IOSR Journal of Computer Engineering (IOSR-JCE)
e-ISSN: 2278-0661,p-ISSN: 2278-8727, Volume 16, Issue 5, Ver. V (Sep – Oct. 2014), PP 24-27
www.iosrjournals.org

Mining Conceptual Relations from Textual Web Content Using Leximancer

[1]Ms.C.Thavamani, [2]Dr.A.Rengarajan,
[1]*Research Scholar, Bharathiar University, Coimbatore.*
[2]*Associate Professor,Veltech Multi Tech SRS Engineering Avadi, Chennai .*

Abstract: *Concept mining is a process that focuses on extracting ideas and concepts found in documents. The approach is somewhat similar to text mining, with the main difference being that mining a text focuses on the extraction of information rather than ideas. In this paper, we propose concept-based text representation, with an accent on using the proposed representation in different applications such as information retrieval, text summarization, and question answering. This work presents a new prototype for concept mining by extracting the concept-based information from a raw text using leximancer. 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 is going to apply to the domain of information retrieval, and give an assertion for proceeding in the right directions of this research.*

Keywords: *Concepts, Similarity, Extraction, Information, Leximancer,Mining.*

I. 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].

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 [2]. 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.

II. 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.
- ✓ Identifying relations between concepts to represent the semantic roles of the extracted concepts.

Most of the data representation techniques are based on word and/or phrase analysis of the text. The statistical analysis of a term (word or phrase) frequency captures the importance of the term within a document. However, to achieve a more accurate analysis, the underlying data representation should indicate terms that capture the semantics of the text from which the importance of a term in a sentence and in the document can be derived. A new concept-based representation that relies on the analysis of the sentence semantics, rather than, the traditional analysis of the document dataset only is introduced.

The proposed conceptual ontological graph representation denotes the terms which contribute to the sentence semantics. Then, each term is chosen based on its position in the proposed representation. Lastly, the selected terms are associated to their documents as features for the purpose of indexing in the text retrieval. Experiments using the proposed conceptual ontological graph representation in text retrieval are conducted. The

evaluation of results is relied on two quality measures, the precision and the recall. Both of these quality measures improved when the newly developed representation is used to enhance the performance of the text retrieval.

2.1 System overview

The proposed system has two main objectives:

- Concept-based text representation
- 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.

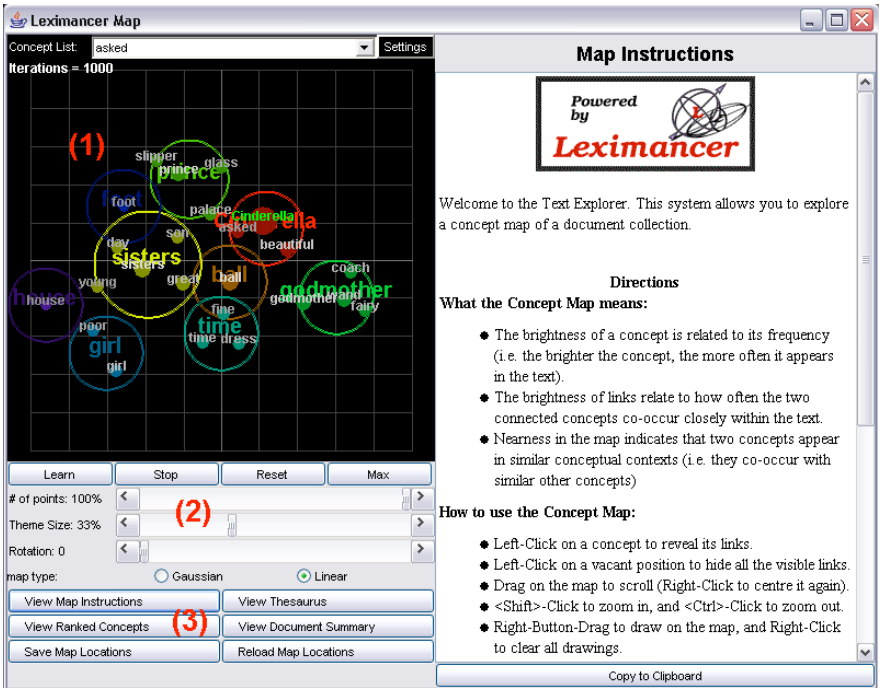
2.2.1 Leximancer

We use the most special tool named "Leximancer" to display the information by means of a conceptual map that provides an overview of the material, representing the main concepts contained within the text and how they are related. Leximancer is text mining software that can be used to analyze the content of collections of textual documents and to visually display the extracted information in a browser. Leximancer combines the discovery of quantified relational information between concepts with the flexibility and dynamics required to analyze natural language in real-life settings.

Operational Capability

Automation

Given the increasing amount of text requiring examination, and the decreasing amount of time available in most work situations, Leximancer can process many different formats and styles of text from many languages, and can automatically select and learn a set of concepts that characterize the text.



Clarity

Leximancer is designed to provide a clear and transparent conceptual analysis of text. It maintains a symbolic representation of each concept in the thesaurus, similar to the headwords in a traditional thesaurus. It is also designed to be transparent: a user can inspect a ranked list of words which make up each concept, and can easily drill down to the text to inspect the validity and nature of the induced abstract relationships.

Deep Meaning

Leximancer generates similarly structured conceptual maps from texts which have similar meanings, even when the documents use different styles, formats, or even languages. This also shows that it is not easy to conceal the patterns of meaning from Leximancer by using veiled speech, dialect, or non-standard grammar.

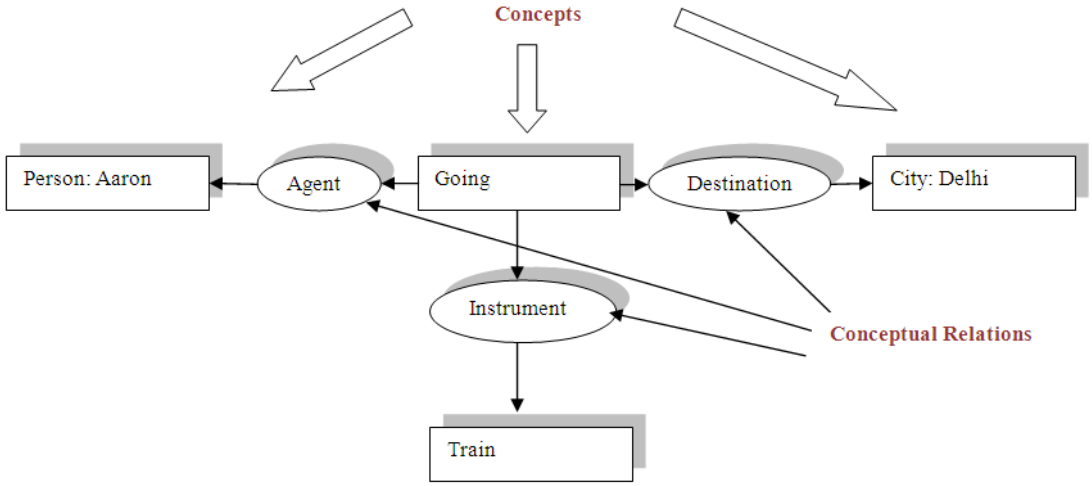
Completeness

This offers a one-stop automatic system to take the user from a large collection of raw text to a concept map, while allowing inspection and validation of all the intermediate steps.

Efficiency

Leximancer system was designed from the outset to be simple and fast. A normal modern PC or notebook computer is quite adequate for analysing 1 GB of text or more. The Leximancer algorithms are scalable and so much larger amounts of text could be processed using more memory and multi-threaded processing.

Conceptual Graph : Example
"Aaron is going to Delhi by train"



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 [6]. 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.

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 [7] for the significant concept (head word) in a constituent. Finally, the Concept Matcher (CM) compares among concepts represented in the COG representation.

2.3 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.

III. Conclusion and Future Work

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 using concept mining tool leximancer. 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 proposed method is promising and indicates that it is going in the right direction, more experiments need to be done in future to evaluate the performance of matching between documents based on their representations.

References

[1]. D. Jurafsky and J. H. Martin, Speech and Language Processing, Prentice Hall Inc., 2000
[2]. D. Gildea and D. Jurafsky, “Automatic labeling of semantic roles,” Computational Linguistics, vol. 28, no. 3,2002.
[3]. V. Bharanipriya & V. Kamakshi Prasad, “Web Content Mining Tools: A Comparative Study” *International Journal of Information Technology and Knowledge Management January-June 2011, Volume 4, No. 1, pp. 211-215.*
[4]. Dunham, M. H. 2003. Data Mining Introductory and Advanced Topics. Pearson Education.
[5]. www.leximancer.com
[6]. http://www.obitko.com/tutorials/ontologies-semantic-web/conceptual-graphs.html
[7]. http://wordnet.princeton.edu/