Graph based Methods for Classification, Similarity and Representation using Noun, Verbs - A Survey

Rima Patel
M.Tech in Computer Engineering-Networking Tech.,
Institute of Technology, Nirma University
Ahmadabad, Gujarat, India

Bharati Kyada
MS in Computer Science,
Illinois Institute of Technology
Chicago, Illinois, USA

***Abstract*—This paper shows the survey of graph based methods for natural language processing [7]. Mainly the work focus on classification methods, similarity between the two phrases and representing this in terms of nouns verbs phrases. Classification done using clustering method, similarity is achieved by measuring similarity measures using various methods. Classification and measuring similarity basically depend on representation of the document which is based on nouns, verbs, phrases.**

Keywords—Natural Language Processing and Understanding, WSD, WSI, Graph.

I. INTRODUCTION

The information is increasing exponential, and the need for quick access is also emerging. This need has ledto natural language processing and understanding. Numbers of terms are required to represent the capabilities of the method, they are as follow: Event resolution, redundancy reduction, labeling, knowledge base, word sense disambiguation, word sense induction semantic[8] relatedness and similarity measures[10].These term if achieve in its best in method for natural language processing we can get best and fast access to the information. The main goal of this paper is to represent the survey on the capabilities of the various graph based method [9] for natural language processing and natural language understanding. It shows the analysis on various methods of classification using clustering technique, measuring similarity using machine learning, verbs phrases tree representation.

The sub methods used in these is following features: Here main concept of document down to certain level of detail. Here we replace all subset text into cluster. Means group of text which have similar meaning. Here we want to measure similarity of group of text and see a different method to find the distance. After finding distance we put the less distance word into one group by this way cluster is design. WSI (word sense induction) [7] is use to find the same meaning of the words and WSD (word sense disambiguation) [7] is use to reduce ambiguous sentence which eliminate multiple meaning sentence.

Section II of this paper describes classification technique, in this we will discuss methods for clustering based on syntax and semantic relation. These clustering methods are based on Chinese whisper[1] algorithm. Section III describes the representation of information using different method based on nouns verbs and their phrases. These methods are important because the way of representing helps in clustering and measuring similarities between the twonodes. Section IV describes the technique for similarity and syntactic measures. The measure here are calculated based on graph based methods.Metric is based on influence of the term, their relation with other term synonym and other similarities, and last the conclusion, which will discuss how these method are important in their context.

II. CLASSIFICATION USING CLUSTERING TECHNIQUE

Clustering is an effective method for classification, inthat we make a group a text into a one cluster base ondifferent parameter. We focus on performance of the allclustering method and key feature. As we know graph is aneffective method for traverse and find word. Use a differentcluster and connect each other and make a graph. For makingcluster Chinese whisper (CV) [1] is an effective method tomake a cluster and its works best n parallel and distributedenvironment.

A. Chinese Whisper [1]:

This algorithm is effective to make cluster of NLP [7]. Here we use a cosine similarity which is use in vector model and some parameter which is given as, V_i is a node or of graph. V_{ij} is edge of graph which gives a similarity of two nodes. So the weight of similarity is represented as edges. Here we focus on undirected graph. Below are the steps for implementing Chinese Wisper [1] algorithm. Step-1 first we put all node in to different cluster and then find an edge of V_i to V_j. Step-2 here we select a node which have a higher V_{ij} then we put it in one cluster. Step-3 repeat this iteration until all node note covered. Here in CW does not cover the nodes which are in between the two different cluster node. So for that type of node CW is fail. One more problem in this method is that it is hard classification means after generating a cluster we are not able to change it.

Method-1:

In this method we put a vertex as words and connect a word with other word when co-occurrence of that word s is more than one time in context. Weight is denoted by the number of occurrence of word. By this way we implemented a graph. This method is dividing into 5 sections all section some work toward the result. (a) In this, connect a word with other word when co-occurrence of that word. (b) Create a cluster with vertex ad words and edge as frequency of co-occurrence. (c) Here we find a quality of cluster by different parameter so that if any unrelated data in cluster just remove it. (d) Here we remove noun if it's put some noise. Noise means it's some nous are use which occur some defect on meaning of target word then remove it. (e) Create a cluster with word pair and their frequency smoothing reduce sparseness of graph.

Method-2:

In this method is use CW which is effective in linear cluster algorithm to implement a graph. Here we put words into cluster and then find a similar meaning word into one cluster. Here we use a weighted undirected graph to represent an edge. Weight is a co-occurrence of that word in context. In document clustering it find a set of similar words and put into one class and find a weight of that co-occurrence. CW algorithm partitions weighted undirected graphs. It finds groups of nodes that broadcast the same message to their neighbors.

Method-3:

Main purpose of this method is to identify the similar sentence and remove that sentence, So that we reduce a number of clusters. For that we select an object from cluster and compare with other cluster object. Here we use two cluster method used to identify the sentence that are similar context .in this method we select an object from cluster and compare that object with other node of cluster rest of the objects of each cluster exceeds a threshold. A threshold [7] is used in both methods for getting approximate minimum or maximum values to decrease the exponential complexity of the methods discussed.

III. METHODS FOR REPRESENTING TEXT IN GRAPH FORMAT

In this methods, we analyze various algorithm used for making graph of relevant information from collection of documents. We take information in form of text and represent it in graph.

Method 1: Make graph relation to find accurateanswers for Questions (user query) fromtext:

It uses subjects, verbs, objects and concepts of text entailment. The subject of sentence and verbs are implemented in graphical structure and use sentence semantic analysis for encoding the parser tree generated from text. We derive information from paired of text between previous two sentence using text entailment. It use Graph-based semi supervised [7] machine learning (SSL)[2] for question-answering of query. Graph summarization algorithm use previous data for assuming vertices of dataset, group of similar data are represented using new vertex and same label.

Performance and conclusion: We apply SSL [2] algorithm for improving the performance of task. It gives effectiveness to real time application. It uses for entailment for text and summarization of similar data representation.

Method 2: Graph based matching algorithm for Syntactic and Semantic Relation assignment:

Most of the verbs, nouns come from the sentence, predicates and Phrases. The semantic relation between nouns and clauses and noun and classes is includes casual, spatial, independent clause, time related and quality. In this method, the most important verb is head Word of the sentence. It takes pair of text form each sentence and assigns relation between them. It also relay on previous pair and also uses user feedback for document and gives more accurate relation for current pair.

Performance and conclusion: It gives better result for inter-clause semantic relation. The structure of this algorithm is subject-object-verb-noun-clause. This methodology depends on which domain it is used.

Method 3: In degree algorithm to improve English word sense disambiguation:

It uses nouns, verb, adverbs and group of words that containverbs and subject for that finding most appropriate meaningfulword from the document. This method based on WSD (Word Sense Disambiguation) [7] which use concept of unsupervised learning. It uses resources of Wordnet[3] and SemCor[3]. This graph creates nodes indicating word and edges that representing the weighted value between two words. The maximum in degree value is selected as sense. Here we remove the word which has more than two meaning like similar attributes, verb groups.

Performance and conclusion: We use WordNet[3] resource forimproving the performance of WSD[3]. Here one limitation found that is SemCor Expansion is still not useful for nounsperformance. It is current research topic.

Method 4: Event indexing using various actors:

It is based on current event, past and present event, time of day,casual, purpose to form relation between different concepts.It use intelligent summarization system based on study ofmental processes (cognitivepsychology [4]) such as languageuse, problem, and attention. It uses summarization [4] approachfor making clusters. Vector representation include actors forcurrent event, relation between two events, occurrences ofevents, goal of user for current event. Time information is comes from the time which is given in sentence. For temporal relation use WordNet [3] lexical resources. Location basednoun words are using spatial information which depend onspace. It also resolve pronoun according to definition. Casualindexing is used for making relation between subject and object.

Performance and conclusion: This model based onsummarization [4] of event indexing. For measuringperformance, we calculate precision and recall. This approach gives best case for text which containing more events andactions.

Method 5: Methods for cause-effect (event) [5] relation repository [5]:

After getting user feedback for document, we have toagain ranking the information for more accurate result. Wedelete and filter some nouns, erroneous terms. It recursivelybuilding pattern for terms, re-rank[5] the terms and thenfiltered. Here we use threshold value for elimination of terms.For more accuracy we use bootstrapping algorithm whichdepends on human based evaluation. We also collect the wholedata in term associated with cause-effect relation and usercan use as input any verb expressing causality. The objectivesof cause-effect relation learning is similar to those of anygeneral open domain relation extraction problems.

Performance and conclusion: It gives semantic relation between nouns, showed 89 percentage accuracy. The success of this framework also open some challenges task.

Method 6: Method for subgraph matching approach for parsed annotated [6] text:

For event recognition it search for graph which is isometrichaving equal values, and dependent relation of previous event.The union of shortest dependent path is given as a result.

Performance and conclusion: Searching for aphorism[6] is NP-hard problem. So we have to merge various events forbetter performance, increasing the precision value. It is usedfor biological event rules.

IV. METHODS MEASURING THE SIMILARITIES AND MERGING CLUSTERS BASED ON SIMILARITY:

Method 1: Tree Pair method used for text entailment:

Tree pair method is the method based on machine learning.In this tree has been created and the similarities betweenthe two trees has been measure by computing the relationbetween these two. Two identify the relation a tree kernelfunction is used. Text entailment is achieved by this as it ismatching the substructures between the two trees. Here weare using supervised learning therefore the matching tree is a hypothesis tree [7](training set). Overview: As large space has been required this approach is not practically possible, prototype for this is not yet created. In this method syntactic tree pair matching has been done.

Method 2: Method based on distance between the nodes in the graph:

Similarities and relationship between the graphs can bemeasure by calculating the distance between the nodes. Manyfunction such as cosine similarities, Euclidean distance are there to measure the distance. In this method a random walk function is used which is merge with shortest path algorithm [7] to optimize the solution. After the algorithm is applied we can get the distance between the two nodes, which gives us the similarity measure between the two nodes. Pseudo inverse la-pelican function [7] is used to calculate the function for defining shortest path algorithm. Algorithm: Create a graphwith [7] n nodes here n is n is word. Now for every n make in edge to its part of speech words, for POS words make an edge to its word sense. In this way connectivity is achieve. Now relate two words depend on their meaning. Single-value distribution [7] is used to discard the eigenvector [7] which is minimum. Overview: The random walk algorithm results in the efficient output for similarity.

Method 3: Method used for modeling the range of influence of the terms based on graph based approach:

Here the measure function is depending on terms. In this method the document has been represented in forms of terms. This term is represented using the series of node. A single node represents a sentence. The main goal of this method is calculate the influence of a particular term in a document. This influence matrix can give us efficient similarity and syntactic mean. Connection strength [7] between two nodes is the weight matrix for an edge. Point wise mutual information [7] is used to measure similarity. Relevance intervals are calculated to achieve the influence of every term. Many methods and prototype are there which are suing related intervals you can see in reference paper [7]. Overview: The unsupervised method base on graph uses weight to calculate influence of a term.

Method 4: Methods to calculate semantic similarity based on Graph Traversing:

Many methods are present for traversing the graph. Inthis method we are using models to traverse the graph whichwill use to find the synonym of the word. In this waysemantic similarity can be achieve. For traversing the, graphwalk models are used. Learned methods are used means thehistory between the two nodes has been saved which will usedin the future calculation.Overview: As history is used and we already have learning setthe search time is reduces in this method.

V. CONCLUSION

The conclusions to this survey show the comparison of the classification technique related to the similarity measures. These two methods depend on how the information is represented. After the completion of this paper you can compare the various methods of Graph Based Method for natural language processing.

REFERENCES

[1] Chris Biemann. Chinese whispers: an efficient graph clustering algorithm and its application to natural language processing problems. In Proceedings of the first workshop on graph based methods for naturallanguage processing, pages 73–80. Association for Computational Linguistics, 2006.

[2] AsliCelikyilmaz, Marcus Thint, and Zhiheng Huang. A graph-based semi-supervised learning for question-answering. In Proceedings ofthe Joint Conference of the 47th Annual Meeting of the ACL and the4th International Joint Conference on Natural Language Processingof the AFNLP: Volume 2-Volume 2, pages 719–727. Association for Computational Linguistics, 2009.

[3] WeiweiGuo and Mona T Diab. Improvements to monolingual English word sense disambiguation. In Proceedings of the Workshop onSemantic Evaluations: Recent Achievements and Future Directions, pages 64–69. Association for Computational Linguistics, 2009.

[4] Yi Guo and George Stylios. An intelligent summarization system based on cognitive psychology. Information Sciences, 174(1):1–36, 2005.

[5] ZornitsaKozareva. Cause-effect relation learning. In Workshop Proceedings of TextGraphs-7 on Graph-based Methods for Natural Language Processing, pages 39–43. Association for ComputationalLinguistics, 2012.

[6] Haibin Liu, RavikumarKomandur, and Karin Verspoor. From graphs to events: A subgraph matching approach for information extraction from biomedical text. In Proceedings of the BioNLP Shared Task 2011Workshop, pages 164–172. Association for Computational Linguistics, 2011.

[7] Michael T Mills and Nikolaos GBourbakis. Graph-based methods for natural language processing and understanding survey and analysis. Systems, Man, and Cybernetics: Systems, IEEE Transactions on, 44(1):59–71, 2014.

[8] Vivi Nastase and Stan Szpakowicz. Matching syntactic-semantic graphs for semantic relation assignment. In Proceedings of the First Workshopon Graph Based Methods for Natural Language Processing, pages 81–88. Association for Computational Linguistics, 2006.

[9] Delip Rao, David Yarowsky, and Chris Callison-Burch. Affinity measures based on the graph laplacian. In Proceedings of the 3rd Text graphsWorkshop on Graph-Based Algorithms for Natural Language Processing, pages 41–48. Association for Computational Linguistics, 2008.

[10] Philip Resnik et al. Semantic similarity in a taxonomy: An informationbasedmeasure and its application to problems of ambiguity in natural language. J. Artif. Intell. Res. (JAIR), 11:95–130, 1999.