Concept Mining using Conceptual Ontological Graph (COG)

Shady Hassan, Fakhreddine Karray, and Mohamed Kamel
Department of Electrical and Computer Engineering
University of Waterloo, Waterloo, Ontario, Canada
Email: {shady,karray,kamel}@pami.uwaterloo.ca

Abstract. Concept mining (CM) is the area of exploring and finding links, associations, relationships, and patterns among huge collections of information. In this paper, we propose concept-based text representation, with an emphasis on using the proposed representation in different application s such as information retrieval, text summarization, and question answering. This work presents a new paradigm for concept mining by extracting the concept-based information from a raw text. At the text representation level, we introduce a sentence based conceptual ontological representation that builds concept-based representations for the whole document. A new concept-based similarity measure is proposed to measure the similarity of texts based on their meaning. The proposed approach is domain independent and it could be applied to general domain applications. The proposed approach has been applied to the domain of information retrieval and preliminary results are promising, and give an affirmation for proceeding in the right directions of this research.

1 Introduction

Text mining is statistical analysis of word frequencies within a document. Text mining generally ignores word order. It is important to note that understanding the meaning of words couldn't be deduced from statistical analysis of word frequencies. Concept mining is related to understand the meaning of text. Inferring what a piece of text is about is a crucial step for machine understanding. Any word we use might have multiple meanings, and we use the context to disambiguate what is meant. Therefore, there is a need for a representation that captures the semantics in text in a formal structure. This need arises from the necessity to perform a variety of tasks that involves the meaning of the linguistic input.

The underlying structure of the predicate argument layout allows the creation of a composite meaning representation from the meaning of the individual parts of a linguistic input. More generally, the predicate argument structure permits the link between the arguments in surface structures of the input text and their associated semantic roles. The study of roles associated with verbs is usually referred to thematic role or case role analysis [1]. Thematic roles, first proposed by Gruber and Fillmore [2], are set of categories that provide a shallow semantic language to characterize the verb arguments.

Recently, there are many attempts to label thematic roles to a sentence automatically. Gildea and Jurafsky [3] were the first to apply a statistical learning technique to the FrameNet database. They present a discriminative model for determining the most probable role for a constituent given the frame, predicator, and other features. These probabilities, trained on the FrameNet database, are based on the verb, the head words of the constituents, the voice of the verb (active, passive), and the syntactic category (S, NP, VP, PP, etc.) and grammatical function (subject, object) of the constituent to be labeled. They evaluated their model on a pre-release version of the FrameNet I corpus, which at that time contained about 50,000 sentences and 67 frame types. Their model was trained by first using the parser of Collins [4], and deriving features from that parse, the original sentence, and the correct FrameNet annotation of that sentence. A machine learning algorithm for shallow semantic parsing has been proposed in [5]. It is an extension of the work in [3] and [6]. Their algorithm is based on Support Vector Machines which gives an improvement in performance over earlier classifiers by [3] . The shallow semantic parsing is formulated as a multi-class classification problem. Support Vector Machines (SVMs) is used to identify the arguments of a given predicate in a sentence and classify them by the semantic roles that they play (AGENT, THEME, GOAL, etc.).

There are numerous efforts for semantic labeling nowadays. However, and to our knowledge, there is no single body that takes the initiative to employ the full potential of the output of the role labeling task. It is important to note that extracting relations between verbs and between their arguments in the same sentence has a promising potential for understanding the meaning of a sentence. Understanding the meaning of text relies on understanding concepts that mentioned in this text.

The main objective of this research work is to introduce a concept mining method that aims to understand the meaning of text based on concept-based understanding for semantic roles. The proposed approach extracts concepts which capture semantics in text. This work aims to extract and use concept-based information, which represents semantics in text, in several applications to demonstrate the improvement of application output after using the concept mining method.

The rest of the paper is organized as follows. Section 2 outlines the proposed system and its main components, in particular the text representation model and the similarity measure. Section 3 presents preliminary results of the test case used to demonstrate the effect of the proposed representation in an information retrieval application. Finally, a conclusion of the work and a future work discussion of the research are given in section 4.

2 The Proposed Approach

The problem of understanding the meaning of text can be divided into three phases: extracting the concepts which represent the meaning of a sentence; representing the captured concepts in a hierarchical structure to divide text meanings into levels; and identifying relations between concepts to represent the semantic roles of the extracted concepts.

2.1 System Overview

The proposed system has two main objectives: Concept-based text representation and matching scheme. The proposed representation consists of concepts and semantic relations which join between concepts. A new matching scheme is needed to match between concepts placed in the representation and to provide a new concept-based similarity measure suitable for the proposed representation.

2.2 Concept-based Text Representation

Any complete and meaningful sentence in English language must consist of main concepts that give an overview about the general meaning of a sentence. A sentence also might have supplementary concepts, beside its main concepts, that provide different levels of details regarding the meaning of a sentence.

Extracting relations between verbs and between their arguments in the same sentence has a promising potential for understanding the meaning of a sentence. Verbs and their arguments represent the concepts that mentioned in text. These concepts represent the semantics in text and semantics represents the meaning of text. Thus, we introduce a new concept-based representation that utilizes the output of the role labeling task and represents verbs with their arguments as concepts with their relations. The proposed text representation captures concepts from a sentence and represents them in a hierarchical manner based on the structure of a sentence. The hierarchal representation presents different levels of details based on the meaning of a sentence. The representation includes semantic relations among concepts based on the role of each concept. The proposed hierarchical representation aims to provide a clear separation between main concepts and other supplementary concepts in a sentence.

We present a new representation called Conceptual Ontological Graph (COG) model. The representation is a conceptual graph where entities, which are the constituents of a sentence, are represented as vertices V, and relations among constituents such as agents, objects, actions, are represented as arcs A. Each node holds information about the entity it represents including its original text, syntactic information, head word, synonyms, concepts, and relations with other nodes. Each sentence in a text is represented by the COG representation as a one to one relationship. This will end up having a conceptual document represented by the COG representation for each text document in a corpus. COG comprises nested conceptual graphs in an

ontological manner. There are two reasons for the hierarchical representation. First, it distinguishes between main concepts and other supplementary concepts listed in a sentence. Second, it presents different levels of depth for understanding the meaning of a sentence.

Due to the proposed representation COG is a conceptual graph, all the operations that are applied to conceptual graphs can also be applied to the COG representation.

2.3 Operations on the COG Representation

The following are some basic operations that can be applied to the COG representation.

- Equivalence rules. The copy rule copies a graph or subgraph. The simplify rule performs the inverse operation of erasing a copy. Let v be any subgraph of a simple graph u; v may be empty or it may be all of u.
- Specialization rules. The restrict rule specializes the type or referent of a single concept node. The join rule merges two concept nodes to a single node. These rules transform u to a graph w that is more specialized than u.
- Generalization rules. The unrestrict rule, which is the inverse of restrict, generalizes the type or referent of a concept node. The detach rule, which is the inverse of join, splits a graph in two parts at some concept node. The last two rules transform u to a graph w that is a generalization of u.

2.4 Rules of Inference

- Equivalence rules. The equivalence rules may change the appearance of a graph, but they do not change its logical status. If a COG u is converted to a COG v by these rules, then u implies v, and v implies u.
- Specialization rules:. The specialization rules transform a COG u to another COG v that is logically more specialized: v implies u.
- Generalization rules. The generalization rules transform a COG u to another COG v that is logically more generalized: u implies v.

By handling the syntactic details of conceptual graphs, the generalization and specialization rules enable the rules of inference to be stated in a general form that is independent of the graph notation.

2.5 System Architecture

A raw text document is the input to our proposed system. Each sentence in a document is labelled automatically based on the PropBank notations. The labelled sentence is the output of the role labelling task. Each original English sentence might have one or more labeled sentences. The number of generated labelled sentences is totally dependent on the complexity of a sentence. In other words, the more the sentence has different labelled output sentences the more the sentence has many verbs associated with their arguments. The objective of the proposed system is to capture the relations among these labelled sentences to represent all the generated labelled sentences in one representation. This representation can provide the meaning of the original sentence based on the concepts that are placed in the representation. The proposed system comprises six main components

- Text Parser (TP),
- Sentence Analyzer (SA),
- Graph Constructor (GC),
- Conceptual Ontological Graph Constructor (COGC),

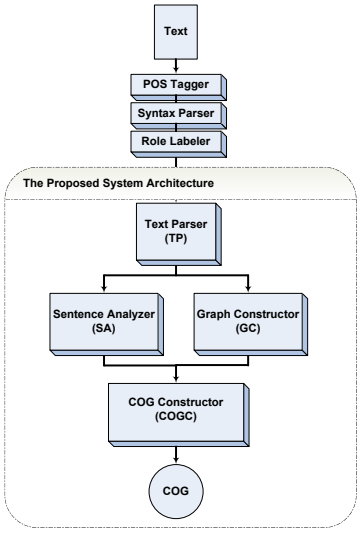

Figure 1: System Architecture

- Synonyms Finder (SF), and
- Concept Matcher (CM).

The proposed system parses labeled sentences using Text Parser (TP) component, which converts each labeled sentence to a sentence called ”parsed sentence”. By using the Sentence Analyzer (SA) component, the proposed system analyzes and categorizes parsed sentences to one or two categories from the following proposed five categories: one-sentence (One-Sent), main sentence (Main-Sent), referenced sentence (Ref-Sent), container sentence (Con-Sent), and unreferenced sentence (Unref-Sent). The Graph Constructor (GC) constructs conceptual graph for each parsed sentence as a basic unit of the COG representation. The COG Constructor (COGC) component combines all the generated conceptual graphs into one Conceptual Ontological Graph (COG) based on the categories of the parsed sentences. For each node in the COG representation, the Synonyms Finder (SF) extracts the suitable senses from WordNet [19] for the significant concept (head word) in a constituent. Finally, the Concept Matcher (CM) compares among concepts represented in the COG representation. Figure 1 depicts the proposed system architecture. The following is the description of each component in the proposed system.

2.5.1 Text Parser (TP)

The output of the role labelling task is the input to the TP component. First, Text Parser (TP) assigns sequential number as a unique identifier to each labelled sentence. Then, TP parses a labelled sentence and stores its elements including the target word and its arguments into database. For each original English sentence in a document, TP component generates a group of parsed sentences for the same original English sentence. The output of the TP component is called *parsed sentences*.

2.5.2 Sentence Analyzer (SA)

This is one of the major components in the proposed system. The output of the TP component is the input to the SA component which comprises two main phases analysis and categorization. The objective of the SA component is to analyze and categorize each parsed sentence to one or two categories from the following five categories: one-sentence (One-Sent), main sentence (Main-Sent), referenced sentence (Ref-Sent), container sentence (Con-Sent), and unreferenced sentence (Unref-Sent). The proposed analyzer algorithm computes the amount of overlapping among words in each constituent including target word and its arguments placed in a parsed sentence. For each parsed sentence, the output of the analysis phase is stored into database.

Based on the output of the analysis phase, the proposed categorizer algorithm during the categorization phase assigns one or two categories for each parsed sentence. The following is the description of each category:

- *One-Sent*: means that there is only one generated parsed sentence, which is called one-sentence.
- *Ref-Sent*: means that there are several parsed sentences. Constituents placed in referenced sentence are referred by constituents placed in main or container sentences.
- *Main-Sent*: means that there are different parsed sentences. Main sentence has the largest number of referent constituents, which refer to other constituents placed in referenced sentences (Ref-Sent).
- *Con-Sent*: means that there are several parsed sentences. Container sentence is not the main sentence and its constituents refer to other constituents placed in other parsed sentences, which are referenced sentence (Ref-Sent).
- *Unref-Sent*: means that there is more than one parsed sentence. Unreferenced sentence is a parsed sentence that its constituents are not referenced by any other constituents placed in any parsed sentence such as Main-Sent or Cont-Sent.

Sentence Analyzer (SA) component keeps track of every constituent whether it refer to or referred by any other constituent placed in other parsed sentence. SA also computes the maximum number of referent constituents for each parsed sentence. SA component comprises two algorithms. The first algorithm analyzes constituents for each parsed sentence. The second algorithm categorizes each parsed sentence to one or two of the following five categories: one-sentence (One-Sent), main sentence (Main-Sent), referenced sentence (Ref-Sent), container sentence (Con-Sent), and unreferenced sentence (Unref-Sent).

The following example shows how a sentence is analyzed and categorized based on the earlier algorithms. There are four parsed sentences for the same sentence "*But leading Socialist dissidents will join Communists and union leaders in Paris to urge people to vote No*".

- *PS1*: But [TARGET leading] [ARG0 Socialist] dissidents will join Communists and union leaders in Paris to urge people to vote No
- *PS2*: [ARGM-DIS But] [ARG0 leading Socialist dissidents] [ARGM-MOD will] [TARGET join] [ARG1 Communists and union leaders] [ARGM-TMP in Paris] [ARGM-PNC to urge people to vote No]
- *PS3*: But leading Socialist dissidents will join Communists and union leaders in Paris to [TARGET urge] [ARG1 people] [ARG2 to vote No]
- *PS4*: But leading Socialist dissidents will join Communists and union leaders in Paris to urge [ARG0 people] to [TARGET vote] No

In this example, parsed sentence number two (PS2) contains maximum number of constituents that refer to other constituents placed in other parsed sentence. Therefore, SA assigns PS2 a main sentence (Main-Sent) category. This means that PS2 contains main concept sentence and it refers to other supplementary concepts placed in other parsed sentences that serve the meaning of the main concepts. Based on the definition of each category for a parsed sentence, PS1 is a referenced sentence because its target word "leading" is placed in constituent "leading Socialist dissidents" placed in PS2. PS1 constituents also don't contain (refer) to any other constituents placed in other parsed sentences. PS3 is categorized to two categories: referenced sentence (Ref-Sent) and container sentence (Con-Sent). PS3 is a referenced sentence (Ref-Sent) because PS3's target word "urge" is referenced by constituent "to urge people to vote No" placed in PS2. PS3 is also a container sentence (Con-Sent) because PS3's constituent "to vote No" refers (contains) to the target word "vote" placed in PS4. Finally, PS4 is a referenced sentence (Ref-Sent) because its target word "vote" is referenced by two constituents which are "to urge people to vote No" placed in PS2 and "to vote No" placed in PS3.

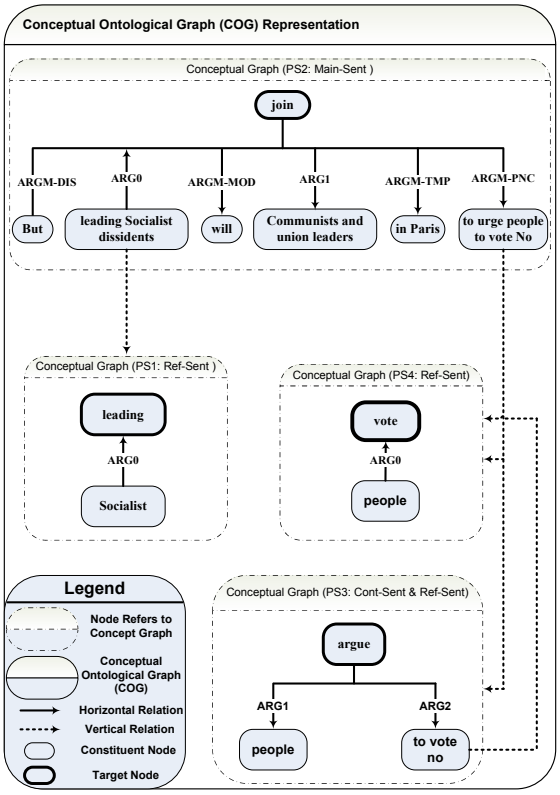

Figure 2: Conceptual Ontological Graph (COG) Representation

2.5.3 Graph Constructor (GC)

The output of the TP component is the input to the GC component. This component constructs a conceptual graph for each parsed sentence. A conceptual graph is the basic unit of the proposed representation COG. A vertex V of the graph G represents any constituent placed in a parsed sentence. A relation R of the graph G represents any role associated with a constituent. A constituent of a parsed sentence could be either a target word or one of the arguments of a target word.

2.5.4 COG Constructor (COGC)

The outputs of the Graph Constructor (GC) and Sentence Analyzer (SA) are the inputs to the COG Constructor (COGC) component. This component constructs one Conceptual Ontological Graph (COG) representation for each original English sentence in a document. The COG representation is constructed from the conceptual graphs generated by the GC and based on the analysis output of the SA component.

There are two kinds of relations in the COG representation: vertical and horizontal relations. Vertical relations represent the semantic roles between the parsed sentences, while horizontal relations represent the relations between each verb and its argument in the same parsed sentence. Figure 2 illustrates the generated COG representation for the same previous example.

The following is the correspondent CGIF for the example in figure 2:

```
[join *a] (ARGM-DIS ?a ?b) [But *b] (ARG0 ?a [leading Socialist dissidents:
[leading *c] [Socialist *d] (ARG0 ?c ?d)]) [will *e] (ARGM-MOD ?a ?e)
[Communists and union leaders *f] (ARG1 ?a ?f) [in Paris *g] (ARGM-TMP ?a ?g)
(ARGM-PNC ?a [to urge people to vote No: [urge *h] [people *i] (ARG1 ?h ?i)
(ARG2 ?h [to vote No: [vote *j] (ARG0 ?j ?i)]) ])
```

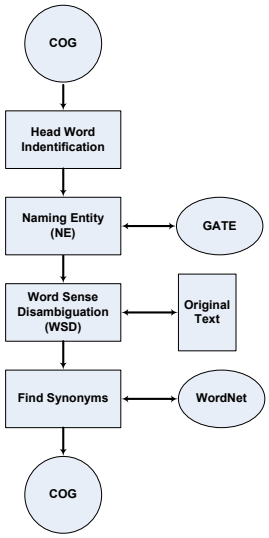

Figure 3: Synonyms Finder (SF) Work Flow

2.5.5 Synonyms Finder (SF)

As we mentioned earlier, a node in the COG representation is a constituent of words. The Synonyms Finder (SF) component provides four phases that are required prior matching between concepts. The first phase identifies the head word of constituent in each node in the COG representation. The second phase is a Naming Entity (NE) phase. The head word might be a ”person”, ”organization”, ”location”, etc... In this case, the head word is named based on its type. GATE (General Architecture of Text Engineering), the project at University of Sheffield, is responsible for naming entities in our proposed system. The third phase is a Word Sense Disambiguation phase (WSD) which extracts the correct sense for each head word in each node in the COG representation from the original text. Finally, the last phase enriches the head words with their synonyms (especially, hyponyms) from WordNet. The synonyms for each head word in a constituent are added to its original node in the COG representation with a membership degree that reflects the certainty of each chosen synonyms to the original head word. The main objective of this membership degree is to solve the pragmatic problem found in the natural language understanding. Figure 3 illustrates the work flow of the Synonyms Finder (SF) component.

2.5.6 Concept Matcher (CM)

In this component, a new conceptual similarity scheme is proposed. CM follows a new matching algorithm to compare among concepts placed in the proposed representation.

2.6 Conceptual Similarity Scheme

As we mentioned earlier, COG representation comprises different levels of conceptual graphs. The root graph consists of the main concepts regarding what a sentence is about. The other nested conceptual graphs provide more details or explanation of the meaning of a sentence. We refer to concepts listed in a main graph as main concepts and concepts listed in other nested graphs as supplementary concepts. The similarity scheme comprises the following layers of similarity among concepts and relations placed in nested conceptual graphs.

2.6.1 Similarity Layers

The first similarity layer called the main layer or the layer of relevance. The main layer compares between main concepts listed in a main graph. The similarity of the main layer presents a fundamental unit of

document relevance to a query sentence. The second similarity layer called the nested layer. It compares between supplementary concepts listed in other nested graphs. The last similarity layer called the relation layer. In this layer, relations are considered in the similarity process. There are two types of concepts; constituent and target word. In the former, there is only one relation associated with each constituent. In the latter, there are two relations associated with each target word.

2.6.2 Matching Scheme

The proposed matching scheme provides new concept-based matching measure. The new matching scheme compares between original concepts and their synonymous placed in the COG representation. There are four basic possible combinations of matching between two COG representations.

For example, if there are two sentences named ”sent A” and ”sent B” which are represented by two COG representations named ”COG A” and ”COG B” respectively. Then, the similarity between ”COG A” and ”COG B” representations comprises one or more forms of the following basic similarity forms between similarity layers: Similarity between main concepts placed in ”COG A” and main concepts placed in ”COG B”, Similarity between main concepts placed in ”COG A” and supplementary concepts placed in ”COG B” Similarity between supplementary concepts placed in ”COG A” and main concepts placed in ”COG B”, and Similarity between supplementary concepts placed in ”COG A” and supplementary concepts placed in ”COG B”

In case one, ”sent A” and ”sent B” are referred to the same general idea. In case two and three, ”sent A” and ”sent B” share common concept that might have different role in each sentence. In case four, ”sent A” and ”sent B” might not related to each other but their supplementary concepts are related. The combination between case one and other cases means that ”sent A” and ”sent B” are conceptually similar to each other. The increase in similarity of nested concepts given that the similarity of main concepts presents a deep semantic similarity among sentences. The similarity is computed based on the level of the concept and its associated relations. First, the matching algorithm compares between the main concepts listed in the main graph. Then it compares between the supplementary concepts listed in the nested graphs. More investigation is required to handle the hierarchy manner of the COG representation.

3 Preliminary Experiment and Results

Search engines play an important role in the World Wide Web. One of the most important features in a search engine is the ranking algorithm for relevant documents. Recently, most of the search engines rank relevant document based on the hyper links that refer to documents. However, ranking for document relevance should be relied on document content. Our preliminary experiment is a search engine that ranks the relevant documents based on the main concepts presented in the content of a document. Some basic operations are required for information retrieval including stemming and eliminating stop words.

3.1 Data Set

The data set is a sample collection of 20 documents extracted from the Reuters-21578. The Reuters data was originally collected and labeled by Carnegie Group, Inc. and Reuters, Ltd. in the course of developing the CONSTRUE text categorization system. The degree of overlap between topics classes is quite high in this data set.

3.2 The Use of the COG Representation

One of the main concerns about a search engine is to avoid retrieving irrelevant documents. As we mentioned earlier, the COG representation provides a clear separation between the main concepts and other supplementary concepts in a sentence. We use the conceptual separation feature to focus on main concepts and to avoid supplementary concepts in a sentence. Each document is conceptually analyzed to extract its main and supplementary concepts for each sentence.

3.3 Concept-based Indexing

Main concepts in a sentence are indexed as features associated with each document. For each query, the system searches for either concepts listed as features or keywords placed in the content of a document. Supplementary concepts are not considered as features. However, supplementary concepts are considered as regular text content. Features of a document encapsulate main concepts listed in main graphs in the COG representation. Searching for main concepts increases the opportunity in retrieving concept-based relative documents to a query and decreases the possibility of retrieving concept-based irrelevant documents to a query. The retrieving algorithm is biased to the features of a document much more than the content of a document. Therefore, the ranking algorithm gives higher score to the document which, its features contain a query. However, the ranking gives lower score to the document which, its content contain a query.

3.4 Information Retrieval Evaluation

Precision and recall are the basic measures used in evaluating search strategies.

$$Precision = \frac{Number of Relevant Records Retrieved}{Total Number of Irrelevant and Relevant Records Retrieved} \tag{1}$$

$$Recall = \frac{Number of Relevant Records Retrieved}{Total Number of Relevant Records in the Database} \tag{2}$$

3.5 Results

Table 4 provides a sample for queries map. The columns represent the retrieved documents based on the system output without using the COG representation, and the rows represent the correct retrieving documents. For example, table 1 says that two documents retrieved by the query ”cotton” were correctly retrieved, and one document that belongs to query ”crude” was incorrectly retrieved by query ”acquisition”. The same applies to Table 6 but with using the COG representation. Tables 5 and 7 describe the precision and recall values with and without using the COG representation.

Figure 4: Sample of Retrieving Documents without COG

Crude	Grain	Corn	Cotton	Trade	Acq.	Retrieved as
2					1	Crude (2)
	2					Grain (2)
		1				Corn (1)
			2			Cotton (2)
				5		Trade (5)
1		1			2	Acq. (3)
	1	1	3	5	1	No Topic (7)

Figure 5: Precision and Recall without COG

Query	Precision	Recall
Crude	0.67	0.67
Grain	0.67	1.0
Wheat	1.0	1.0
Corn	0.33	1.0
Sorghum	1.0	1.0
Cotton	0.4	1.0
Trade	0.5	1.0
Rice	1.0	1.0
Acquisition	0.67	1.0
Overall	69.3%	96.3%

We assigned a threshold value = 0.09. If the scoring value for a retrieved document is less than 0.09 a document will not be retrieved.

3.6 Discussion

In the first case (without COG), the search engine retrieve any document which its content contains the query keyword without considering whether the meaning of a document is related to the keyword query or not. Therefore, the search engine has very high recall 96.3% because the system retrieves all documents either related or not. Moreover, the search engine has low precision 69.3% because the system is not accurate enough to retrieve only the documents that are conceptually related to the query keyword.

In the second case (with COG), the situation is reversed. The search engine has a very high precision 100% because the system retrieves only documents that are conceptually related to the query. However, the

Figure 6: Sample of Retrieving Documents with COG

Crude	Grain	Corn	Cotton	Trade	Acq.	Retrieved as
2						Crude (2)
	2					Grain (2)
		1				Corn (1)
			2			Cotton (2)
				2		Trade (5)
					1	Acq. (3)
				3	2	No Topic (7)

Figure 7: Precision and Recall with COG

Query	Precision	Recall
Crude	1.0	1.0
Grain	1.0	1.0
Wheat	1.0	1.0
Corn	1.0	1.0
Sorghum	1.0	1.0
Cotton	1.0	1.0
Trade	1.0	0.4
Rice	1.0	1.0
Acquisition	1.0	0.33
Overall	100%	86%

system has a moderate recall 86% because some documents that are conceptually related to the query are not retrieved by the search engine to ensure its precision.

4 Conclusion and Future Work

Most of recent text mining systems consider only the presence or the absence of keywords in text. Statistical analysis of word frequencies is not sufficient for representing the meaning of text. Concept mining is the area targeting the meaning of a word rather than a word itself. The main contribution of this work lies in the proposed concept mining model which captures and represents the semantics in text based on concepts. The concept mining model discovers structured knowledge to be utilized in several applications. At the knowledge representation level, the system extracts and joins between verbs and their arguments with semantic relations. The representation structure is a projection of the syntactic data represented in verbs and their argument to conceptual information represented in concepts and their semantic relations. The whole framework is based on concepts. Hence, the framework could be applied to unstructured, structured, or semi-structured data. Although the quality of the results is promising and indicates that it is going in the right direction, more experiments need to be done to evaluate the performance of matching between documents based on their representations.

There are number of issues that have to be addressed in future work. One direction is to use fuzzy logic in choosing the appropriate sense from WordNet based on the context. Another direction is to formulate and develop a similarity measure that can handle the hierarchical manner in the COG representation. One interesting future direction is to transform the COG representation to an ontology language especially OWL. By this transformation, the semantics of the content of a web page represented in COG will be available in a standard format for the World Wide Web applications especially for semantic web applications.

References

[1] D. Jurafsky and J. H. Martin, *Speech and Language Processing*, Prentice Hall Inc., 2000.

[2] C.J. Fillmore, "The case for case," in *Universals in Linguistic Theory*, E. Bach and R.T. Harms, Eds. 1968, pp. 1–88, Holt and Rinehart and Winston.

[3] D. Gildea and D. Jurafsky, "Automatic labeling of semantic roles," *Computational Linguistics*, vol. 28, no. 3, 2002.

[4] M. Collins, *Head- Driven Statistical Model for Natural Language Parsing*, Ph.D. thesis, University of Pennsylvania, 1999.

[5] S. Pradhan, W. Ward, K. Hacioglu, J. H. Martin, and D. Jurafsky, "Shallow semantic parsing using support vector machines," in *the Proceedings of the Human Language Technology/North American Association for Computational Linguistics (HLT/NAACL-2004)*, Boston, 2004.

[6] M. Surdeanu, "Using predicate argument structures for information extraction," in *ACL*, Sapporo, Japan, 2003.