ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 4, April 2015

An Approach for Keyword Searching in Uncertain Graph Data

Nikita B. Zambare[1], Snehalata S. Dongre[2]

M. Tech Student, Department of Computer Science and Engineering, GHRCE, Nagpur, India[1]

Assistant Professor, Department of Computer Science and Engineering, GHRCE, Nagpur, India[2]

ABSTRACT: Keyword searching is generally used for retrieving the relevant data from the database. For input query, the related data is retrieved. But it is tedious task to search keyword on uncertain graph. In this paper, the keyword searching technique over uncertain graph is introduced. The Keyword routing method is used to route the keywords to relevant source. In this approach two methods are included. The keyword relationship graph deduces the relationship between keywords and the element mentioning them. The scoring mechanism computes the score of keywords at each level which reduces the ambiguity. The result will include the subtree of the entire graph which includes all keywords of input query having high score and in addition it retrieves the most relevant data . Effective results are derived from employed method.

KEYWORDS:Keyword searching, Uncertain graph, algorithm, Keyword routing, graph data, Keyword query.

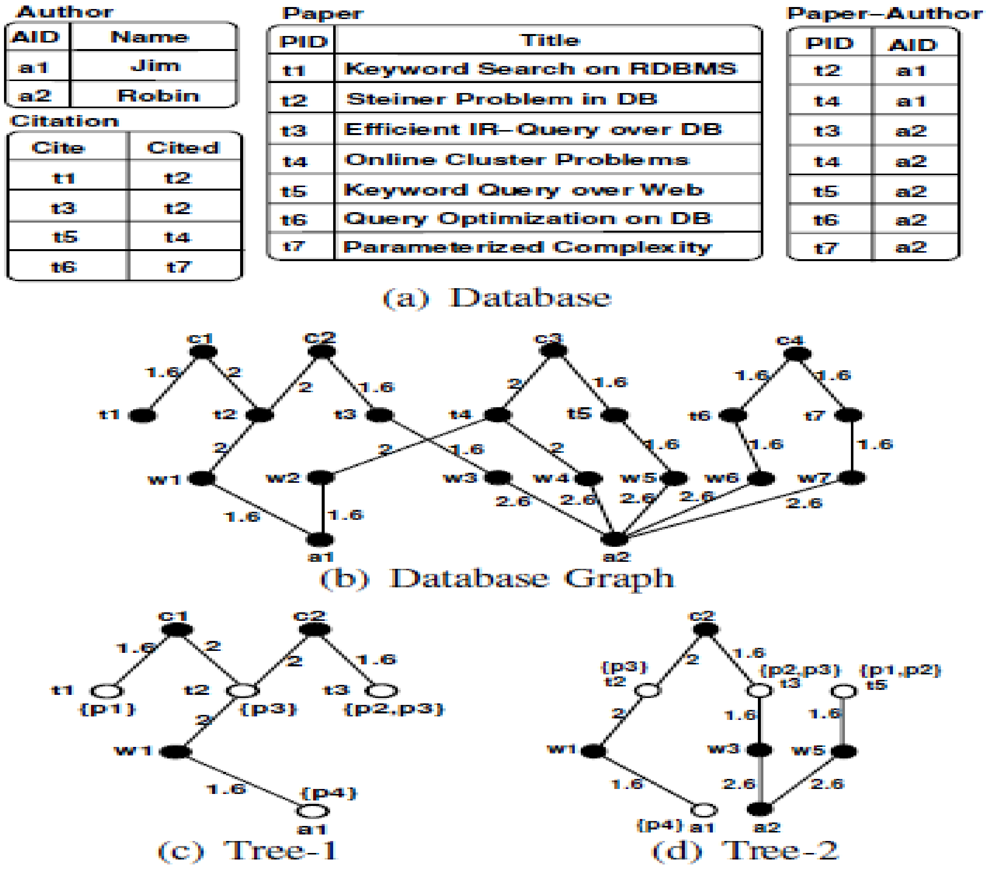
I. Introduction

Keyword search has been deduced to retrieve useful data from database, graph data. Keyword search has major advantage i.e. it is easy to operate. Users do not have to understand the query language and the database schema, and can gain the knowledge very quickly how to use information retrieval.Now a days, the study of keyword search technology based on Graph data has become a hot spot, and it is generally applied to the field of information retrieval. In the field of traditional graph database, the research on keyword searchhas already gained some achievement, but in the field ofuncertain graph data, the study on keyword search has barelystarted. Especially recently, quite a lot efforts have been put for keyword search over graphs, However, all graphs in the database are assumed to be certain or precise, and this assumption is often not valid in real-life applications. As RDF data and XML data can be highly unreliable due to errors in the web data or data expiration.

In the application of the data integration, it is needed to incorporate such RDF data from various data sources into an incorporated database. Uncertainties or inconsistencies often exist in this case. Like In social networks, each link between any two persons is often associated with a probability that represents the uncertainty of the link or the strength of influence a person has over another person in viral marketing. XML data having graph or tree form, uncertainties are integrated in XML documents known as probabilistic XML document (p-document).Keyword searching in RDF data, social networks and XML data have many important applications.

For data with relational and XML schema, specific query languages, such as SQL and XQuery, have been developed for information retrieval. In order to query such data, the user must master a complex query language and understand the underlying data schema. In relational databases, informationabout an object is often scattered in multiple tables due to normalization considerations, and in XML datasets, the schema are often complicated and embedded XML structures often create a lot of difficulty to express queries that are forced to traverse tree structures. Furthermore, many applications work on graph-structured data with no obvious, well-structured schema, so the option of information retrieval based on query languages is not applicable. Both relational databases and XML databases can be viewed as graphs. Specifically, XML datasets can be regarded as graphs when IDREF/ID links are taken into consideration, and a relational database can be regarded as a data graph that has tuples and keywords as nodes.

In the data graph, for example, two tuples are connected by an edge if they can be joined using a foreign key; a tuple and a keyword are connected if the tuple contains the keyword. Thus, traditional graph search algorithms, which extract features (e.g., paths [12], frequent-patterns [13], sequences [11]) from graph data, and convert queries into searches over feature spaces, can be used for such data.Therefore, it is necessary to relax the strict assumption of Deterministic or well certain graphs and study keyword search over uncertain graphs. Keyword Query Analysis is the ultimate goal of research on uncertain graph data management to retrieve the useful data from uncertain graph data.



Author

AID	Name
a1	Jim
a2	Robin

Citation

Cite	Cited
t1	t2
t3	t2
t5	t4
t6	t7

Paper

PID	Title
t1	Keyword Search on RDBMS
t2	Steiner Problem in DB
t3	Efficient IR–Query over DB
t4	Online Cluster Problems
t5	Keyword Query over Web
t6	Query Optimization on DB
t7	Parameterized Complexity

Paper–Author

PID	AID
t2	a1
t4	a1
t3	a2
t4	a2
t5	a2
t6	a2
t7	a2

(a) Database

(b) Database Graph

(c) Tree-1

(d) Tree-2

Figure 1. A Motivation Example

In Fig1.[4] the relational database is considered for keyword searching. The database includes author data which provides the information about author's id and his name. Next it includes paper data which provides its id and title. The database also includes the relation data between paper and author data which includes paper id and author id. In next, the relationship is represented among these data via graphical structure.Whatever the input keyword query is entered, the keywords are searched in graph and routes are found out to reach keywords and shows the routed subgraph in results.

II. Literature Review

The previous work on uncertain graph data introduced the various approaches which also provides the effective results but the employed methods having some drawbacks which will be overcome in our approach. Some keyword searching techniques and pruning techniques are reviewed.

Preprocessing Techniques:

1.Named Entity Recognition (NER)-

This task [14] parses each word in statement and grouping each of them according to predefined classes. e.g. Prof. Kulkarni taught NLP during February 2014. Here "Prof. Kulkarni" will come under predefined class "Person" and "February 2014" will come under predefined class "Date".

2.Word Sense Disambiguation (WSD)-

Sometime it happens in English language same word has different meaning, and according to sentence its meaning also change. So WSD [15] will recognize correct sense of term or word in particular text sentence. E.g. 1) The Sheep is in the pen. 2) The red ink is in the pen. Here word "pen" has two meaning first it is an enclosure where animals are kept. Second meaning is it is the pen uses for writing purpose.

3. Stemming-

Stemming task [16] derives a word to their root, base form, or stem. It correlated number of words by mapping them to the same stem. E.g. decided (Adjective), decision (Noun), decidedly (Adverb), all these words are different form of the same word. These are derived from the same word "decide". So decide is the stem word.

4. Part Of Speech (POS)-

It is the technique [17] of annotating term in a text corresponding to a particular part of speech, depend on both its interpretation and its circumstance—i.e. relationship with neighbor and associated words in a phrase, sequence of statements or paragraph.E.g. Raj saw the ship, Here, in example in first line sentence is given and in second line their respective POS tagging are also given. NNP denotes proper noun, VBD denotes verb, etc. POS tagging has large and major application in "preprocessing of text" that can be done by various methods and algorithm.

5. Chunking-

When there is need to remove unused words in the sentences, chunking is used. Chunking [18] will find out more specific and particular information.E.g. A Central jail in the City of Nagpur, Here, just next to chunking operation output will be shown as "Central jail Nagpur", it extracts only significant word from the statement. Chunking has major application in preprocessing of text to enhance the work of POS tagging

Keyword Searching Techniques:

Keyword searching is a challenging task for retrieving the useful and more relevant data over an uncertain graph data. Some techniques provide more efficient keyword searching.

Wangchao Le et al. [1] developed an effective summarization algorithm which summarizes the RDF data. RDF data are simulated as graphs. It can be highly unreliable. Search algorithm provides accurate results. This algorithm constructs a brief summary at the type level of RDF data. On query appraisal, the summary is effectively leveraged to prune the significant part of RDF data on the search space, and to elaborate SPARQL queries for efficiently accessing the graph data. As data get updated, the proposed summary can be updated.

George Kollios et al. [2] perform inclusive study on clustering the probabilistic graphs by using edit distance metric. The expected edit distance is minimized from the input probabilistic graph by the problem of finding the cluster graph. The optimum number of clusters are derived algorithmically. By establishing a connection with correlated clusters, the problem of finding cluster graph is efficiently estimated. A framework is established to compute deviations of a random world to the proposed clustering. But the output clusters are noisy.

In [7], Keyword search method is implemented on uncertain database which includes different tuple levels. The single table and multi-table uncertain data problems processed with keyword search method and results in optimized ranking function. Under the possible world semantics and the correlation with query keywords, the top-k query results having highest ranking scores are effectively evaluated. The proposed method is more effective and efficient.

Bolin Ding et al. [8] studied large directed/Undirected graphs, and confirmed that the optimal GST-1 can be achieved by proposed algorithm with high efficiency and achieve high efficiency and high quality for computing GST-k. The illustration is given in figure 1.

ZhaonianZou et al. [5] investigated the problem of mining uncertain graph data and focuses on mining frequent subgraph patterns on an uncertain graph database. By introducing a new measure known as expected support, the frequent subgraph pattern mining problem is validated. To find out the imprecise dominant subgraph patterns having relative error tolerance on expected supports, the approximate mining algorithm i.e. MUSE is proposed. MUSE has high efficiency, scalability, and accuracy, and the optimization techniques adopted by MUSE are efficient and so much effective.

Jun Gao et al. [4] introduced a FEM framework to bridge over the gap between graph operations and relational operations. To improve the performance of FEM framework, new feature of SQL standards viz. window function and merge statements are introduced in this paper. An edge weight aware graph partitioning schema and design a bi-directional restrictive BFS (breadth-first-search) over partitioned tables, are proposed which helps to improve the scalability and performance by avoiding extra indexing overheads. Likewise, keyword searching over uncertain graph data becomes easier.

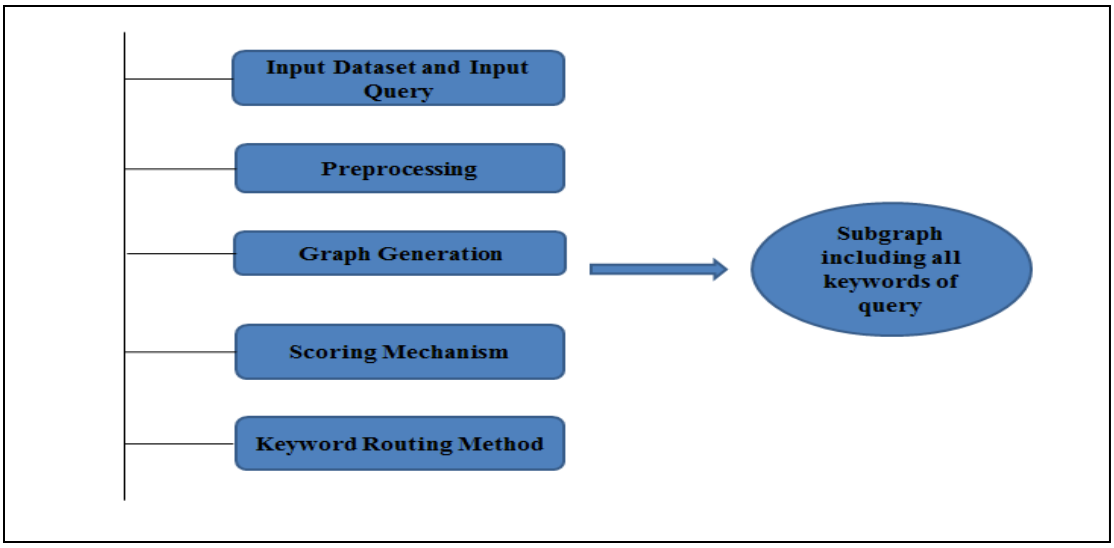
Thanh Tran and Lei Zhang [8], employs new keyword searching method i.e. keyword routing method which routes the keywords to the relevant and useful data sources, it reduces high cost of processing keyword search queries over various sources. Keyword element relationship summary and multilevel scoring mechanism are employed for routing plan. It greatly helps to improve the performance of keyword search without compromising the result quality.
The top-k subtrees

Ye Yuan et al [9] developed some pruning techniques which helped to minimize the complexity of keyword search over uncertain graph data. The filtering and verification strategy is adopted to speed up the search. In filtering phase, a probabilistic inverted index, PIndex which is based on subgraph features obtained by an optimal feature selection process are used. And finally in verification phase, the remaining candidates are validated using exact algorithm with tight bounds which also provide final results.

The work in [4] presented efficient keyword searching over uncertain graph data using filtering and verification methods. Where filtering includes three pruning phases existence, path-based and tree-based probabilistic pruning phases. And for verification, the sampling algorithm is used. In existence probabilistic pruning, all uncertain information is removed from the graph. In path based probabilistic pruning, probabilistic keyword index (PKIndex) stores all shortest path in graph. So, for each keyword w, PKIndex stores the top-k score probabilities of P that can reach w. In tree-based probabilistic pruning all subtrees are arranged in non-increasing order.

III. PROPOSED DESIGN

The main objective of our approach is to search keyword in uncertain graph data and in addition to retrieve the relevant data for input query.


Input Dataset and Input Query
Preprocessing
Graph Generation
Scoring Mechanism
Keyword Routing Method
Subgraph including all keywords of query

The above mentioned modules are used in our approach to search keywords in an uncertain graph data and routes to reach the query keywords and finally showssubtree in result which includes all keywords entered by users and in addition it shows most relevant data related to query keywords.

Phase I:

1.Input dataset- Initially the input dataset browsed by user, the text data file is selected as an input dataset. In our approach the twitter dataset is used. The twitter dataset includes tweets from different ids.

2. Preprocessing- As the input dataset is used, the complete text file is preprocessed. Here, the POS tagging and Chunking technique is used for preprocessing the data. POS tagging and chunking tasks are so important in document text preprocessing. It will be better to use both POS tagging and chunking for text preprocessing. It will show optimized result if Chunking will used after POS tagging operation on text. POS focused on assigning each word with unique tag which represents syntactic role. It does it by using features like multiple words bigrams, trigrams, etc. with preceding and following tag context, and manual features to handle unknown words. In Chunking, this is also called as a shallow parsing to focus on assigning sentence segment with syntactic role such as verb or noun phrases. Chunking assigns only one unique tag, often called as a begin-chunk or inside chunk tag. For chunking and POS tagging [19], there is a need to assign tag to each word in a sentence. consider the whole sentence for tagging each word in the sentence by producing local features for each word of the sentence and integrates these features into a global feature vector using Neural Network which can then be send to standard affine layers.
Enter text for Preprocessing: ***The fisherman went to the bank*.**
Preprocessed Text after chunking will be: ***fisherman went to bank***
Here by using chunking, important and relevant word are extracted as efficient search keyword through which we can get exact and desired document in search result.

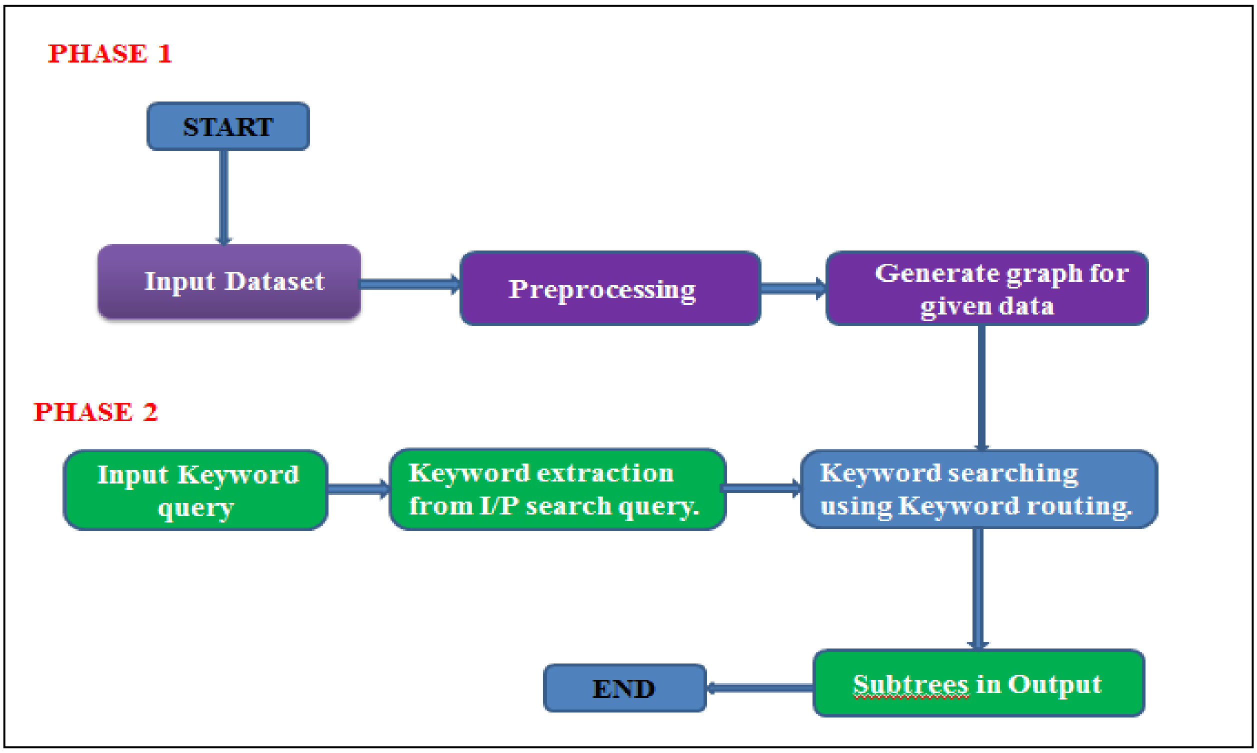

PHASE 1
START
Input Dataset
Preprocessing
Generate graph for given data
PHASE 2
Input Keyword query
Keyword extraction from I/P search query.
Keyword searching using Keyword routing.
Subtrees in Output
END

Fig.3. Proposed Approach

3.Graph Generation- In our approach we generate the level-wise tree structured graph. By using score based graph generation method, the frequency of each word is manipulated according to its occurrences, those keywords having highest frequency will be allotted at highest level. The keyword at the highest level has highest score. Thus a complete tree structured level-wise graph is generated for entire dataset.

Phase II:

4. The input query including keyword is given by a user. If the input query is sentence then this sentence is initially preprocessed and keywords are fetched for efficient searching.

5.Routing Method-

Routing method routes the keyword to highly relevant data sources within some instant of time. While in keyword searching on all sources, it reduces the high cost required for query processing. Firstly in this method, the selected sources are preprocessed (pruned) then the keyword graph is generated for more relevant sources. According to the routing plan, the query including keywords is processed and delivers only the most relevant and matching information needed. As the keyword searching using other approaches is problematic when the number of keywords is large in a query. But routing method can be used for large keywords in a query because if the information need is well described then only more relevant data can be retrieved.

In our approach as per the input query keywords, the algorithm scan the entire graph from root node to leaf nodes till reaching to the all keyword. It maintains an index to store all the routes reaching to the keywords and finally shows the subtree in output result.

6.Scoring Mechanism-In addition we are showing the relevant data in result to input keyword query. For example we consider the twitter dataset which includes tweets. These keywords are searched in twitter dataset. Those tweets having these keywords, only that tweets will be shown in results. But for effective results they are ranked by scoring them for each tweet. By calculating the score of keywords for every tweet, that score is again manipulated by comparing it with the graph levels. If we do not set the score with respect to the graph, then we will get normal re-ranking without proper score. So with proper score, relevant tweets are re-ranked and efficient results are generated.

IV. CONCLUSION

Keyword search provides a simple but user-friendly interface to retrieve information from complicated data structures. Since many real life datasets are represented by trees and graphs, keyword search has become an attractive mechanismfor data of a variety of types. Because of the underlying graph structure, keyword search over graph data is much more complex than keyword search over documents. So proposed work is about searching keyword in an uncertain graph data with preprocessed keyword query.The keywords are searched on graph and generate the subtree which includes all keywords.

REFERENCES

[1] Wangchao Le, Feifei Li, AnastasiosKementsietsidis, SongyunDuan,"Scalable Keyword Search on Large RDF Data",IEEE2013.

[2] George Kollios, MichalisPotamias, and EvimariaTerzi," Clustering Large Probabilistic Graphs", IEEE vol. 25, NO. 2, February 2013

[3] Ye Yuan, Guoren Wang, Lei Chen, and HaixunWang,"Efficient Keyword Search on Uncertain Graph Data",IEEE vol. 25, no. 12, December 2013.

[4] Jun Gao, Jiashuai Zhou, Jeffrey Xu Yu, and TengjiaoWang,"Shortest Path Computing in Relational DBMSs", IEEE vol. 26, no. 4, April 2014.

[5] ZhaonianZou, Jianzhong Li, Member, IEEE, Hong Gao, and ShuoZhang,"Mining Frequent Subgraph Patternsfrom Uncertain Graph Data", IEEE vol. 22, no. 9, September 2010.

[6] LifangQiao, Yu Wang,"A Keyword Query Method for Uncertain Database", 2nd International Conference on Computer Science and Network Technology, IEEE, 2012.

[7] Bolin Ding, Jeffrey Xu Yu, Shan Wang, Lu Qin, Xiao Zhang, Xuemin Lin," Finding Top-k Min-Cost Connected Trees in Databases", IEEE 1-4244-0803-2/07/2007.

[8] Thanh Tran and Lei Zhang, " Keyword Query Routing", IEEE vol. 26, no. 2, February2014.

[9] Ye Yuan, Guoren Wang, Haixun Wang, Lei Chen," Efficient Subgraph Search over Large Uncertain Graphs". In Proceedings of the VLDB Endowment, Vol. 4,pp. 876-886, 2011.

[10] Hao He, Haixun Wang, Jun Yang, Philip S. Yu," BLINKS: Ranked Keyword Searches on Graphs", SIGMOD'07, June 2007.

[11]Haoliang Jiang, HaixunWang, Philip S. Yu, and Shuigeng Zhou GString: A novel approach for efficient search in graph databases. In *ICDE*, 2007.

[12] DennisShasha, Jason T.L.Wang, and RosalbaGiugno. Algorithmics and applications of tree and graph searching. In *PODS*, pages 39–52, 2002.

[13] Xifeng Yan, Philip S. Yu, and Jiawei Han. Substructure similarity search in graph databases. In *SIGMOD*, pages 766–777, 2005.

[14] Branimir T. Todorovic, Svetozar R. Rancic, Ivica M. Markovic, Eden H. Mulalic, Velimir M. Ilic, "Named Entity Recognition and Classification using Context Hidden Markov Model," 9th Symposium on Neural Network Application in Electrical Engineering, NEUREL, pp. 43-46, 2008.

[15] Dekai Wu, Weifeng Su and Marine Carpuat, "A Kernel PCA Method for Superior Word Sense Disambiguation," Proceedings of the 42nd Meeting of the Association for Computational Linguistics, pp. 637-644, 2004.

[16]AbdelazizZitouni, AsmaDamankesh, ForooghBarakati, Maha Atari, Mohamed Watfa, FarhadOroumchian, "Corpus-based Arabic Stemming Using N-grams," Asia Information Retrieval Symposium - AIRS, vol. 6458, pp. 280-289, 2010.

[17] Hassan Mohamed, Nazlia Omar, MohdJuzaidinAb Aziz, "Statistical Malay Part-of-Speech (POS) Tagger using Hidden Markov Approach," International Conference on Semantic Technology and Information Retrieval, pp. 231-236, June 2011.

[18]Saike He, Taozheng Zhang, XueBai, Xiaojie Wang ,Yuan Dong, "Incorporating Multi-task Learning in Conditional Random Fields for Chunking in Semantic Role Labeling," International Conference on Natural Language Processing and Knowledge Engineering, pp. 1-5, 2009.

[19] Rahul Dudhabaware,Mangala M. Madankar, "Review on Natural Language Processing Tasks for Text Documents", IEEE International Conference on Computational Intelligence and Computing Research, December 2014.