UGC Approved Journal

IARJSET

ISSN (Online) 2393-8021
ISSN (Print) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

ISO 3297:2007 Certified

Vol. 4, Issue 7, July 2017

Labeling Document Clusters with Thematic Phrases

Dr. Y. Sri Lalitha[1], Dr. N. V. Ganapathi Raju[2], Dr. O. Srinivasa Rao[3]

Professor, GRIET, Hyderabad, India[1]

Professor, GRIET, Hyderabad, India[2]

Associate Professor, Dept of CSE, JNTUK, Kakinada, India[3]

Abstract: Document clustering is a powerful technique to detect topics and their relations for information browsing, analysis, and organization. However, clustered documents require post-assignment of descriptive titles to help users interpret the results. Existing techniques often assign labels to clusters based only on the terms that the clustered documents contain, which may not be sufficient for some applications more over term labeling will not give clear meaning of the clustered contents. To solve the problem, a phrase based cluster labeling is considered in this work. The work considers embedding external knowledge to terms using WordNet and provides an approach to derive a theme in the group of documents and label that group with the most appropriate Phrase. Number of experiments conducted on benchmark datasets and observed that results produced are very accurate to the clusters formed.

Keywords: Thematic Phrases, Document clustering, Information Browsing, Analysis, and Organization.

1. INTRODUCTION

Sophisticated technologies and storage devices, allowed storing of huge volumes of data, but, the ability to understand and utilize this information, remains constant. The task of organizing and categorizing these documents to the diverse need of the user by manual means is a complicated job; hence a machine learning technique named clustering is very useful. Document Clustering is an approach to organize this huge collection of documents into partitions of related documents. This facilitates location of a relevant document. Document Clustering Phases involves Collection of Document Datasets, Preprocessing the documents for removal of noisy and irrelevant data, Apply clustering technique and evaluate the cluster quality. After forming the Clusters can we exactly say what kind of information a cluster contains? The least emphasized step of Clustering Process is Labeling the Cluster.

Cluster Labeling

Rapid growth of the World Wide Web has caused an explosion of research aimed at facilitating retrieval, browsing and organization of on−line text documents. Much of this work was directed towards clustering documents into meaningful groups. In many applications of flat clustering and hierarchical clustering, particularly in analysis tasks and in user interfaces, human users interact with clusters. The abundance of information in text form has been both a blessing and a plague. There is always a need to summarize information into compact form that could be easily absorbed. The challenge is to extract the essence of text documents and present it in a compact form that identifies their topic(s). In such Application, Cluster Labelling with most relevant description of Cluster Contents is necessary. Often, given a set or hierarchy of document clusters, a user would prefer to quickly browse through the collection to identify clusters of interest without examining particular documents in detail.

This paper is organized as follows. The next section deals with Related Work. Section 3 presents our approach to label the clusters, section 4 explains experiment setup, results and analysis and section 5 deals with Conclusions and future work.

2. RELATED WORK

Clusters are often labeled by Prominentwords in the cluster after removal of stop words [1−3]. The lists of the most frequent words often reveals the topic at a high level, but can fail to depict cluster specific details as they are diluted with what we call collection specific stop words. E.g., in a collection of computer science research papers, terms such as paper, method, result, system, or present are very frequent and are common to most computer science sub-disciplines, therefore giving no additional information to someone who already knows that all of the documents are computer science research papers. One could use the words that are the most predictive of a given cluster. These are equivalent to what have been called the most salient words in word sense disambiguation [4]. However, relatively

DOI10.17148/IARJSET.2017.4703

infrequent words usually have high salience and are not suggestive of any topic. They are often simply words that are misspelled, if low frequency words are not removed. There exists a promising body of work [9-12] aimed at learning or organizing hierarchical topics of words or document collections, which assumes that words in a document follow distributional patterns and can be ascribed to different word−generating components ranging from the general to specific. The above research provides ways of labeling clusters which reduce but do not eliminate the problems mentioned above. Furthermore, such methods can be difficult to implement. If labeling is all that is needed, one would prefer to use simpler methods to label clusters. Clusters can also be labeled using titles of the papers most central to a cluster [1], or most cited, but these labels can be quite idiosyncratic and also fail to provide insights into the quality of clusters under consideration. Following description reveals two word based Labeling techniques.

2.1 Word Based Labeling

2.1.1 X^2 Method

This method assumes the existence of a document hierarchy, either manually constructed and/or populated, or a hierarchy resulting from application of a hierarchical clustering algorithm. The X^2 test is well suited for testing dependencies when count data is available. The method use X^2 tests for each word at each node in a hierarchy starting at the root and recursively moving down the hierarchy to determine a set of words that are equally likely to occur in any of the children of a current node. Such words are general for all of the sub-trees of a current node, and are excluded from the nodes below. Label the node corresponding to a cluster with the prominent word at the node [1,3].

2.1.2. Frequent and Predictive Words Method

For the "frequent and predictive words" method, words are selected as labeling based on the product of local frequency and predictiveness:

$$P(\mathit{word}/\mathit{Class}) \times \frac{P(\mathit{word}/\mathit{Class})}{P(\mathit{word})}$$

This combined use of local frequency and predictivenesswas used to select the most important words in categories for illustrating word sense disambiguation[15]. The formula consists of two parts each having a well defined meaning: the first term, predictiveness, **p(word | class) / p(word)** is similar to a mutual information estimator and TF−IDF measure used in information retrieval in that it distributes more weight to the words occurring frequently in a given cluster and less weight to the words occurring frequently in all of the clusters; **p(word | class)** is frequency of the word in a given cluster and **p(word)** is the word's frequency in a more general category or in the whole collection.

Labeling the Clusters with word reveal less information as compared to phrase in most of the cases. For example consider "Information" is the prominent word in the cluster, and labeling the cluster with word "Information" will not clearly conclude which information it is indicating. Similarly "Apple" is the prominent word; we can't conclude either it as fruit or Apple as product. Rather if we take phrase into consideration, it will reveal more information as compared to single word.We will elaborate the above two examples. In first example, consider "Information" as the cluster label.The documents in such cluster contains word "Information" either preceded or succeeded by other word such as "Information Retrieval, Information Extraction, Information Systems" or "History Information, Mathematics Information or Geography Information". Labeling the clusters with such phrases will give more Information than single word. This work deals to label the clusters not only with Key Phrases, but meaningful Key Phrases by incorporating the Background Knowledge.

2.2Key Phrase Extraction Techniques

A **phrase** often refers to any group of words. In linguistics, a phrase is a group of words (or sometimes a single word) that forms a constituent and so function as a single unit in the syntax of a sentence. A phrase is lower on the grammatical hierarchy than a clause. Most phrases have an important word defining the type and linguistic features of the phrase. This word is the head of the phrase and gives its name to the phrase category. The heads in the following phrases are in bold: too **slowly** - Adverb phrase (AdvP), very **happy** - Adjective phrase (AP), the massive **dinosaur** - Noun phrase (NP), **at** lunch - Preposition phrase (PP), **watch** TV - Verb phrase (VP)

Key phrase extraction, which is a text mining task, extracts highly relevant phrases from documents. Key Phrase Extraction from unstructured text documents is becoming important. Literature lists over a dozen applications that utilize key phrase extraction. For example, providing mini-summaries of large documents, highlighting Key phrases in text, text compression, constructing human-readable key phrase-based indexes, interactive query refinement by suggesting improvements to queries, document clustering, and document classification are few usecases. Works on automatic Keyphrases extraction started rather recently. First work to approach this goal is based on heuristics [16]. The work use heuristics to extract significant phrases from document for learning rather than use standard mathematical techniques. However Keyphrases generated by this approach, failed to map well to author assigned keywords indicating

DOI10.17148/IARJSET.2017.4703

that the applied heuristics were poor ones [16]. Key phrase extraction techniques are often categorized as supervised and unsupervised learning. Key phrase Extraction from single document, which is often called as supervised learning task and Key phrase extraction from a set of documents is called as unsupervised learning task. Unsupervised task tries to discover the topics rather than learn from examples.

2.2.1 Supervised Learning

The Work in [13] was the first approach to address the problem of Keyphrases extraction as a supervised learning problem which is called GenEx. It regards Keyphrases extraction as classification task. The GenEx is a Hybrid Genetic Algorithm for Keyphrases extraction which has two components, the Genitor genetic algorithm proposed in [17] and the Extractor which is a Keyphrases extraction algorithm proposed in [13][19]. The Extractor takes a document as input and produces a list of Keyphrases as output. The Extractor has twelve parameters that determine how it processes the input text. In GenEx, the parameters of the Extractor are tuned by the Genitor genetic algorithm to maximize performance on training data. Genitor is used to tune the Extractor, but Genitor is no longer needed once the training process is complete. Basing on the work of [13][19], another algorithm proposed in [20] called Keyphrases extraction algorithm (KEA). This latter uses Naïve Bayes learning model using a set of training documents with known Keyphrases. Then, the model will be used to determine which sentences of an original document are likely to be Keyphrases.

2.2.2 Unsupervised Learning

A graph-based ranking method[21], constructs a word graph according to word co-occurrences within the document, and then use random walk techniques to measure word importance. After that, top ranked words are selected as Keyphrases. [14] presents a highly accurate method for extracting key phrases from multi-document sets or clusters, with no prior knowledge about the documents. The algorithm is called CorePhrase, and is based on finding a set of core phrases from a document cluster. CorePhrase works by extracting a list of candidate key phrases by intersecting documents using a graph-based model of the phrases in the documents. This is facilitated through a powerful phrase-based document indexing model. Features of the extracted candidate key phrases are then calculated, and phrases are ranked based on their features. The top phrases are output as the descriptive topic of the document cluster. The KP-Miner system proposed in [22]for Arabic language contains three logical steps: Candidate Keyphrases selection, Candidate Keyphrases weight calculation and finally Candidate Phrases List Refinement extracts Arabic Key Phrases[18]. Another work is Suffix Tree based Phrase extraction. A Suffix Tree is a data structure that allows efficient string matching and querying. It has been studied and used extensively, to fundamental string problems such as finding the longest repeated substring, strings comparisons, and text compression[23]. It often deals with strings as sequences of characters, or with documents as sequences of words. A Suffix Tree of a string is simply a compact tree of all the suffixes of that string. The work in [24] is the first work used fortext documents. It's linear time clustering algorithm that is based on identifying the shared phrases that are common to some Document's in order to group them in one cluster. A phrase in our context is an ordered sequence of one or more words. To our knowledge Phrase based cluster Labeling has very few works. In this study we present an approach to Label Clusters using Suffix Tree and enrich the words with external knowledge usingWordNet.

WordNet is a database of semantic dictionary for the English language that provides the sense information of words [25]. Different from traditional dictionaries, WordNet is organized as a semantic network. The whole corpus consists of four lexical databases, for nouns, verbs, adjectives and adverbs [26]. The basic unit in each database is a set of synonyms called a synset. A synset represents a meaning and all words that have such a meaning will be included in this synset. If a word occurs in several synsets, then it is polysemous. Each synset is assigned a definition or gloss to explain the meaning of it. There are various relationships defined between two synsets. For example, the relationship hypernym indicates one synset is a kind of another synset (IS-A relationship), the hyponym relationship is the inverse of hypernymy. In the semantic network, each synset is represented as a node and the relationships among the synsets are represented as arcs. WordNet allows the meaningfully related words and concepts to be navigated using a browser, making it useful for natural language processing. It provides a user friendly approach in representing and interpreting the sense of text and provides well-organized and integrated access to information. WordNet's recognition and analysis of semantic equivalents are used by many researchers in the context of Information Retrieval, Text Classification and Text CategorizationDocument features, enriched with different relationships of WordNet has high possibility of deriving better results when incorporated in document Categorization, Clustering or Retrieval process.

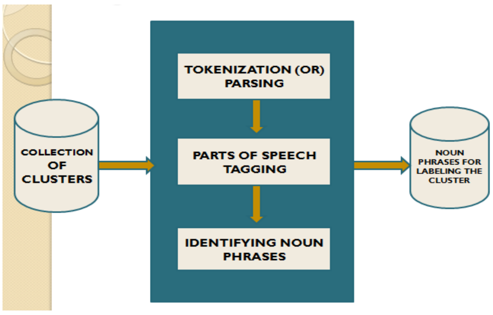
A set of different terms that having the same meaning (Synonyms) is termed as a concept. In general, frequent terms are used to cluster documents, where as in FCDC (Frequent Concepts based Document Clustering), frequent concepts are employed. In this approach frequent concepts are identified by applying Apriori algorithm on feature vector that constitutes concepts [28]. The approach uses WordNet hierarchy to determine semantic relations between words and allows creating low dimension feature vectors [29]. It is observed that often text clustering approaches employ VSM, which treats documents as BOW, thus, ignoring the word sequence of a document. But, word sequences of a document convey more meaning than considering terms alone. The Clustering with Frequent Word Meaning Sequences (CFWMS) [27] generates a Suffix Tree to determine the FWMS from the document collection. This work explored WordNet synonym/hypernym relations to build meaning sequences and applied Suffix Tree Clustering.

3. OUR APPROACH

In Labeling the clusters, we prefer to use Noun Phrases over other phrases. Since the noun phrases are more frequent in English Language and are easily understandable than other Phrases. This section provides the approach to extract the noun phrases, incorporate the semantic knowledge and identifying the accurate phrase to label the cluster.

3.1 Noun Phrase Extraction Technique

Noun phrases are extracted from the preprocessed documents datasets.It is composed of the following steps :Tokenizing the documents, Tagging each token with Parts of Speech, identify Noun Phrases with POS tagging. In Natural Language Processing Noun Phrases are contained in the combinations of "Noun + Noun", "Adjective + Noun", "Noun + ing "and so on. Extraction of Noun Phrases and Identifying frequent Noun Phrases and Labeling the Cluster.



The following is a list of different types of Noun Phrases.

- "Noun followed by noun", i.e., Noun + noun combination. Ex. "The car driver" or "the driver of the car", "Data abstraction", "History information" and "Power supply".
- "Adjective followed by noun" i.e.,adjective + noun combination. Ex. "ambiguous statement", "slow motion picture", "statistical data", "higher education", "internal representation", "interactive session".
- "Adverb followed by noun", i.e., Adverb + noun. Ex. "terminally ill", "technically strong".
- "Adverb followed by adjective followed by noun", Adverb+ Adjective + noun. Ex. "inaccurately entered data", "statistically derived information", "economically backward class","newly published book", "scientifically proven theory".
- "Noun followed by gerund", i.e., noun + gerund (ing). Ex. "Data mining", "method overriding", "horse riding", "sky diving" and "reverse engineering".
- "gerund followed by noun", i.e., gerund (ing)+ noun. Ex. "Typing skils", "Marketing strategies", "Boiling point", "Freezing point", "Waiting area", "Incoming call" and "Outgoing batch"
- "Noun followed by gerund followed by noun", i.e., noun + gerund (ing) + noun. Ex. "Software testing methodologies", "Method overloading concepts", "Object-oriented programming language", "Weight lifting equipment".
- Noun followed by series of nouns, i.e., noun + noun + … . Ex. "Java programming language specification for 7th edition", "Important Relational Database management systems concepts"
- Pronoun followed by noun, i.e., Pronoun + noun, Ex. "That girl", "This definition", "These concepts".
- Noun followed by Pronoun, i.e., Noun + Pronoun. Ex. "A friend of mine", "The idea of yours".

3.2 Suffix Tree Based Algorithm

A suffix tree of a string is simply a compact trie of all the suffixes of that string[26]. Here we treat documents as sequences of words, not characters. A suffix tree T for an m-word string S is a rooted directed tree with exactly m leaves numbered 1 to m. Each internal node, other than the root, has at least two children and each edge is labeled with a nonempty sub-string of words of S. The label of a node is defined to be the concatenation of the edge-labels on the path from the root to that node. No two edges out of a node can have edge labels beginning with the same word. For each suffix s of S, there exists a suffix node whose label equals S. The suffix trees are fast, incremental and are constructed in linear time of the suffixes generated. This work considers, suffix tree data structure to represent the word sequences from documents. The suffix tree that contains all the documents in the data set is called a Generalized Suffix Tree (GST). The GST is explored to extract unique suffix phrases.

Construction of suffix trees for documents :A text document D is viewed as a sequence of words, so that it can be represented as D = (w1, w2, w3 . . .), where w1, w2,w3, . . . are words appearing in D. Like a frequent itemset in the association rule mining of a transaction data set [24], a word set is frequent when at least the specified minimum number (or percentage) of documents contains this word set. A frequent word set containing k words is called frequent k-word set. A frequent k-word sequence is an FS with length k, such as FS = (w1, w2. . . wk), and it has two frequent subsequences of length k - 1, which are (w1, w2, . ., wk-1) and (w2,w3, . . ,wk). [31].The suffix of each line of the document is generated. After finding all the suffixes of all the documents we extracted unique suffixes and built compact suffix tree with these unique suffixes in a cluster. For each suffix, a search is done to see whether a part or whole of it already exists in the suffix tree. If this is true, only the TF and IDF counts for the respective node are updated. If only a partial match of the suffix is found, then it asks for mismatch option and based on user input it allows for words mis-match. The k value is determined automatically from the length of the Phrase. In the end, the suffix tree node will contain all the possible suffixes of every sentence in the document collection, together with the term and document frequency count for each node. If Filtering by frequent word sequences is enabled, list of frequent words (i.e., words that appear in frequent word sequences) is compiled prior to POS tagging.

3.3. Incorporating Semantic Knowledge

We use different terms to express the same meaning called synonyms. A term meaning in WordNet is represented by a synonym set (SynSet) a set of terms which are synonyms. In this paper, a SynSet is denoted by SyS, e.g. SyS1 = {chair, seat}, ‘‘seat’’ is a synonym of ‘‘Chair’’, so they are interchangeable in documents. When the term matching is performed to find the frequency of a term in a text database, the frequency of terms ‘‘seat’’ and ‘‘chair’’ can be summed up if we treat these two terms as same in a document. If it appears in different documents, then these documents may be placed in same cluster. Synsets are interconnected with semantic relations forming a large semantic network such as hyponym/hypernym which defines isArelationship between concepts also called as subset/superset relationship. For ex., the hypernym of a synset {armchair, folding chair} is {Chair}, and the hypernym of {Chair} is {Furniture}. Documents containing ‘‘arm chair’’ may share the same topic with other documents containing ‘‘folding chair’’ or ‘‘Chair’’. With simple term matching, we may lose the relatedness in these terms. If we use ‘‘→’’ to represent this relationship, then these three synsets could be related as {arm chair, folding chair}‘‘→’’{Chair} ‘‘→’’{Furniture}. In this case, {chair} is a direct hypernym of {arm chair, folding chair} and {furniture} is an inherited hypernym of {arm chair, folding chair}. Such binding of a synset with its hypernym in this paper is called as Synset Hypernym Component denoted by SHC. The SHC and SyS relation is denoted as SySi Є SHCj. In this work we propose text clustering approach with SyS and SHC relation, as this is the main relation group of WordNet hierarchy [27].

Since more than one synonym exists for terms, identifying the right synonym that a term expresses in a certain context is a difficult task. Here, we use a Synonym Group (SG), which is a collection of SHC, to find the correct meaning of a term. The relationship between SHC and SG is denoted as SHCj Є SGh. For a SySi, if SySi Є SHCj and SHCj Є SGh, then it is denoted as SySi Є SGh. For example there is a SG, SG1 = {SHC1,SHC2}, where SHC1 = {SyS1→SyS2}, SHC2 = {SyS3 →SyS4}, SyS1 = {arm chair}, SyS2 = {chair}, SyS3 = {folding chair, lounge }, and SyS4 = {furniture}. We expect that one of the SyS belonging to SG1 is the real term meaning of the term ‘‘chair’’ in this document. In our approach unrecognized terms in WordNet are taken as it is into the unique terms list as these terms may capture important information and is helpful in clustering[29].

In WordNet multiple synonyms of a given term are ordered from the most to the least frequently used. For each SyS selected, atleast two SyS with one direct hypernym synset is retrieved. In this way, each term has its SG containing at least one SHC. For example, a document d1 = <t1,t2,t3> can be converted to a sequence of SG as d1‘ = <SG1,SG2, SG3>, where SG1 = {SyS1→SyS2, SyS3→SyS4}, SG2 = {SyS5→ SyS8, SyS6 → SyS7}, SG3 = {SyS9 →SyS5}[25].

3.4 Our Methodology

Steps involved in Cluster labeling

1. Build a Suffix Tree of words.
2. Identify Frequent Key Phrases
3. POS Tag the words of Phrases.
4. Identify the valid Noun Phrases.
5. Apply Semantic Knowledge
6. Determine Frequent Meaningful Key Phrases.
7. Label the Cluster.

3.3 WordNet

WordNet isa lexical database for English language. It groups English words into sets of synonyms called synsets, provides short, general definitions, and records the various semantic relations between these synonym sets.

DOI10.17148/IARJSET.2017.4703

We can increase the efficiency and effectiveness of the System by integrating it with WordNet. The Noun Phrase Extractor is efficient in and of itself; however, we need the help of WordNet in order to deal with the complex and complicated issues of noun phrases which may not seem to be following the algorithm implemented for the system.
WordNet offers help on synonymy and hypernyms along with information on the appropriateness of a noun phrases in terms of context. Taking advantage of all this, we can make the NPE more efficient and the accuracy level can be more when synonym and hypernym related information is taken into consideration.
For example:

"Information Retrieval"

"Information extraction"

These two sentences are Noun Phrase with different words. When we consider manually, these two sentences give the same meaning. So we need to group these two sentences if we applied WordNet. Then we can enhance the meaning of the label of the cluster.

4. RESULTS

The datasets considered in this work are Classic, 500AB and DT. Classic Dataset is a bench mark dataset in the area of Text Mining. It is a collection containing Scientific and Medical works on Computer Hardware and Software. 500AB and DT datasets are the collection of abstracts from Internet. The datasets have 1000, 500 and 200 documents respectively.The work here applies two clustering techniques to partition the datasets. KMeans and Suffix Tree Clustering. On the results of these clustering we have applied Noun Phrase Based and Suffix Tree Phrase Based Labelling approaches. In these approaches to incorporate external knowledge we used WordNet and derived Semantic Groups.

Noun Phrase Based Labeling

DATASET		K-Mismatch	Semantic Groups + K-Mismatch
Classic	CAC	Computer Science Scientific Research	Computer Science Research
	CIS	Information Science Computing	Technical information
DT	DT	Based methods	Data mining
	IP	Dimensional feature space	Image processing
	NS	Tolerant authentication broadcast authentication protocol	Ad hoc networks
	ST	Test plans improvement using simulated defect removal	Software testing
500 AB	DM	JDBC driver implementation	Spatial database
	IP	Digital image Processing Component.	Image processing
	NS	Flow system	Ad hoc network
	ST	Rich base stations	Structural testing

Suffix Tree Based Labeling

DATASET		Without Semantic Groups	Semantic Groups
Classic	CAC	Batch Processing	Data processing
	CIS	Annual Review Information Science Technology	Information Science Technology
DT	DM	Data Mining Research	Data Mining Techniques
	IP	Digital Image Processing	Image Processing
	NS	Denial service attacks	Denial service attacks
	ST	Software testing	Software Testing/Software Maintenance costs
500 AB	DM	Large spatial database	Spatial Data mining
	IM	Digital image processing	Image processing
	NS	Network routing protocols	Sensor networks
	ST	Instant world Wide Web enabled application website quality reliability	World Wide Web Testing Software

5. CONCLUSIONS

This work experimented with Noun Phrase Labelling and Suffix Tree Phrase Labelling. It is observed that Noun Phrase Labelling with K-Mismatch has shown good results in certain datasets and is quite fast to extract Labels but in certain cases it has derived lengthy Labels. The Semantic Groups approach specified in this work has shown compact and

DOI10.17148/IARJSET.2017.4703

meaningful Labels. We have derived Labels with Suffix Tree Phrases applied WordNet to derive Semantic Groups. Semantic Group based Labelling is giving meaningful labels. It is observed that Noun Phrase Extraction requires lot of time for preprocessing the data, whereas in Suffix Tree approach we giveweight to each phrase and consider the phrase with good weight. In the process there is a chance of losing some Information. In our future work we would like to explore the semantic information methods

REFERENCES

[1] Banko, Michele, Mittal, Vibhu O., &Witbrock, Michael J.Headline generation based on statistical translation. ACL 2000.

[2] Yuen-Hsien Tseng, Generic title labeling for clustered documents, elsevier.com/locate/eswa.

[3] Brill, E, Marcus, M. Tagging an Unfamiliar Text with Minimal Human Supervision. Intelligent Probabilistic Approaches to Natural Language. Fall 1992 Symposium.

[4] Church, Kenneth (1988), "A Stochastic Parts Program and Noun Phrase Parser for Unrestricted Text",Proceedings of Second Conference on Applied Natural Language Processing, pp. 136-143.

[5] Cutting, D. R., Karger, D. R., Pedersen, J. O., and Tukey, J. W., Scatter/Gather: aCluster−Based Approach to Browsing Large Document Collections. In Proceedingsof the 15th International ACM SIGIR Conference on Research and Development inInformation Retrieval, pages 318−29 (1992).

[6] Baker, L. D., and McCallum, A. K., Distributional Clustering of Words for TextClassification, Proceedings of the 21st Annual International ACM SIGIR Conferenceon Research and Development in Information Retrieval , pp. 96 − 103 (1998).

[7] Sahami, M., Hearst, M., and Saund, E., Applying the Multiple Cause Mixture Modelto Text Categorization, In ICML−96: Proceedings of the Thirteenth InternationalConference on Machine Learning, pp. 435−443, San Francisco, CA: MorganKaufmann (1996).

[8] Yarowsky, D.,"Word−Sense Disambiguation Using Statistical Models of Roget'sCategories Trained on Large Corpora ,"In Proceedings, COLING−92. Nantes, pp.454−460 (1992).

[9] Pereira, F., Tishby, N., and Lee, L., Distributional Clustering of EnglishWords,"30th Annual Meeting of the Association for Computational Linguistics, pp.183−190 (1993).

[10] McCallum, A., Rosenfeld, R., Mitchell, T., and Ng, A.Y, Improving TextClassification by Shrinkage in a Hierarchy of Classes, In Proceedings: the FifteenthInternational Conference on Machine Learning, Morgan Kaufmann (1998).

[11] Hofmann, Th., Learning and Representing Topic: A hierarchical Mixture Model forWord Occurrences in Document Databases, Conference for Automated Learningand Discovery, Workshop on Learning from Text and the Web, CMU (1998).

[12] McCallum, A., Nigam, K., Rennie, J., and Seymore, K., Building Domain−SpecificSearch Engines with Machine Learning Techniques, AAAI−99 Spring Symposium onIntelligent Agents in Cyberspace (1999).

[13] Turney, P.D.: Learning algorithms for keyphrase extraction. Information Retrieval 2 (2000) 303–336

[14] Hammouda, K., Matute, D., and Kamel, M., " CorePhrase: Keyphrase Extraction for Document Clustering ", IAPR International Conference on Machine Learning and Data Mining MLDM' 2005, Leipzig, Germany, pp265-274, 2005.

[15] D. Yarowsky Unsupervised word sense disambiguation rivaling supervised methods, Published in Proceeding, ACL '95 Proceedings of the 33rd annual meeting on Association for Computational Linguistics, Pages 189-196.

[16] B. Krulwich, and C. Burkey (1996). Learning user information interests through the extraction of semantically significant phrases. In M. Hearst and H. Hirsh, editors, AAAI 1996 Spring Symposium on Machine Learning in Information Access. California: AAAI Press.

[17] D. Whitley, (1989). The GENITOR algorithm and selective pressure. Proceedings of the Third International Conference on Genetic Algorithms (ICGA-89), pp. 116-121. California: Morgan Kaufmann.

[18] IssamSahmoudi, Hananefroud and AbdelmonaimeLachkar, A New Keyphrases Extraction Method Based On Suffix Tree Data Structure For Arabic Documents Clustering, in International Journal of Database Management Systems (IJDMS) Vol.5, No.6, December 2013.

[19] P.D. Turney, (1997). Extraction of Keyphrases from Text: Evaluation of Four Algorithms. National Research Council, Institute for Information Technology, Technical Report ERB-1051.

[20] Ian H. Witten,Gordon W. Paynter, Eibe Frank, Carl Gutwin and Craig G. Nevill-Manning ,"KEA: Practical Automatic Keyphrase Extraction".

[21] R. Mihalcea, "Graph based Ranking Algorithms for Sentence Extraction, Applied to Text Summarization",ACL 2004.

[22] El-Beltagy, S. R. (2006) KP-Miner: A Simple System for Effective Keyphrase Extraction. In proceedings of the 3rd IEEE International Conference on Innovations in Information Technology (IIT '06), Dubai, UAE.

[23] Rodeh, M., Pratt, V. R. and Even, S. Linear algorithm for data compression via string matching. Journal of the ACM, 28(1):16-24, 1981.

[24] O. Zamir and O. Etzioni, "Grouper: a dynamic clustering interface to Web search results", Computer Networks, 31(11-16), pp. 1361-1374, 1999.

[25] Fellbaum C.(Ed.), "WordNet : An Electronic Lexical Database", MIT Press, 1998

[26] Wang, Qiang, Yunming Ye, Joshua Zhexue Huang and Shengzhong Feng, "Post-processing strategies for improving local gene expression pattern analysis", International Journal of Data Mining and Bioinformatics, 2013.

[27] Yanjun Li., and Chung S.M., "Text Document Clustering Based on Frequent Word sequences". Journal of Data and Knowledge Engineering, 64, 381-404 pp., 2008.

[28] RekhaBaghel and RenuDhir., "A Frequent Concepts Based Document Clustering Algorithm", International Journal of Computer Applications, 4(5), 6-12 pp., 2010.

[29] Y. Sri Lalitha, Dr. A. Govardhan "Semantic Framework for Text Clustering with Neighbors ", S.C. Satapathy et al. (eds.), ICT and Critical Infrastructure: Proceedings of the 48th Annual Convention of CSI - Volume II, Advances in Intelligent Systems and Computing 249, DOI: 10.1007/978-3-319-03095-1_29, © Springer International Publishing Switzerland 2014 pp.261-271.

[30] D. Gusfield. In Algorithms on strings, trees and sequences: computer science and computational biology. Cambridge University Press, 1997.

[31] R. Agrawal, R. Srikant, Fast algorithms for mining association rules, in: Proceedings of the 20th VLDB Conference, 1994.