International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395 -0056
Volume: 04 Issue: 01 | Jan -2017 www.irjet.net p-ISSN: 2395-0072

SEMANTIC BASED DOCUMENT CLUSTERING USING LEXICAL CHAINS

SHABANA AFREEN[1], DR. B. SRINIVASU[2]

[1]M.tech Scholar, Dept. of Computer Science and Engineering, Stanley College of Engineering and Technology for Women, Telangana-Hyderabad, India.

[2]Associate Professor, Dept. of Computer Science and Engineering, Stanley College of Engineering and Technology for Women, Telangana -Hyderabad, India.

Abstract – *Traditional clustering algorithms do not consider the semantic relationships among documents so that cannot accurately represent cluster of the documents. To overcome these problems, introducing semantic information from ontology such as WordNet has been widely used to improve the quality of text clustering. However, there exist several challenges such as extracting core semantics from texts, assigning appropriate description for the generated clusters and diversity of vocabulary.*

In this project we report our attempt towards integrating WordNet with lexical chains to alleviate these problems. The proposed approach exploits the way we can identify the theme of the document based on disambiguated core semantic features extracted and exploits the characteristics of lexical chain based on WordNet. In our approach the main contributions are preprocessing of document which identifies the noun as a feature by performing tagging and lemmatization, performing word sense disambiguation to obtained candidate words based on the modified similarity approach and finally the generation of cluster based on lexical chains.

We observed better performance of lexical chain based on the chain evaluation heuristic whose threshold is set to 50%. In future we can demonstrate the lexical chains can lead to improvements in performance of text clustering using ontologies.

Key Words*: *Text clustering, WordNet, Lexical chains, Core semantic features, cluster, semantic, candidate words.

1. INTRODUCTION

This section provides detail about semantic web mining and a more explained introduction to text clustering followed by few examined problems facing for text clustering and their possible solutions.

1.1 Text Clustering

With the increasing information on Internet, Web mining has been the focus of information retrieval. Now a day's Internet is being used so widely that it leads to a large repository of documents. Text clustering is a useful technique that aims at organizing large document collections into smaller meaningful and manageable groups, which plays an important role in information retrieval, browsing and comprehension [1].

1.2 Problem Description

Feature vectors generated using Bow results in very large dimensional vectors. The feature selected is directly proportional to dimension.

Extract core semantics from texts selecting the feature will reduced number of terms which depict high semantic conre content.

The quality of extracted lexical chains is highly depends on the quality and quantity of the concepts within a document. i.e., larger the chain clearer the concepts.

Several other challenges for the clustering results:

synonym and polysemy problems. There has been much work done on the use of ontology to replace the original word in a document with the most appropriate word called word sense disambiguation(WSD).

Assign distinguished and meaningful description for the generated clusters

1.3 Basics and background knowledge

1.3.1WordNet

WordNet is the product of a research project as Princeton University. it classify the word in two categories i.e., content word and function word, content words dealt with noun, verb, adverb and adjective which form a set of synonyms called synsets. The synsets are organized into senses, giving thus the synonyms of each word, and also into a hyponym/hypernym (i.e., is-A) and meronym/holonym (i.e., Part-of) relationships[2].

1.3.2 Semantic similarity

Semantic similarity plays an important role in natural language processing, . In general, all the measures can be grouped into four classes: path length based measures, information content based measures, features based measures and hybrid measures [3].

1.3.2.1 Path-based Measures

The main idea of path-based measures is that the similarity between two concepts is a function of the length of the path linking the concepts and the position of the concepts in the taxonomy.

The various path based measure include Shortest path measure, wu and palmer's measure, leakcock and chodorow's measure and li's measure.

The information based measure are resnik's measure,lin's measure, jiang's measure

Feature based measure is independent on the taxonomy and the subsumers of the concepts. The final approach i.e., hybrid measure combined the ideas presented above[4,5,6,7].

1.4 Word sense disambiguation

We adopt the WSD procedure which is given by [8], aim to identify the most appropriate sense associated with each noun in a given document based on the assumption that one sense per discourse. The WSD approach can be described as follows. Let N= $\{n_1, n_2, ...n_p\}$ denote the set of all senses associated with the noun n_i according to the WordNet ontology. We determine the most appropriate sense of a noun n, by computing the sum of its similarity to other noun senses in d as follows.

$$c_i = \max_{c_{ik} \in C_i} \sum_{n_j \in d} \max_{c_{jm} \in C_j} s(c_{ik}, c_{jm}) \quad \text{(Eq1.10)}$$

Where $s(c_{ik},c_{jm})$ is the similarity between two senses. We restrict to the first three senses for each synset to participate in this computation for several reasons as given by [9].

First, the senses of a given noun in the WordNet hierarchy are arranged in descending order according to their common usage. Furthermore we compare the clustering results on using only the top three senses against using all senses of a noun, the former yields similar clustering results at a reduced computation cost to the latter.

1.5 Lexical chains

Lexical chains are groups of words which exhibit lexical cohesion. Cohesion is a way of getting text together as a whole. Lexical cohesion is exhibited through cohesive relations. They have classified these relations as:

1. reiteration with identity of reference
2. reiteration without identity of reference
3. reiteration by means of super ordinate
4. systematic semantic relation
5. non systematic semantic relation
6. The first three relations involve reiteration which includes repetition of the same word in the same sense, the use of a synonym for a word and the use of hypernyms for a word respectively. The last two relations involve collocations i.e., semantic relationships between words that often co-occur. Lexical chains in a text are identified by the presence of strong semantic relations between the words in the text[10].

2. LITERATURE SURVEY

2.1 Introduction

Most of the existing text clustering methods use the bag of words model known from information retrieval where single terms are used as features for representing the documents and they are treated independently. Some researches recently put their focus on the conceptual features extracted from text using ontologies and have shown that ontologies could improve the performance of text mining.

2.2 Semantic ontology for text clustering

Current approaches for semantic ontology for text clustering can be divided into two major categories, namely, concept mapping and embedded methods[11].

Concept mapping methods simply replace each term in a document by its corresponding concepts(s) extracted from an ontology before applying the clustering algorithm. These methods are appealing because they can be applied to any clustering algorithm[12]. Furthermore, the mapping of terms into concepts incurs only a one-time cost, thus allowing the clustering algorithm to be invoked multiple times (for different cluster initialization, parameter settings etc.,) without the additional overhead of re-creating the concepts. However, their main limitation is that the quality of the clusters is highly dependent on the correctness of the WSD procedure used. Embedded methods, on the other hand, integrate the ontological background knowledge directly into the clustering algorithm. This would require modifications to the existing clustering algorithm, which often leads to substantial increase in its both runtime and memory requirements. Instead of performing WSD explicitly, these methods assume the availability of a distance/ similarity matrix for all pairs of words in a text corpus computed based on the WordNet's concept hierarchy. Since a word can be mapped to several synsets, multiple pairs may exist between any two words and the algorithm has to decide which path to use when computing the distance measure. Thus embedded methods are still susceptible to incorrect mapping issues related to WSD.

2.3 Proposed approach

The proposed approach exploits the relations to provide a more accurate assessment of the similarity between terms for word sense disambiguation. Furthermore, we implement lexical chains to extract a set of semantically related words from texts, which can represent the semantic content of the texts. Although lexical chains have been extensively used in text summarization, their potential impact on text clustering

problem has not been fully investigated. Our integrated way can identify the theme of documents based on the disambiguated core features extracted, and in parallel downsize the dimensions of feature space.

The work exploits this characteristic of lexical chains - each lexical chain is an indicator of a topical strand and the segments connected by it belong to the same topic. We work on the premise that, instead of just computing the lexical chains with respect to a single document, if we compute the chains across documents, we are in effect chaining together those documents belonging to a topical strand.

3. ARCHITECTURE

This section deals with the method used, architecture and statistics of semantics for the construction of lexical chains based on semantic knowledge database such as WordNet.

3.1 System architecture

Text document clustering can greatly simplify browsing large collections of documents by reorganizing them into a smaller number of manageable clusters.
Preprocessing the documents is probably at least as important as the choice of an algorithm, since an algorithm can only be as good as the data it works on. While there are a number of preprocessing steps, that are almost standard now, the effects of adding background knowledge are still not very extensively researched.

We preprocess the document by running it through a tokenizer; we then filter out all non-noun words identified in the WSD stage. (The WSD here referred as the process of word sense disambiguation where original word is being replaced by the most appropriate sense based on the similarity sense from WordNet).The result is a sequence of nouns which appear in the text along with its sense. We refer to these as 'candidate words'. We base our algorithm on the WordNet lexical Database. WordNet is used to identify the relations among the words. We use identity, synonym, hypernym, meronym relations to compute the chains. Our algorithm works by maintaining a global set of lexical chains, each of which represents a topic.

We now compute the lexical chains corresponding to each of the candidate words by looking up the synsets for the word from WordNet. We then traverse the global list of lexical chains to identify those chains with which it has a identity, synonym, hypernym and meronym relations. We refer to these identified lexical chains as potential chains. we refer to these identified lexical chains as potential chains. If the candidate word has no relation with any of the chains in the global list, a new potential chain is created.

A chain is selected from the global set based on the score of representative i.e., a threshold value. We have selected the threshold in such a way that the chain which is greater than the threshold value is selected as the potential chain for the document.

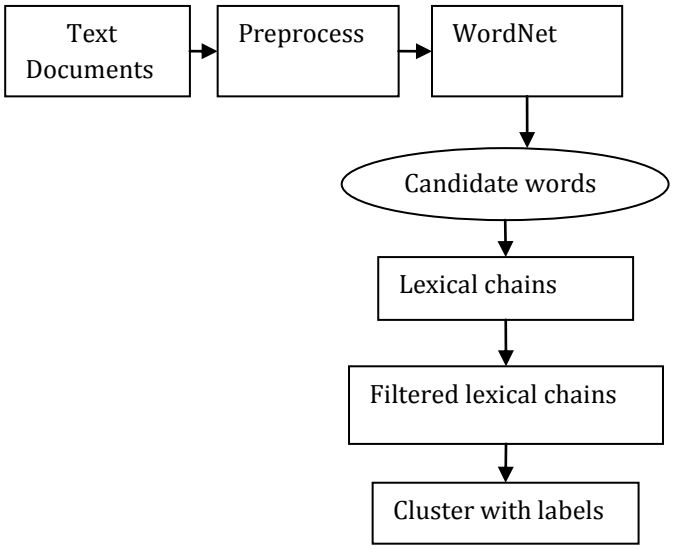

Figure 3.1 Architecture

3.2 Candidate word generation

It implements a method to obtain description of a synsets,

Description of synsets: Let C={c_1, c_2, c_3, c_4.....c_k} be the set of synsets in a document, c_i € C. let lemma(c_i) be the set of words that constitute a synsets of c_i. let gloss(c_i) be the definition and examples of usages of c_i. let related(c_i) be the union of the synonym, hypernym, hyponym.

Scoring mechanism:
The scoring mechanism assigns an n word overlap the score of n^2. This gives an n- word overlap a score that is greater than the sum of the scores assigned to those n words if they had occurred in two or more phrases, each less than n words long.

$$\text{Score}(\text{DES}(c_i), \text{DES}(c_j)) = \sum_k (N_k)^2$$

Modified measure: In order to take the full advantage of both explicit and implicit semantic relations between synsets such as is-a and has-par links, we define a new similarity measure that combines both measures as below.

$$\delta(c_p, c_q) = \frac{2d + S}{L_p + L_q + 2d + S}$$

Where S= log(score(DES(c_p),DES(c_q))+1), this method not only reflects structure information of synsets such as distance, but also incorporates content meaning of synsets in the ontology. It integrates well with explicit and implicit semantic between synsets in ontology.

3.3 Lexical chain

Lexical chain represents semantic relations among the selected word senses of the words appearing in that lexical chain. Each node in a lexical chain is a word sense of a word, and each link can be identity, synonym, hypernym, meronym relation between two word senses.

In order to extract the core semantics, the semantic importance of word senses within a given document should be evaluated first, generally, let N= {n_1, n_2, n_3 .....n_k} be the set of nouns in a document d and let f={f_1,f_2,f_3 .....f_p} be the corresponding frequency of occurrence of nouns in d. let C={c_1, c_2, c_3, ...c_k} be the set of disambiguation concepts that corresponding to N, given a document d, a set of nouns N, a set of frequencies F and a set of concepts C, let W={w_1, w_2, w_3 .....w_n} as the set of corresponding weight of disambiguated concepts in C, if c_i(c_i€C) is mapped from n_k and n_m(n_k, n_m €N). then the weight of c_i us computed by based on the weighted concepts, we give following definition.

$W_k=f_k+f_n$

Score of concept:

Let C={c_1, c_2, c_3 .....c_n} be the set of disambiguated concepts(word senses), and let W={ w_1, w_2, w_3 .....w_q } be the set of corresponding weight of disambiguated concepts in C. let RN={identity, synonym, hypernym, meronym} be the set of semantic relations, and let R ={r_1 r_2=r_1, r_3, r_4 } be the set of the corresponding weight of relation in RN. Then the score of a concept c_i(c_i€C) in a lexical chain is computed by

$$S(c_i) = w_i \times r_1 + \sum_{k=3}^{4}\sum_{p=1}^{q}\{w_p \times H(c_i, c_p, k) \times r_k\}$$

$$\text{where } H(c_i,c_p,k) = \begin{cases} 1 & \text{if there exits an edge of } RN_k \text{ between } c_i \text{ and} c_p \\ 0 & \text{otherwise} \end{cases}.$$

A large value of S(c_i) indicates that c_i is a semantically important concept in a document. The relation weight r(r €R) depending on the kind of semantic relationship and it is in the order listed: identity, synonym, hypernym, meronym (thus r_1=r_2>r_3> r_4).

Score of lexical chain:

Let L= {L_1, L_2, L_3..... L_m} be a set of lexical chain of a given document. L_i€L, let c_i = { c_{i1}, c_{i2}, c_{i3} .....c_{iq} } be a set of disambiguated concepts in L_i let S(c_i) be the score of concept c_{il}(c_{il} € c_i). Then , the score S(c_{i1}) of lexical chain in a document is define as

$$S(L_i) = \sum_{l=1}^{q} S(c_{il})$$

Score of representative lexical chain:

Let Let L= {L_1, L_2, L_3..... L_m} be a set of lexical chains of a given document. Let L^R ={L_1^R, L_2^R .....L_n^R} (n≤m) be a set of representative lexical chains that satisfy the following criterion:

$$S(L_i^R) \geqslant \alpha \cdot \frac{1}{m}\sum_{j=1}^{m} S(L_j) \quad (i = 1,2,\ldots,$$

Where ∞ is a weighting coefficient that is used to control the number of the representative lexical chains to be considered.

We extract the weighted concepts in the lexical chain L^R composing set of core semantic features for the given document. It is these concepts can be then used to cluster the document.

4. IMPLEMENTATION

4.1Dataset Description

We have run our experiments on a small dataset of 30 documents retrieved from 20**Newsgroup** dataset. This was because we were unable to get a pre-clustered dataset and comparing the results would have been difficult. Hence, we limited our experiments to a small dataset to keep our experiments humanly tractable.

We keep the number of documents small in order to be able to do a qualitative analysis of the clusters formed as opposed to a quantitative one.

4.1 Pre-processing

Documents are process by passing through tokenizer by setting space as delimiter, and then we obtain the tokens. These tokens are further passed to a Tagger to assign the category for each token, in our case it is noun finally we pass to a lemmatizer to obtain the root form of the tagged word.

Here we are using wordNet lemmatizer for a part of speech tagging output. Lemmatizer return the same word if it is not found in WordNet or else returns the lemma.

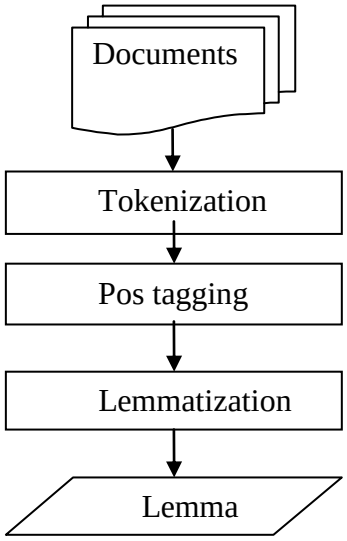

Figure 4.1: preprocessing

4.2 Performing Word Sense Disambiguation

Here we determine the most appropriate sense associated to each lemma by computing the similarity among three sense, the sense which assigned the highest score is consider as the probable sense. Finally we replace our lemma with the synset of that sense which is termed as candidate words.

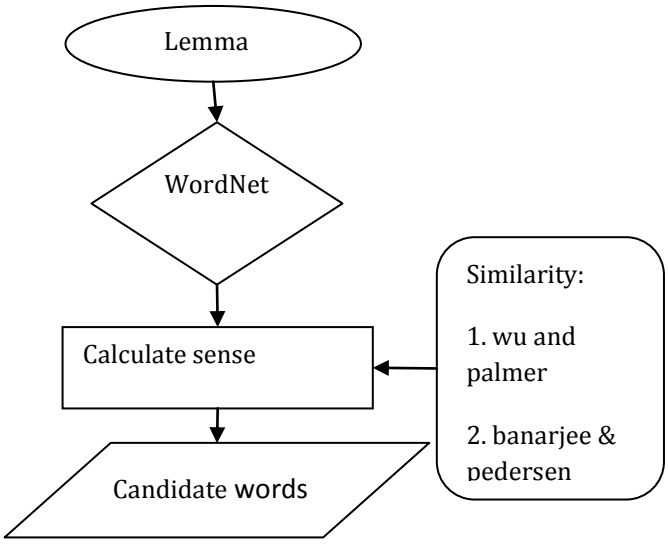

Figure 4.2 Calculating word sense disambiguation

4.3 Cluster Generation

The candidate words are assigned the weights and generated the score of concept by considering the relations. Then summing up all the relation score will give the concept score. This concept score further used to compute the score of lexical chain. we then evaluate the filtered chains, in order to select a subset of chains to which the document is added. We select all those lexical chains whose length exceeds heuristic value.

Our algorithm works on the assumption that lexical chains represent the theme of the document. And grouping documents based on these lexical chains results in the documents being clustered based on its theme. This work explores if and how the two following methods can improve the effectiveness of clustering through semantic similarity approach.

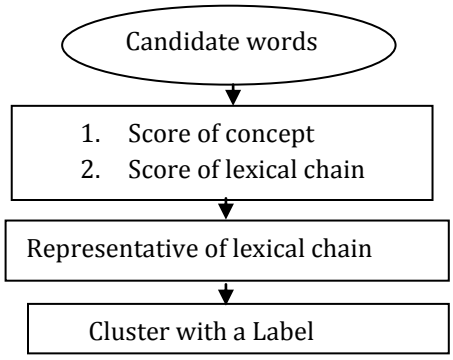

Figure 4.3 Generation of cluster

5. EXPERIMENT RESULTS

The results are analyzed after pre-processing step is illustrated in the below table 5.1. the first column specify the document id, second column specify the tokens and the third column specify the number of nouns obtained.

Table 5.1: sample for pos tagging.

Document id	Number of tokens	Number of Nouns
9	106	26
10	129	19
29	49	26
30	70	19

The below table shows the chains obtained for the documents. The first columns specify the document ids and their respective chains.

Table 5.2: Sample for lexical chains.

Document id	Chains

29	1. Bent, hang, knack 2. Design, designing 3. Station 4. Wagon, wagon 5. Tercel, tercelet, tierce 6. Corolla 7. Architect, designer 8. Sense, signified
10	1. Earn 2. Livestock, carcass 3. Argentina, Philippines 4. Bond, American, Canada
30	1. Earn 2. Nyse, take, income, stock, rate

Table 5.3 Number of Lexical chains generated.

Document id	Number of chains
29	8
30	2
10	4

6. CONCLUSIONS

Documents contain multiple topics and clustering them using hard clustering is very unnatural. Documents should ideally be clustered using soft clustering algorithms. Unfortunately, these algorithms work only for very small dimensions. The nature of lexical chains makes them suitable for clustering documents. Each lexical chain is considered as a topical strand and if two documents share a chain, then they contain the same topics.

This work presents a methodology for clustering using disambiguated concepts and lexical chains. A modified term-based semantic similarity measure is proposed for word sense disambiguation, and lexical chains are employed to extract core semantic features that express the topic of documents, which determining the number of clusters, and assigning appropriate description for the generated clusters. More importantly, we show that the lexical chain features (core semantics) can improve the quality significantly with a reduced number of features in the document clustering process. Although lexical chains have been widely used in many application domains, this study is one of the few researches which try to investigate the potential impact of lexical chains on text clustering.

In future work, we would like to perform our method on a larger knowledge base, such as Wikipedia. Moreover, since we have demonstrated that the lexical chains can lead to improvements in text clustering, the next work we plan to explore the feasibility of lexical chains in the text mining task.

ACKNOWLEDGEMENT

To Ammi
Affu is nothing without you
All that I am, all that I do, all that I ever had, I owe it all to you.
To pappa
I'm glad how your words have always given me life's real view,
"Winning is great, sure but if you want to do something in life run a marathon of life. I am there."
It is with a sense of gratitude and appreciation that I feel to acknowledge any well wishers for their king support and encouragement during the completion of the project.

I would like to express my heartfelt gratitude to my Project guide ***Dr. Srinivasu Badugu,*** *Coordinator of M.Tech in the Computer Science Engineering department, for encouraging and guiding me through the project. His extreme energy, creativity and excellent domain knowledge have always been a constant source of motivation and orientation for me to gain more hands-on experience and hence get edge over. I am highly indebted to him for his guidance and constant supervision as well as for providing necessary information regarding the project & also for his support in completing the project.*

REFERENCES

1. *Wei, Tingting, et al. "A semantic approach for text clustering using WordNet and lexical chains." Expert Systems with Applications 42.4 (2015): 2264-2275.*
2. *Miller, George A., and Walter G. Charles. "Contextual correlates of semantic ++*
3. *similarity." Language and cognitive processes 6.1 (1991): 1-28.*
4. Resnik, Philip. "Semantic similarity in a taxonomy: An information-based measure and its application to problems of ambiguity in natural language." *J. Artif. Intell. Res.(JAIR)* 11 (1999): 95-130.
5. Sánchez, David, et al. "Ontology-based semantic similarity: A new feature-based approach." *Expert Systems with Applications* 39.9 (2012): 7718-7728.
6. Budanitsky, Alexander, and Graeme Hirst. "Evaluating wordnet-based measures of lexical semantic relatedness." *Computational Linguistics* 32.1 (2006): 13-47.

7. Budanitsky, Alexander, and Graeme Hirst. "Evaluating wordnet-based measures of lexical semantic relatedness." *Computational Linguistics* 32.1 (2006): 13-47.
8. *Amine, Abdelmalek, Zakaria Elberrichi, and Michel Simonet. "Evaluation of text clustering methods using wordnet." Int. Arab J. Inf. Technol. 7.4 (2010): 349-357.*
9. Jayarajan, Dinakar, Dipti Deodhare, and Balaraman Ravindran. "Lexical chains as document features." (2008): 111-117.
10. Fodeh, Samah, Bill Punch, and Pang-Ning Tan. "On ontology-driven document clustering using core semantic features." *Knowledge and information systems*28.2 (2011): 395-421.
11. Termier, Alexandre, Michèle Sebag, and Marie-Christine Rousset. "Combining Statistics and Semantics for Word and Document Clustering." *workshop on ontology learning*. 2001.
12. Chen, Chun-Ling, Frank SC Tseng, and Tyne Liang. "An integration of WordNet and fuzzy association rule mining for multi-label document clustering." *Data & Knowledge Engineering* 69.11 (2010): 1208-1226.

BIOGRAPHIES

SHABANA AFREEN received her B.E and M.tech degree in computer science from Osmania university. Her research interests are natural language processing, semantic web.