Concept Based Text Document Clustering

K.Gunavathi
Department of computer science and engineering
Kumaraguru College of technology
Coimbatore, India

M.Manikandan
Department of computer science and engineering
Kumaraguru College of technology
Coimbatore, India

S.Thilagavathi
Department of computer science and engineering
Kumaraguru College of technology
Coimbatore, India

***Abstract*—A cluster is a collection of data objects that are similar to one another. A cluster of data objects can be treated collectively as one group and so it may be considered as form of data compression. Clustering is also called as data segmentation in some applications because clustering partitions large data sets into groups according to their similarity. Indexing of documents is based on the related or semantically related keywords. Topic based weighting scheme is proposed to index the text. It involves with identifying topic candidates, determine their importance, and detect similar and synonymous topics. The indexing algorithm uses topic frequency to determine their importance and existence of the topics. Concept based weighting scheme is used to index the document, it identifies topic candidates, determine their importance, detect the similar and synonymous topics. In this system the numbers of medical documents are collected, and then the documents are taken for document pre-processing which includes tokenization and stop word removal. Finally compare the topic based weighting scheme with other indexing schemes and prove that topic based indexing reduces the dimensionality of the data which is efficient even for very large databases and provides an understandable description of the discovered clusters by their frequent term sets.**

***Keywords*— Clustering algorithms, Indexing, Topic based weighting scheme, Concept based weighting scheme and MeSH ontology**

I. INTRODUCTION

A cluster is a collection of data objects that are similar to one another. A cluster of data objects can be treated collectively as one group and so it may be considered as form of data compression. Clustering is also called as data segmentation in some applications because clustering partitions large data sets into groups according to their similarity. Clustering can also be used for outlier detection. As a data mining function, cluster analysis can be used as a stand-alone tool to gain insight into the distribution of data, to observe the characteristics of each cluster. The purpose of text mining is to process unstructured information, extract meaningful numeric indices from the text, and make the information contained in the text accessible to various data mining algorithms. Information can be extracted to derive summaries for the words contained in the documents or to compute summaries for the documents based on the words contained in them.

Data mining is the process of analyzing data from different perspectives and summarizing it into useful information, Data mining uses the information from past data to analyze the outcome of a particular problem or situation that may arise. Data mining works to analyze data stored in data warehouses that are used to store that data that is being analyzed. It allows users to analyze data from many different dimensions, and summarize the relationships identified. Technically, data mining is the process of finding correlations or patterns among dozens of fields in large relational databases.

Clustering and classification are both fundamental tasks in data mining. Classification is used mostly as a supervised learning method, whereas clustering is used for unsupervised learning. The goal of clustering is descriptive, that of classification is predictive. Since the goal of clustering is to discover a new set of categories, Clustering groups data instances into subsets in such a manner that similar instances are grouped together, while different instances belong to different groups. The instances are thereby organized into an efficient representation that characterizes the population being sampled.

Document clustering has many important applications in the area of data mining and information retrieval. Many of the existing document clustering techniques uses the bag-of-words model to represent the content of a document. However, this representation is only effective for grouping related documents when these documents share a large proportion of lexically equivalent terms. In other words, instances of synonymy between related documents are ignored, which can reduce the effectiveness of applications using a standard full-text document representation.

The Purpose of text mining is to process unstructured information, extract meaningful numeric indices from the text, and to make the information contained in the text accessible to the various data mining algorithms. Information can be extracted to derive the summaries of the documents based on the words contained.

A novel concept based mining model is proposed. The goal of this approach is to mine the text through the analysis of higher level characteristics called concepts for minimizing the vocabulary problem and the effort necessary to extract useful information. Instead of applying the text mining techniques on terms or keywords , the discovery process works over concepts extracted from the text. Concepts represent real world attributes and as seen in discourse analysis, they help to understand ideas and ideologies present in the text.

Biomedical ontology improves the clustering quality for MEDLINE articles. The controlled vocabulary contains several different types of terms, such as descriptor, Qualifiers, Publication trees, Geographic, and entry terms. MeSH descriptors are organized in a MeSH tree, which can be seen as a MeSH concept Hierarchy.

Topic frequency is used to determine the importance and existence of the topics within a document. The weight of the term is calculated by the frequency of the terms that are used to calculate the dynamic weight of the topic in a document based on the semantic relations such as identity, synonymy, hypernymy and meronymy using the domain ontology as the background knowledge.

II. Literature Review

Luis Filipe da Cruz Nassif Et All (2013) proposed Concept-Based Mining Model for Enhancing Text Clustering that bridges the gap between natural language processing and text mining disciplines. The first component is the sentence based concept analysis which analyzes the semantic structure of the sentence. The second component, document based concept analysis, analyzes each concept at the document level. The third component analyzes the concepts on the corpus level using the document frequency. The fourth component is the concept based similarity measure which allows measuring the importance of each concept with respect to semantics of the sentence, topic of the document and the discrimination among documents in a corpus.

Junjie Wu (2013) proposed Model Based Method for Projective Clustering, in which high dimensional data projective clustering has been defined as an extension to traditional clustering that attempts to find projected clusters in subsets of the dimensions of a data space. In this method, a probability model is proposed to describe projected clusters in a high dimensional data space. A model based fuzzy projective clustering method is used to discover the clusters with overlapping boundaries in various projected subspaces.

Duc Thang Nuyen (2013) proposed Analyzing and Visualizing Web Opinion Development and Social Interactions with Density-Based clustering, the density based clustering algorithm and the scalable distance based clustering techniques are used for web opinion clustering. This web opinion clustering technique enables the identification of themes in web social networks and their development. Some interactive visualization tools are also developed, which make use of identified topic clusters to display the social network development, the network topology similarity between the topics.

Xiping Liu (2011) proposed Returning Clustered Results for Keyword Search on XML Documents, the problem of returning cluster-based search results for XML keyword search is investigated. New answer semantics for XML keyword query which is based on conceptually related relationship between nodes is proposed. Then a novel clustering methodology based on the notion of keywords matching pattern is used. Here two approaches are used. The first one is a conventional one, which does clustering in post phase. The second one is novel in that it performs clustering in an active way. The generated clusters can be further improved by organizing clusters into a hierarchy.

Zhiang Wu (2012) proposed Incremental Learning for Information –Theoretic Text Clustering Information which is used to exploit information-theoretic measures as clustering criteria. A common practice on this topic is called Info K-means, which performs K-means clustering with KL-divergence as the proximity function. The effectiveness of this replacement is guaranteed by an equivalent mathematical transformation in the objective function of Info K-means. Eduardo Raul Hruschka (2013) proposed Document Clustering for Forensic Analysis. This approach is illustrated by carrying out extensive experimentation with six well-known clustering algorithms, which is applied to five real-world investigations. The effectiveness of this measure is evaluated on several real-world entities for text classification and clustering problems. The results show that the performance is better than the conventional clustering methods.

III. Existing System

A. *Clustering Process*

Clustering is the task of grouping a set of objects in such a way that the objects in the same group are more similar to each other than to those in other groups.

The Clustering process consist of

- Representation Model
- Clustering Algorithm
- Similarity Measures
- Performance Measures.

1) Representation Model: Representation model is used when fusion processes are performed on linguistic values. The limitation of this model is the loss of information. The loss of information implies a lack of precision in the final results from the fusion of linguistic information. The various Representation Models are.

- Vector Space Model(VSM)
- Document Index Graph(DIG)
- Suffix Tree Document(STD)

2) Clustering Algorithm: Clustering is the process of grouping physical or abstract objects into classes of similar objects. Clustering algorithms are based on the cluster model. Clustering algorithms are broadly classified as.

- Hierarchical Algorithms
- Partition Algorithms

a) Hierarchical Algorithms: This algorithm produces a set of nested clusters organized as a hierarchical tree. There are many times when clusters have sub-classes within them, and which in turn have sub-classes of their own. Two types of hierarchical algorithm are.

- Agglomerative

- Divisive

b) Partition Algorithms: Partitioning methods relocate instances by moving them from one to another, starting from an initial partitioning. Such methods typically require that the number of clusters will be pre-set by the user. Two types of Partition Algorithms are.

- K-means Algorithm
- Bisecting K-means Algorithm

3) Similarity Measures: A similarity measure or similarity function is a real-valued function that quantifies the similarity between two objects. The notion of similarity measures has been successfully used in many domains such as data analysis,pattern recognition and machine learning.

4) Performance Measures: Performance measures are the process of collecting, analyzing and reporting information regarding the performance of an individual, group, organization, system or component.

Performance Measures consist of

- Precision
- Recall
- F-measure
- FM-Index

a) Precision: Precision is the fraction of retrieved instances that are relevant.

$$precision = \frac{|\{relevant\ documents\} \cap \{retrived\ documents\}|}{|\{retrived\ documents\}|} \quad (1)$$

b) Recall: Recall is the fraction of relevant instances that are retrieved. Both precision and recall are based in an understanding and measure of relevance.

$$recall = \frac{|\{relevant\ documents\} \cap \{retrived\ documents\}|}{|\{relevant\ documents\}|} \quad (2)$$

c) F-measure: It is a measure that combines the precision and recall. In statistical analysis of binary classification, the F-Measure is a measure of a test's accuracy. The F-Measure is often used in the field of information retrieval for measuring search, document classification, and query classification. The F-measure is also used widely used in the natural language processing.

$$F = 2.\frac{precision.recall}{precision+recall} \quad (3)$$

d) FM - Index: FM-Index is used to find the number of occurrences of a pattern within the compressed text, as well as locate the position of each occurrence. Both the query time and storage space requirements are sub linear with respect to the size of the input data.

5) Clustering Models: The various clustering models help to improve the clustering process.

The clustering model consists of

- Term Based Model
- Phrase Based Model
- Concept Based Model

a) Term Based Model: In this model, the vector space model(VSM) is used to represent text as a vectors of identifiers. The most popular term frequency measure is often considered as the default weighting scheme.

However, this scheme is pure statistical and does not incorporate any information about semantic or category that belongs to a term.

b) Phrase Based Model: In this model, it translates the phrase into atomic units. It includes suffix tree (STD) and document index graph (DIG). The suffix tree is a data structure that presents the suffix in a way that allow for particularly fast implementation of many important string operations. It uses the local context in translation.

c) Concept Based Model: The concept-based model can effectively discriminate between unimportant terms with respect to sentence semantics and terms which hold the concepts that represent the semantic of the sentence. A concept-based model that analyzes terms on the sentence and document levels rather than the traditional analysis of document.

B. Issues In The Existing Models

The limitations of the conventional document representation models are

- In VSM, the order in which the term appear in the document is lost.
- Difficulty in extracting semantically important indices.
- Synonymy and polysemy problems.

IV. Proposed System

A. Topic Based Indexing

Indexing of documents is based on the related or semantically related keywords. Topic based weighting scheme is proposed to index the text. It involves with identifying topic candidates, determine their importance, and detect similar and synonymous topics. The indexing algorithm uses topic frequency to determine their importance and existence of the topics. A concept based weighting scheme computes the importance of the underlying text by converting the document into a bag_of_concepts. Document clustering has been used for better document retrieval, document browsing and text mining.

Medical Subject Headings (MeSH), published by the National Library of medicine mainly consists of the controlled vocabulary and a MeSH Tree. The controlled vocabulary contains several different types of terms, such as descriptor, Qualifiers, Publications Types, Geographic and Entry terms. MeSH descriptors are organized in a MeSH tree, which can be seen as a MeSH concept hierarchy. In a MeSH tree, there are 15 categories (e.g.catagory A for anatomic terms) and each category is further divided into sub categories. For each subcategory, corresponding descriptors are hierarchically arranged from most general to most specific. Though the descriptors normally appears in more than one place in a tree, they are represented in a graph rather than a tree.

B. *Index Table*

The index table consists of all descriptor terms in the ontology for the particular concept.

Index table= (ID, descriptor terms, entry terms, value, weight)

Here ID represents the term.The descriptor terms are main concepts or main headings. Entry terms are the synonyms or the related terms to the descriptors. Value is represented as 'n' in a tree and it denotes parent, augmented by the index of 'n' among its siblings, adding dot to separate them.

C. *Topic Based Weighting Scheme*

A topic-based weighting scheme is used to index the text in the document. Documents are indexed based on the terms that are presented in the ontology. Tokenization and stop word removal are done during the pre-processing stage of the document. A keyword and an abstract are given as input to the system. The document is getting pre-processed and the words are getting stored in a text file. The given keyword is searched in the MeSH ontology for its existence. If it exists, the corresponding path and its synonyms will also get captured. All the terms in the ontology are compared with the given abstract. Topic frequency is used to determine the importance and existence of the topics.

D. *Comparison With Other Indexing Systems*

In the other indexing schemes, individual terms were considered for indexing and also only two relations, identity and synonymy were used for calculating the weight of the topics. But in proposed system, the indexing of the document is based on related keywords. The proposed system also reduces the dimensionality of the data which is efficient even for very large databases and provides an understandable description of the discovered clusters by their frequent term sets.

E. *Ontology Based Clustering*

The Ontology Based Clustering is a new approach for applying background knowledge during pre-processing in order to improve the clustering results. The input data is pre-processed by applying an ontology-based heuristics for feature selection and feature aggregation. Thus, a number of alternative text representations are constructed. Based on these representations, multiple clustering results using K-means are computed. The results may be distinguished and explained by the corresponding selection of concepts in the ontology. The results compare favorably with a sophisticated baseline pre-processing strategy.

F. *Mesh Ontology*

Ontology is very important for biomedical documents clustering. First, biomedical literature is usually composed of many complicated biomedical concepts with name variations containing usually more than one word. Second, bag-of-words model suffers from "the curse of dimension" and lacks interpretation power to clustering results.

Medical Subject Headings (MeSH) mainly consists of the controlled vocabulary and a MeSH Tree. The controlled vocabulary contains several different types of terms, such as Descriptor, Qualifiers, Publication Types, Geographic, and Entry terms.

Descriptor terms are main concepts or main headings. Entry terms are the synonyms or the related terms to descriptors. For example, "Neoplasm's" as a descriptor has the following entry terms {"Cancer," "Cancers," "Neoplasm," "Tumours'", "Tumours", "Benign Neoplasm," "Neoplasm, Benign"}. As a result, descriptors terms are used in this research. MeSH descriptors are organized in a MeSH Tree, which can be seen as the MeSH Concept Hierarchy.

Terms in each document are mapped to the entry terms in MeSH.The selected Entry terms are mapped into MeSH Descriptors to remove the synonyms. The indexing system matches the terms in each document to the entry terms in MeSH and then maps the selected entry terms into MeSH Descriptors.

Instead of searching all entry terms in the MeSH against each document, 1-to-5 gram words are selected as the candidates of MeSH Entry terms. Then, the candidate terms that are chosen, are matched with MeSH entry terms. Those semantically similar entry terms are replaced with the Descriptor term to remove synonyms.

V. System Design

A. *Modules*

- Document Pre-Processing
- Concept Mapping
- Experimental Setup

B. *Modules Description*

1) Document Preprocessing: Documents may be represented by a wide range of different feature descriptions. The most straightforward description of documents relies on term vectors. A term vector for one document specifies how often each term from the document set occurs in that document.

It includes two steps. They are

- Tokenization
- Stopword Removal

a) Tokenization: Tokenization is the process of breaking a stream of text into words, phrases, symbols, or other meaningful elements called tokens. The list of tokens becomes input for further processing such as parsing or text mining. Tokenization is useful both in linguistics and in computer science, where it forms part of lexical analysis. In tokenization process, each and every word to be separated individually are contained in particular document. The medical related Document such as asthma, anemia, cancer, eye disease, polio is collected from the Pub Med. The Pub Med documents acts as the input to the tokenization process. Each and every medical term from that document is tokenized and they are stored in a separate file.

b) Stop Word Removal: Stop words are the words which are filtered, after processing of natural language data. There is no definite list of stop words that are always used. Any group of words can be chosen as the stop words for a given purpose. For some search machines these are some of the most common short function words, such as the, is, at, which, and on. In this case, stop words can cause problems when searching for phrases that include them, particularly in names such as 'The Who', 'the', or 'Take That'.

2) Concept Mapping: The first database is maintained with medical terms, number of medical terms, count of the number of times that the medical terms occurred, and weight of the medical terms. In the second database, the concept term from the documents are collected and the frequency for the collected documents are calculated. The second database is compared with the first database and the concept weight is calculated. Finally the relevant document is produced with the document that contains the most relevant term weight.

TABLE I. MESH DATABASE

Index	Concept	Synonym 1	Synonym 2	Synonym 3
1	Anemia	Acquired Autoimmune	Anemia, Hemolytic, Cold Antibody	Anemia, Hemolytic, Idiopathic Acquired
1.1	Favism	Gasser Syndrome	Gasser's Syndrome	Chlorosis
1.1.1	Hemolytic anemia	Leukemia, Chronic Myelogenous	Neutrophilic Leukemia, Chronic	Leukemoid Reaction
1.2	Leukemia, Erythroblastic	Polycythemia	Erythremia	Erythremia
1.2.1	Leukemoid Reaction	Thrombocythemia	Hemorrhagic Thrombocythemia	Agnogenic Myeloid Metaplasia

TABLE II. TF/IDF

Doc id	Term	Tf/idf
D1	Neoplasm	1.4237
D2	Leukemia	1.2589
D3	Polycythemia	1.3789

TABLE III. CONCEPT WEIGHT

Doc No	Cancer	Anemia	Asthma
D1	1.0292	0.3472	0.1584
D2	4.0390	0.1791	0.1564

3) Experimental Setup: For experimental purpose ten documents are collected from Pub Med and after pre-processing 725 individual words are found. The medical related Documents such as asthma, anemia, cancer, eye disease, polio are collected from the Pub Med. The Pub Med documents acts as the input for the tokenization process. Each and every medical term from that document is tokenized and they are stored in a separate file. The words are compared with the MeSH ontology and for those words the weight is calculated. The term based weight and topic based weight is also calculated. In this, the topic weight is constant and the term weight is changed for different terms that are used. The higher topic weight shows that, it is more related to the corresponding topic. Here five documents are considered for for calculating the term weight and topic weight. The weight of topic is high when compare to term weigh.

a) PUBMED: For the experimental purpose the number of medical documents from the PubMed is collected. The documents are taken for document pre-processing. Then the tokenization has been performed and the stop words are removed.

K-means Clustering performance has been experimented using.

- FM Index
- Silhouette Index
- Jaccard Index

b) MeSH: MeSH (Medical Subject Heading) is a tree structure, that represents the hierarchy of the keywords occurred in the medical documents. The hierarchy is based on the weight of the term value.

The MeSH dataset is collected and the dataset is maintained in one index table. The index table contains the document name, term, number of times occurred, weight value. In another index table it contains the term, synonyms, weight value such as,

Index-table = (ID, descriptor term, entry term, value, weight)

Here ID represents the term number, descriptor terms are main concepts, and Entry terms are the synonyms or the related terms to descriptors. Value is represented as n in a tree and it denote parent, augmented by the index of n among its siblings, adding dot to separate them.

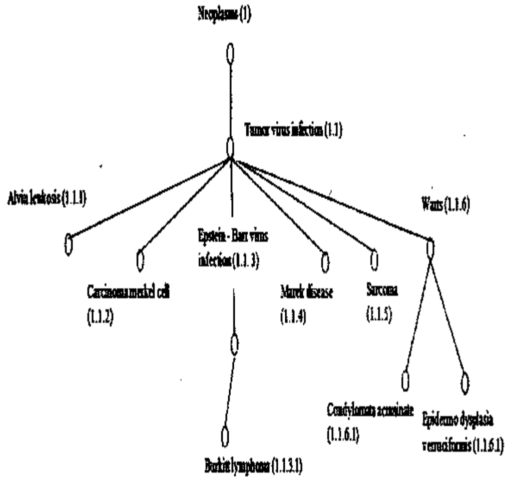
Neoplasms (1)
Tumor virus infection (1.1)
Alvia leukosis (1.1.1)
Carcinoma merkel cell (1.1.2)
Epstein - Barr virus infection (1.1.3)
Marek disease (1.1.4)
Sarcoma (1.1.5)
Warts (1.1.6)
Burkitt lymphoma (1.1.3.1)
Condylomata acuminate (1.1.6.1)
Epidermo dysplasia verruciformis (1.1.6.1)

Fig. 1. Descriptor terms in the form of tree.

c) Topic Based Weighting Scheme: A topic base weighting scheme is used to index the text in the document. Documents are indexed based on the terms that are presented in the ontology. Tokenization and stop word removal are done during the pre-processing stage of the document. A keyword and an abstract are given as input to the system. The document is getting pre-processed and the words are getting stored in a text file. The given keyword search in the MeSH ontology for its existence. If it exists, the corresponding path and its synonyms will also get captured.

All the terms in the ontology are compared with the given abstract is calculated by using the frequency of the term and their weight. The total weight S for the topic is calculated is shown in equation.

$$.S(topic) = \frac{\sum_{i=1}^{N} S_{word(w_i)}}{N} \qquad (4)$$

The weight of the individual words can be calculated as follows.

$$S_{word\,(w_i)} = nR^k_{w_i} \times SR^k_{w_j} + nR^k_{w_j} \times SR^k_{w_j} \qquad (5)$$

Where N denotes the number of unique words in the abstract,

k = represents the relation of the word either identity or synonymy,

nR = represents the no of occurrence of relation,

SR = denotes the weight assigned to that relation,

The highest score represents the important of the word in the abstract.

a) Term Selection: It produces a low dimensional representation. Selection of terms is based on the information retrieval measure tf(i,j). Let tf(i,j) be the term frequency of term j in a document di . Let df(j) be the document frequency of term j that counts in how many documents term j appears. Then tf/idf (term frequency / inverted document frequency) of term j in document is defined by.

$$W_j = Total\ no.of\ N \times \frac{tf}{idf}(i,j) \qquad (6)$$

4) Results

TABLE IV. COMPARISON OF TERM BASED WEIGHT AND TOPIC BASED WEIGHT

Documents	**Term name**	**Occurrence of the term in the given document**	**Term based weight**	**Topic based weight**
Document 1	Cell Carcinoma Burkitt	1 1 1	0.0769 0.0723 0.0793	0.8846 0.8815 0.8846
Document 2	Barr Virus Human	2 3 1	0.1666 0.25 0.0833	1.1 1.1 1.1
Document 3	Marek Disease Virus	2 4 1	0.2867 0.5714 0.1428	0.7571 0.7514 0.7542

The term based weight and topic based weight values are shown in form of graph in Fig.2

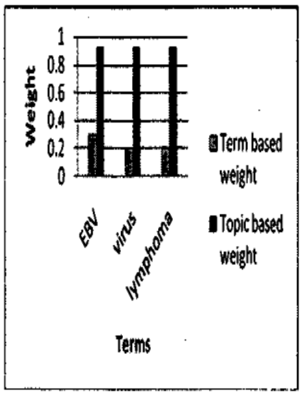

Fig. 2. Wight Vs Terms for document.

TABLE V. MESH DATABASE

ID	**Concept name**	**Term**	**Frequency**	**Weight**
1	Asthma	Carcinoma	2	1.913636
2	Anemia	Favism	3	3.820404
3	Cancer	Cysts	1	0.156162584

CONCLUSION

In topic based indexing, the indexing of document is based on related or semantically related keywords. Concept based weighting scheme is used to index the document, it identifies topic candidates, determine their importance, detect the similar and synonymous topics. Topic based indexing reduces the dimensionality of the data which is efficient even for very large databases and provides an understandable description of the discovered clusters by their frequent term sets.

FUTURE WORK

In existing system, there are some limitations such as order in which the term appear in the document is lost. The other difficulty is extracting semantically important indexes, synonymy and polysemy problems. In this system the numbers of medical documents are collected, and then the documents are taken for document pre-processing which includes tokenization and stop word removal.

The terms occurred in the documents are represented by using MeSH ontology tree structure. But in this MeSH ontology tree structure, it has repeated medical terms in which the terms appear more than one place in the tree structure. So in future work the graph structure is considered rather than tree structure because the graph structure will achieve the consistency of the terms.

REFERENCES

[1] P. Tan, M. Steinbach, and V. Kumar, Introduction to Data Mining, 1st ed. Addison Wesley, May 2005.

[2] A. Brandi and P. Blonde, “A survey of fuzzy clustering algorithms for patternt recognition.” IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics, vol. 29, no. 6, pp. 778–785, Dec. 1999.

[3] A. Silic, M.-F. Moans, L. Zmak, and B. Basic, “Comparing document classification schemes using k-means clustering,” vol. 5177, pp. 615–624, 2008.

[4] J. Han, Data Mining: Concepts and Techniques. CA,USA: Morgan Kaufmann Publishers Inc., 2005.

[5] Y. Zhao, G. Karypis, and U. Fayyad, “Hierarchical clustering algorithms for document datasets,” Data Min.Knowl. Discover. vol. 10, no. 2, pp. 141–168, Mar. 2005.

[6] Y. Zhao and G. Karypis, “Evaluation of hierarchical clustering algorithms for document datasets,” in Proceedings of the eleventh CIKM ’02, NY, USA, 2002, pp. 515–524.

[7] Y.J. Li, C. Luo, and S.M. Chung, “Text Clustering with Feature Selection by Using Statistical Data,” IEEE Trans. Knowledge and Data Eng., vol. 20, no. 5, pp. 641-652, May 2008.

[8] W. Lian, D.W.-L. Cheung, N. Mamoulis, and S.-M. Yiu, “An Efficient and Scalable Algorithm for Clustering XML Documents by Structure,” IEEE Trans. Knowledge and Data Eng., vol. 16, no. 1,pp. 82-96, Jan. 2004.

[9] Z. Liu and Y. Chen, “Identifying Meaningful Return Information for XML Keyword Search,” Proc. ACM SIGMOD Int’l Conf.Management of Data, 2007.

[10] Z. Liu and Y. Chen, “Reasoning and Identifying Relevant Matches for XML Keyword Search,” Proc. VLDB Endowment, vol. 1, no. 1,pp. 921-932, 2008.