking algorithm and Wordnet are used. Topic tracking is used for the assignment of a concept in each leaf node. For this, a topic based on a Rocchio algorithm will first be assigned for each document. Then, distinct topics are determined that appear in each leaf node. If only a single topic appears in a leaf node, this topic is simply assigned as a concept in the leaf node. However, if more than one topic appears in a leaf node, a majority of documents of a leaf node will be associated with a specific topic and this topic will be assigned as a concept to this leaf node. In cases in which the majority rule cannot be applied, generic concept from WordNet using all the topics is chosen and assigned. If association between keyword and concept is 1:N, concept disambiguation is required. For disambiguation of concepts, vector for each sense is constructed and cosine similarity is applied.

Another approach, which is related to intention - extension concept pair, is commonly referred to as ontology population task. Guiliano and Gliozzo [35] present an approach that a system decides whether *w* entails *e* or not, given two generic words *w* and *e*, and a context H. Their method broadly consists of three steps; first, named entities in coarse-grained categories are recognized; second, from the domain corpus, contextual lexicalizations of coarse-grained type are extracted and trained while occurrences of the named entities are classified; third, a distinct category is assigned to the entities referring to the same phrase in the list. The basic idea is that an entity belongs to a specific category if its occurrences entail a particular superclass "more often than expected by chance". The expectation is modeled on the basis of the overall distribution of fine-grained category labels, assigned during the second step, in the corpus.

Concept extraction is a bridging process between term extraction and taxonomy extraction in a sense that this process is melted in either the term extraction process or the taxonomy extraction process in many cases.

4.1.3. Taxonomy extraction

Most methods are based on application of lexico-syntactic patterns, distributional similarity and hierarchical clustering. The output of this task is a hierarchical structure of terms and/or term clusters which may indicate concepts in the domain.

4.1.3.1. Taxonomy extraction using lexico-syntactic patterns

An approach of concept hierarchy formulation is based on application of lexico-syntactic patterns. In this approach, regular expressions of class – superclass relationship are modeled and recurring expressions in documents are matched against the modeled patterns. Some of the common patterns of such relationship are 'such NP as NP_n (and/or) NP', 'NP_n (and/or) other NP', 'NP including NP (and/or) NP', and 'NP especially NP (and/or) NP' [36].

Cimiano and colleagues [37] present a approach to learning taxonomic relations between terms by considering multiple and heterogeneous sources of evidence. In order to normalize the source combination, they exploit a machine-learning approach, representing all the sources of evidence as first-order features and training standard classifiers. In their method, they consider, in particular, different features derived from Wordnet, an approach matching Hearst-style patterns in a corpus and on the Web. In particular, they explore different classifiers as well as various strategies for dealing with unbalanced datasets. They evaluate the approach by comparing the results with reference taxonomy for the tourism domain.

Maedche, Pekar, and Staab [38] adopt the idea of using lexico-syntactic patterns in the form of regular expressions for the extraction of taxonomic relations. In this lexico-syntactic ontology learning approach, the text is scanned for instances of distinguished lexico-syntactic patterns that indicate the taxonomic relation. Thus, the underlying idea is to define a regular expression that captures re-occurring expressions and map the results of the matching expression to a semantic structure, such as taxonomic relations between concepts.

Sundbald [39] explore how lexical and ontological relations can be acquired automatically from natural language questions. The focus in this paper is on identifying hyponym and meronym relations by using simple pattern matching. It is shown that natural language questions can provide a significant source for ontological information.

Hahn and Marko [40] introduce a methodology for automating the maintenance and growth of domain-specific concept taxonomies and grammatical class hierarchies simultaneously, based on knowledge capture from natural language texts. Their approach proposes integrated approach for learning lexical-syntactic and conceptual knowledge. New concepts are acquired and positioned in the concept taxonomy. Also, the grammatical status of their lexical correlates is learned taking two knowledge sources into account., i.e., domain knowledge and grammatical knowledge. Domain knowledge provides a concept and role taxonomy which serves as a comparison scale for judging the plausibility of newly derived concept descriptions in the light of that prior knowledge. Grammatical knowledge contains a type hierarchy of lexical classes which make increasingly restrictive grammatical constraints available for linking an unknown word with its corresponding word class.

Although using the lexico-syntactic pattern for the hierarchical relationship is commonly used techniques the proposed patterns are not always available in natural language documents.

4.1.3.2. Taxonomy extraction using distributional similarity and hierarchical clustering

The basic idea of this approach is that terms that bear semantic similarity tend to occur in a similar context in documents [41]. The terms are represented as a vector consisting of contextual features and vectors are computed for similarity.

Ryu and Choi [42] propose a method to assign taxonomic relations among technical terms. They describe an approach to the problem that relies on term specificity and similarity measures. Term specificity and similarity are necessary conditions for taxonomy learning, because highly specific terms tend to locate in deep levels and semantically similar terms are close to each other in taxonomy. In their method, new taxonomy starts with empty state, and changes to enrich taxonomic structure with the repeated insertion of terms. Terms to be inserted are sorted by term specificity values.

Cimiano [43] describes a conceptual clustering method based on *Formal Concept Analysis* for automatic taxonomy construction. He claims that the method he describes outperforms agglomerative hierarchical clustering as well as with Bi-Section-KMeans. Instead of using a totally unsupervised approach, he has also examined the possibility of applying a *weakly supervised agglomerative clustering algorithm* which exploits a hypernym oracle to guide the clustering process.

Since hierarchical clustering algorithm tends to build binary nodes under a concept, it seems unreasonable when there are more than two natural categories under one concept. Moreover, distance relationship among the categories as a result of hierarchical clustering may not correctly represent conceptual distance.

4.1.4. Relation extraction

Most methods are based on syntactic dependency, i.e., problem of acquiring selection restrictions for verb arguments in natural language processing, and association rules based on sentence based term co-occurrence.

Suchanek and colleagues [44] present LEILA, a system that can extract instances of arbitrary given binary relations from natural language Web documents without human interaction. The core algorithm proceeds in three phases; In the Discovery Phase, linkages in which an example pair appears are identified. It replaces the two words by placeholders, thus producing a pattern. These patterns are collected as positive patterns. Then, the algorithm runs through the sentences and also identifies negative patterns. In the Training Phase, statistical learning is applied to learn the concept of positive patterns. Finally, in the Testing Phase, the

algorithm considers again all sentences in the corpus. For each linkage, it generates all possible patterns by replacing two words by placeholders. If the two words form a candidate and the pattern is classified as positive, the produced pair is proposed as a new element of the target relation.

Specia and Motta [45] present an approach for extracting relations from texts that exploits linguistic and empirical strategies. They follow a pipeline method involving a parser, part-of- speech tagger, named entity recognition system, pattern-based classification and word sense disambiguation models, and resources such as ontology, knowledge base and lexical databases. The authors suggest that the use of knowledge intensive strategies to process the input text and corpus based techniques to deal with unpredicted cases and ambiguity problems allows to accurately discover the relevant relations between pairs of entities in that text.

Kavalec and colleagues [46] propose a technique for extraction of lexical entries that may give cue in assigning semantic labels to otherwise 'anonymous' relations. They implement the technique as extension to the existing Text-to-Onto tool, and test on a collection of texts describing worldwide geographic locations from a tour–planning viewpoint.

Chklovski and Pantel [47] present an automatic method for extracting fine-grained semantic relations, addressing relations between verbs. They detect similarity, strength, antonymy, enablement, and temporal relations between pairs of verbs with high mutual information using lexico-syntactic patterns over the Web. On a set of 26,118 strongly associated verb pairs, the extraction algorithm yields 56.5% accuracy, while, on the relations strength and similarity, it achieves 79.6% and 66.7% accuracy respectively.

Xu [27], inspired by Hearst [48], propose a system learning lexico-syntactic patterns indicating paradigmatic relations. Instead of using initial seeds of patterns, they employ the existing semantics relations provided GermaNet.

Gamallo and colleagues [49] present a corpus-based method for extracting semantic relations between words. The method is based on two sequential procedures. First, it automatically classifies syntactic dependencies according to their selection restrictions. Those dependencies that require the same selection restrictions are put together into the same semantic group. Then, interpretation rules are applied on the classified syntactic dependencies, in order to learn the specific semantic relations underlying syntactically related words.

Maedche and collegues [50], based on their generic architecture, describe a case study in telecommunication domain for mining ontologies from text using methods based on dictionaries and natural language text. Supporting the overall text ontology engineering process, their approach combines dictionary parsing mechanisms for acquiring a domain-specific concept taxonomy with a discovery mechanism for the acquisition of non-taxonomic conceptual relations.

Relation extraction may be considered the least advanced task process within the ontology learning framework. It is a very complex to automatically recognize semantic relations since natural language expressions are not straightforward.

4.2. Tools and Systems of Ontology Learning

We discuss a few ontology learning systems that provide functions of term extraction, concept extraction, taxonomy extraction, and relation extraction.

4.2.1. Text2Onto

Text2Onto is described in Cimiano and Volker [26] and Maedche and Staab [51]. The system provides term extraction and taxonomy extraction functions using statistical analysis, conceptual clustering, patterns, and WordNet. Relation extraction is described in terms of association rules and subcategorization frames.

4.2.2. OntoLearn

OntoLearn is discussed in studies conducted at the University of Rome [52, 53]. Term extraction and interpretation is conducted using shallow parsing, term disambiguation, and compositional interpretation. Relations are identified through classification of the relation between terms in a compound into predefined set of relations.

4.2.3. OntoLT

OntoLT is a plug-in for Protégé ontology engineering environment. The system is described in Buitelaar, Olejnik, and Sintek [54]. Terms are extracted using shallow parsing and statistical analysis. Taxonomy and semantic relation are extracted through shallow parsing and manually defined mapping rules.

4.3. Ontology vs Topic Map

Studies have shown that topic map models and ontology data models, e.g., RDF, may interoperate at a fundamental level [7, 55-57]. Since both standards are concerned with defining relationships between entities with some type of identity, basic idea of knowledge representation is similar. Main endeavor focuses on how each component of both data models can be mapped to one another with a minimum semantic distortion.

5. DISCUSSION & CONCLUSION

In this paper, we reviewed cybercrime and problems of conducting cybercrime research. As a potential solution, we reviewed topic map data model and how we can automate some portion of topic map creation using ontology learning approaches. In this section, we address some questions identified in reviewing automatic topic map generation using ontology learning approaches. We also discuss future work and research directions based on the reviewed literature items.

We address two problems in regards to using ontology learning approaches; lack of theoretical background in mapping 'concept' to 'topic' and technical immaturity in relation extraction methods. In Section 4, we assumed that we use 'concept' in the ontology architecture as 'topic' in the topic map structure based on current practices. However, concepts can be loosely identical to research topics in the academic research context. In other words, not all domain concepts are research topics, and not all research topics are domain concepts. Mapping a relation to an association seems more straightforward. However, current maturity of relation extraction techniques is, in many cases, in its infancy and requires human validation for more accurate extraction. One important presumption that ontology learning approaches make in the first place is that ontology learning is not a perfect solution but a supportive tool to decrease human effort in managing abundance of information. Therefore, ontology learning techniques for topic map creation should be treated as such; expert involvement should be emphasized in the refining stage of a topic map design.

Future research in relation to this review is to enrich the literature review and design a pilot experiment. Some of the future work are; 1) defining different classes of users and exploring various use cases for a cybercrime research topic map, 2) investigating topic map visualization and use of the visual analytics into a topic map design, 3) collecting data based on the decision of what to collect and how to collect them, 4) conducting a pilot test with different configurations of ontology learning tools and techniques, and 5) exploring evaluation and analysis methods.

Research directions in regards to further project deem to pursue answers that should be addressed in seeking solutions to questions below;

RQ1a: How a research topic is represented in cybercrime publications?

RQ1b: Is there a discernable pattern in the relationship between research topics and topic representation in cybercrime publications?

RQ2: How can we embed the pattern to automated cybercrime research topic map creation approach?

In order to follow the appropriate tracks to answers for the research questions, innate characteristics of cybercrime research should be defined in relation to relevant research fields. Then, a set of rules and algorithms should be designed to recognize appropriate topics and associations among the topics of cybercrime research. Finally, proper analytic approach should be identified.

More and more cybercrime related conferences aim for academic research and, as a consequence, papers are published through those conferences and journals devoted to cybercrime related issues. However, still, a significant number of conventions are devoted to industry and government, e.g., law enforcement and security. Unfortunately, products of such discussions seem to be kept among or within the interest groups except for the limited number of academic windows to such discussions. Therefore, it is sometimes difficult to acquire research data that are products of proprietary conventions. Creating a cybercrime nomenclature will facilitate researcher communication, addressing proper research problems, and promote sharing knowledge among groups of different stakeholders.

REFERENCES

1. ISO/IEC_JTC1, *ISO/IEC 13250 Topic Maps: Information Technology Document Description and Processing Languages*. 2002.

2. Rath, H.H. and S. Pepper. *Topic Maps: Introduction and Allegro*. in *XML 1999*. 1999. Philadelphia, PA.

3. Biezunski, M. and S.R. Newcomb, *XML Topic Maps: Finding Aids for the Web*, in *IEEE Multimedia*. 2001. p. 104-108.

4. Dicheva, D. and C. Dichev, *TM4L: Creating and browsing educational topic maps.* British Journal of Educational Technology, 2006. **37**(3): p. 391-404.

5. Stümpflen, V., R. Gregory, and K. Nenova. *From Biological Data to Biological Knowledge*. in *TMRA '06 International Conferences onTopic Maps Research and Applications*. 2006. Leipzig, Germany.

6. Smolnik, S. and L. Nastansky, *K-Discovery: Using Topic Maps to Identify Distributed Knowledge Structures in Groupware-based Organizational Memories*, in *35th Hawaii International Conference on System Sciences*. 2002.

7. Ciancarini, P., et al. *Metadata on the Web: On the integration of RDF and Topic Maps*. in *Extreme Markup Languages 2003*. 2003. Montréal, Québec.

8. Schweiger, R. and J. Dudeck. *Improving Information Retrieval Using XML and Topic Maps*. in *TMRA '05 International Workshop on Topic Maps Research and Applications* 2004. Leipzig, Germany

9. Garshol, L.M., *Metadata? Thesauri? Taxonomies? Topic maps! Making sense of it all.* Journal of Information Science, 2004. **30**(4): p. 378-391.

10. Dicheva, D., et al., *Authoring Topic Maps-based Digital Course Libraries*, in *International Workshop on Ontologies and Semantic Web for E-Learning* 2004: Eindhoven, The Netherlands.

11. Dicheva, D. and C. Dichev, *Educational Topic Maps*, in *3rd International Semantic Web Conference* 2004: Hiroshima, Japan.

12. Barta, R. *Towards a Formal TMQL Semantics*. in *TMRA '06 International Conferences onTopic Maps Research and Applications*. 2006. Leipzig, Germany.

13. Grand, B.L. and M. Soto, *Topic Maps Visualization*, in *XML Topic Maps*, J. Park, Editor. 2003, Addison-Wesley: New York, NY.

14. Cregan, A., *An OWL DL construction for the ISO Topic Map Data Model*. 2005.

15. Garshol, L.M. *Towards a Methodology for Developing Topic Maps Ontologies*. in *TMRA '06 International Conferences onTopic Maps Research and Applications*. 2006. Leipzig, Germany.

16. Gordon, S. and R. Ford, *On the definition and classification of cybercrime.* Journal in Computer Virology, 2006. **2**(1): p. 13-20.

17. Lu, C., W. Jen, and W. Chang, *Trends in Computer Crime and Cybercrime Research During the Period 1974-2006: A Bibliometric Approach*, in *Intelligence and Security Informatics*. 2007. p. 244-250.

18. Liu, W.Y., et al., *An antiphishing strategy based on visual similarity assessment.* IEEE Internet Computing, 2006. **10**(2): p. 58-65.

19. Rath, H.H., *Technical Issues on Topic Maps*, in *Metastructures Conference*. 1999: Montreal, Canada.

20. Roberson, S. and D. Dicheva. *Semi-automatic ontology extraction to create draft topic maps*. in *45th Annual Association for Computing Machinery Southeast Conference* 2007. Winston-Salem, NC.

21. Hofmann, T. *Probabilistic Topic Maps: Navigating through Large Text Collections*. in *Third Symposium on Intelligent Data Analysis* 1999. Amsterdam, The Netherlands

22. Böhm, K., et al., *Topic map generation using text mining.* Journal of Universal Computer Science, 2002. **8**(6): p. 623-633.

23. Yang, H.-C. and C.-H. Lee. *Building Topic Maps Using a Text Mining Approach*. in *International Syposium on Methodologies for Intelligent Systems (ISMIS)*. 2003. Maebashi City, Japan.

24. Abramowicz, W., T. Kaczmarek, and M. Kowalkiewicz, *Supporting topic map creation using data mining techniques.* Australasian Journal of Information Systems, 2003/2004. **Special Issue**: p. 63-78.

25. Barthes, R., *Element of semiology*. 1968, New York: Hill and Wang.

26. Cimiano, P. and J. Völker. *Text2Onto - A Framework for Ontology Learning and Data-driven Change Discovery* in *10th International Conference on Applications of Natural Language to Information Systems (NLDB)*. 2005. Alicante, Spain.

27. Xu, F., et al. *A Domain Adaptive Approach to Automatic Acquisition of Domain Relevant Terms and their Relations with Bootstrapping*. in *LREC2002, the Third Conference on Language Resources and Evaluation*. 2002. Las Palmas, Spain.

28. Gupta, A. and T. Oates. *Using ontologies and the web to learn lexical semantics*. in *International Joint Conferences on Artificial Intelligence (IJCAI-07)*. 2007. Haderabad, India.

29. Nichols, E., et al. *Multilingual ontology acquisition from multiple MRDs*. in *2nd Workshop on Ontology Learning and Population: Bridging the Gap between Text and Knowledge in conjunction with COLING/ACL 2006*. 2006. Sydney, Australia: Association for Computational Linguistics.

30. Agirre, E., et al. *Enriching very large ontologies using the WWW*. in *First Workshop on Ontology Learning OL'2000 in conjunction with the 14th European Conference on Artificial Intelligence ECAI'2000*. 2000. Berlin, Germany.

31. Alfonseca, E. and S. Manandhar. *An Unsupervised Method for General Named Entity Recognition and Automated Concept Discovery*. in *1st International Conference on General Wordnet*. 2002. Mysore, India.

32. Lame, G. *Knowledge acquisition from texts towards an ontology of French law*. in *EKAW'2000, 12th International Conference on Knowledge Engineering and Knowledge Management* 2000. Juan-les-Pins, France.

33. Moigno, S.L., et al. *Terminology extraction from text to build an ontology in surgical intensive care*. in *Workshop on Machine Learning and Natural Language Processing for Ontology Engineering*. 2002. Lyon, France.

34. Khan, L. and F. Luo. *Ontology Construction for Information Selection*. in *14th IEEE International Conference on Tools with Artificial Intelligence (ICTAI 2002)*. 2002. Washington, DC.

35. Giuliano, C. and A. Gliozzo. *Instance Based Lexical Entailment for Ontology Population*. in *Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*. 2007. Prague.

36. Hearst, M.A. *Automatic Acquisition of Hyponyms from Large Text Corpora*. in *COLING-92*. 1992. Nantes.

37. Cimiano, P., et al., *Learning Taxonomic Relations from Heterogeneous Sources of Evidence*, in *Frontiers in Artificial Intelligence and Applications* P. Buitelaar, P. Cimiano, and B. Magnini, Editors. 2005, IOS Press. p. 59-73.

38. Maedche, A., V. Pekar, and S. Staab, *Ontology learning part one - On discovering taxonomic relations from the web*, in *Web Intelligence*, N. Zhong, J. Liu, and Y.Y. Yao, Editors. 2003, Springer-Verlag. p. 301-321.

39. Sundblad, H. *Automatic Acquisition of Hyponyms and Meronyms from Question Corpora*. in *Workshop on Machine Learning and Natural Language Processing for Ontology Engineering*. 2002. Lyon, France.

40. Hahn, U. and K.G. Markó. *Joint knowledge capture for grammars and ontologies*. in *1st international Conference on Knowledge Capture* 2001. Victoria, British Columbia, Canada.

41. Harris, Z., *Distributional structure.* Word, 1954. **10**(23): p. 146-162.

42. Ryu, P.-M. and K.-S. Choi. *Taxonomy learning using term specificity and similarity*. in *2nd Workshop on Ontology Learning and Population: Bridging the Gap between Text and Knowledge in conjunction with COLING/ACL 2006*. 2006. Sydney, Australia: Association for Computational Linguistics.

43. Cimiano, P., *Ontology learning and population from text: Algorithms, evaluation and applications*. 2006, New York: Springer.

44. Suchanek, F.M., G. Ifrim, and G. Weikum. *LEILA: Learning to Extract Information by Linguistic Analysis*. in *2nd Workshop on Ontology Learning and Population: Bridging the Gap between Text and Knowledge in conjunction with COLING/ACL 2006*. 2006. Sydney, Australia: Association for Computational Linguistics.

45. Specia, L. and E. Motta. *A hybrid approach for extracting semantic relations from texts* in *2nd Workshop on Ontology Learning and Population: Bridging the Gap between Text and Knowledge in conjunction with COLING/ACL 2006*. 2006. Sydney, Australia: Association for Computational Linguistics.

46. Kavalec, M., A. Maedche, and V. Svatek. *Discovery of Lexical Entries for Non-taxonomic Relations in Ontology Learning*. in *SOFSEM 2004: Theory and Practice of Computer Science* 2004. Czech Republic

47. Chklovski, T. and P. Pantel. *Large-Scale Extraction of Fine-Grained Semantic Relations between Verbs* in *Mining for and from the Semantic Web Workshop*. 2004. Seattle, WA.

48. Hearst, M.A., *Automated Discovery of WordNet Relations*, in *WordNet: An Electronic Lexical Database and Some of its Applications*, C. Fellbaum, Editor. 1998, MIT Press.

49. Gamallo, P., et al. *Mapping Syntactic Dependencies onto Semantic Relations*. in *Workshop on Machine Learning and Natural Language Processing for Ontology Engineering*. 2002. Lyon, France.

50. Maedche, A. and S. Staab. *Mining Ontologies from Text*. in *EKAW-2000 12th International Conference on Knowledge Engineering and Knowledge Management*. 2000. Juan-les-Pins, France.

51. Maedche, A. and S. Staab. *The text-to-onto ontology environment*. in *International Conference on Complex Systems*. 2000.

52. Navigli, R., et al. *Automatic ontology Learning: Supporting a per-concept evaluation by domain experts*. in *Workshop on Ontology Learning and Population in conjunction with 16th European Conference on Artificial Intelligence*. 2004. Valencia, Spain.

53. Velardi, P., et al., *Evaluation of OntoLearn, a methodology for automatic learning of domain ontologies*. 2003.

54. Buitelaar, P., D. Olejnik, and M. Sintek, *A Protégé Plug-In for Ontology Extraction from Text Based on Linguistic Analysis*, in *The Semantic Web: Research and Applications*. 2004. p. 31-44.

55. Cregan, A. *Building Topic Maps in OWL-DL*. in *Extreme Markup Languages*. 2005. Montréal, Québec.

56. Moore, G. *RDF and Topic Maps: An exercise in convergence*. in *XML Europe*. 2001. Berlin, Germany.

57. Vatant, B. *Ontology-driven topic maps*. in *XML Europe*. 2004. Amsterdam, The netherlands.