# FREQUENCY BASED SEMANTIC RELATIVE ANALYTIC HIERARCHY TREE INDEXING FOR TEXTUAL WEB CONTENT MINING

**D.Dhayalan[1], Dr.C.Chandrasekar[2]**

*[1]Research scholar, Department of Computer Science, Dravidian University kuppam, A.P, India*
*and*
*Assistant Professor, Vel Tech High Tech Dr Rangarajan Dr Sakunthala Engineering College Avadi Chennai-62 ,*
*[2]Professor, Department of Computer Science, Periyar University, Salem, Tamil Nadu, India*

*[1]filldhayainbox@gmail.com, [2]ccsekar@gmail.com*

***Abstract***

***The quality of extracted features is the key issue to textual web content mining due to the large number of words, sentences, phrases, and noise. Most existing textual web content mining methods are based on contextual information or content credibility which extract attributes from a training set for describing relevant information. However, the quality of the extracted attributes in contents may be not high because of lot of noise in contents. For several years, researchers made use of various content credibility approaches that have more semantics than single words to improve the relevance. But, many experiments do not support the effective use of content credibility since they have low frequency of identifying the most significant factors, and include redundant and noisy contextual information due to data sparseness. In this paper, we propose a novel textual web content mining method using semantic relations, called, Domain Frequent-Analytic Hierarchy Tree (DF-AHR). This method initially discovers domain frequent knowledge in textual web content for identifying the most informative contents by performing pre-processing. By applying domain frequent model irrelevant words are removed, therefore achieving dimensionality reduction. The identified contents are then utilized to extract useful features for textual web content mining. Finally, a novel Analytic Hierarchy Tree method based on semantic relations which allows to combines the pieces of document to discover the knowledge (features) according to the web user specific needs is presented. Substantial experiments on Freebase Data Dump demonstrate that the proposed method achieves encouraging performance in terms of accuracy, computational time and computation complexity.***

***Keywords:*** ***Web Content Mining, Semantic Relations, Data Sparseness, Domain Frequent, Analytic Hierarchy Tree***

## I. INTRODUCTION

Textual web content necessitates the creation of significant numerical indices from the unstructured text and then processing these indices with the help of several data mining techniques. A Quantitative predictive model was designed in [1] for Web Content Credibility, based on human evaluations called as, Quantitative Predictive model for Web Content Credibility (QP-WCC). Initially, to minimize the dimensionality, Latent Semantic Analysis (LSA) and Singular Value Decomposition (SVD) were used. Followed by the dimensionality reduced values, clustering was done with the help of expectation-maximization clustering algorithm.

Finally, to fasten the analysis, unsupervised learning methods were used. From the experiments conducted, the QP-WCC method ensured credibility evaluation through assigned labels. Hence, the predictive models designed on the basis of label frequency achieved high quality level with the aid of random forest approach that formed a basis to create sound models of Web content credibility. Despite accuracy, being achieved, the methods for estimating the most significant factors remained unsolved. To address this issue, in this work, Domain Frequent Pre-processing model is applied to extract the most

significant features (i.e. domains). Hence, with this most significant domain frequent feature identified, the time consumed for textual content web mining is said to be minimized.

Community Question Answering Knowledge Bases (CQA-KB) [2] used semantic class, for browsing community question answering archives. The CQA-KB extracted features on the basis of early finding that semantic relations, between terms conveyed in the same session, were found to be communicative for the classification task. In CQA-KB, the effectiveness extracted from the target search query was studied and also the preceding historical elements with respect to the corresponding transaction were analyzed. This was performed by applying eleven different groups of features covering surface, syntactic and semantic characteristics.

Due to this the retrieval of related answers were improved by enhancing the classification time and accuracy. Despite improvement found in the classification time and accuracy, data sparseness caused by less frequent semantic categories remained unsolved. To address this issue, in this work, data sparseness is addressed by applying Analytic Hierarchy Tree that maximizes the sparsity so that good domain specific groups are formed. In this way well defined answers to the web user specific needs are addressed.

Based on the aforementioned problem, in this work a method for textual content web mining based on machine learning technique using semantic relations, called, Domain Frequent-Analytic Hierarchy Tree (DF-AHR) method is presented. Based on the two delimiters, irrelevant contents present in the documents are removed. Followed by this Domain Frequent Pre-processing model is applied to extract frequent occurrences of domains. Finally, Domain Sense Disambiguation (DSD) is applied to mine correct semantic relation based on domain usage. From our experimental results, it can be verified that the use of the DF-AHR method provides an efficient means for textual web content mining using semantic relations.

The rest of the paper is structured as follows. First, Section 2 discusses related work and provides the theoretical framework for designing the objectives of our work. Then, in Section 3, the proposed method Domain Frequent-Analytic Hierarchy Tree (DF-AHR) is described in detail. In Section 4, the experimental results are provided followed by which in Section 5, discussion is made. Finally, Section 6 concludes the paper.

### 1.1 Related works

Applications in which a user communicate with several item set include, social media browsing through e-commerce websites, audio/video streaming and so on. In such applications, computerized generation of item recommendations has become a crucial element, making a user to investigate a large set of options. In [3], matrix factorization method was applied resulting in the increased accuracy of prediction.

Frequent subgraph mining was investigated in [4] to achieve elegant mining performance. However, with higher amount of uncertainty being involved, accuracy was said to be compromised. To address this issue, crowd sourcing [5] was applied to top-k queries minimizing the crowd time. Yet another efficient algorithm called, Top K Utility item sets (TKU) [6] provided an insight into greater amount of scalability. K Nearest Neighbour approach based on Various Widths Clustering (KNN-VWC) [7] on the other hand reduced the computational cost involved in mining.

Suggestions based on the keywords have become one of the most significant features of commercial Web search engines. With the submission of a keyword, the user may not be satisfied with the results. In

those cases, a set of keyword that is highly likely to refine the user's search in the correct direction is said to be more preferable.

A weighted keyword document graph [8] provided means for semantic relation between keywords and therefore resulting in the minimization of response time. Yet another semantic web mining was implemented for e-learning in [9]. Tripartite Citation Analysis was investigated in [10] for the identification of authors based on similar themes and therefore improving the similarity rate. Supplemented Latent Semantic Analysis (SLSA) [11] efficiently analyzed the word correlations in documents, therefore improving the classification accuracy.

In the current era, one of the most essentials for human is the e-commerce website. Besides, its popularity, this vision has become complicated because the prevailing e-commerce systems are not able to identify the expected product. Hence, there arises a need for search personalization. A conceptual model based on long term behaviour signals was presented in [12]. Yet another, effect of semantic web technologies on distance education was presented in [13]. A systematic mapping study was investigated in [14]. A rule based compositional approach for factuality assignment was designed in [15].

Examining huge user-generated microblogs is extremely critical in several fields, attracting many researchers to study. However, it is very demanding to exercise such noisy and short microblogs. In [16], a method to measure structure similarity was presented using Laplacian matrix to model semantic relations between microblogs. A review of mining social semantics on web was presented in [17]. An efficient algorithm for measuring semantic similarity of text was presented in [18] therefore improving the extraction of relevant sentences. Yet another, artificial neural network and genetic algorithm-based approach was investigated in [19], therefore improving the true positive rate.

The extensive amount and multiplicity of the content shared on social media can create a problem for any business in identifying the prospective customers. In [20], both unsupervised and supervised learning methods were used to classify the target audience on Twitter with minimal annotation efforts. This was said to be achieved using Twitter Latent Dirichlet Allocation (LDA). Followed by which, Support Vector Machine (SVM) was then ensembled using content from users of several topic domains identified by Twitter LDA. This in turn identified the target audience with high accuracy.

However, by missing the most significant features and limitations related to data sparseness, web users are often overwhelmed with information when trying to analyze various reviews from textual web content. So far, many researchers have tacked the problem of providing meaningful analysis for textual web content mining from a large number of reviews relying on data-mining tools. Considering a similar problem, this work is an effort to provide a machine learning technique by extracting the most significant feature using domain frequent pre-processing and constructing analytic hierarchy tree for textual web content mining using semantic relations.

### 1.2 Domain Frequent Analytic Hierarchy Tree

Despite huge amount of data in Web, each web page denotes similar information in a different manner. How to identify semantically similar data is a very challenging task with several practical applications available. The focus of our work remains in the extraction of data from textual web content. With the efficient extraction of such data, helps in providing services to the clients. Several methods are available for textual web content mining using semantic relations. In this work, machine learning technique, namely, Domain Frequent-Analytic Hierarchy Tree (DF-AHR) is applied for fast retrieval of information from textual web content using semantic relations. Figure 1 shows the flow diagram of the DF-AHR method.

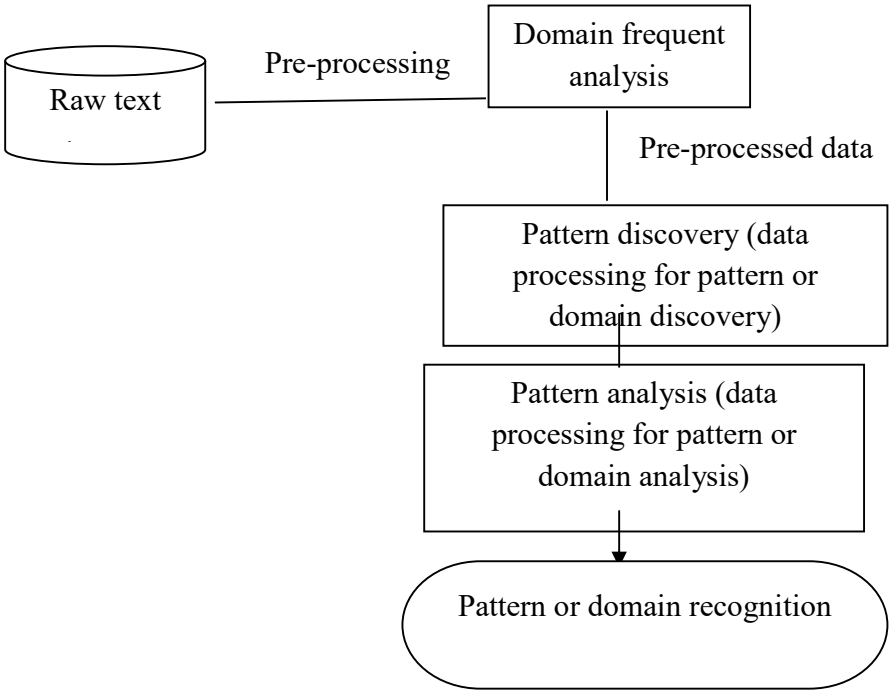


**Figure 1 Flow diagram of Domain Frequent-Analytic Hierarchy Tree**

As illustrated in the figure, the proposed method starts with the raw text web content as input. The pre-processing step in the figure prepares data for pattern extraction. In this step, raw text web content is transformed into certain data representation format that forms the input for pattern discovery algorithms. In this work, domain relevant document are extracted by applying Domain Frequent Pre-processing (DFP) model. As the tasks accomplished in the pre-processing step are critical for the success of the entire textual web content mining process, the data characterization using DFP conserves the patterns hidden in the documents in such a way that mining is carried out in an efficient manner in the succeeding step.

With the pre-processed content, the Analytic Hierarchy Tree Construction algorithm is applied to extract the hidden patterns or domains for further classification. In this way, by using semantic relations, occurrences of similar domain in the document or similar domain in other document are analyzed. As a result, textual web content mining using semantic relations is said to be performed based on the position. Therefore, textual web content recognition rate is said to be improved at a faster rate. The elaborate description of the DF-AHR method is described below.

### 1.3 Domain Frequent Pre-processing model

The first and foremost step in the DF-AHR method is the process of cleaning with the objective of preparing the raw web content data for efficient mining. Textual web content includes lots of noise and irrelevant parts such as scripts, tags, advertisements and so on. Besides, several words in the sentence do not have positive influence on mining and keeping those words results in higher dimensionality making the mining process more difficult. Hence, in this work, pre-processing of raw text web content is

performed to make the mining process simpler. With this, the performance of the mining is said to be improved and also speeds up the textual content retrieval process.

To start with, parsing is initially done, where the raw web content data comprises of multiple fields. With the most widely used separator characters as comma ',' and space '()', removal of these are initially performed. Followed by which the data cleansing is performed. The data cleaning cleans the raw text web content by removing irrelevant contents with the objective of only preserving the required data. Figure 2 shows the block diagram of Domain Frequent Pre-processing model

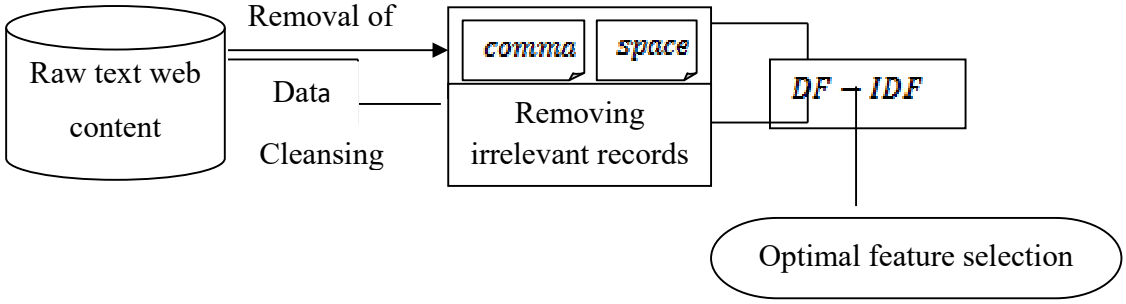


**Figure 2 Block diagram of Domain Frequent Pre-processing model**

As illustrated in the above figure, Domain Frequent Pre-processing model is used. The Domain Frequent Pre-processing model parses html tags that extract each text tag separately. It then partitions them on the basis of dot for each specific statement, removing stop words and stemming them. Finally, the frequency of each domain is calculated, therefore achieving feature selection. The Domain Frequent Pre-processing is mathematically evaluated as given below.

$$\mathbf{DF - IDF} = \mathbf{N_{dom\_occ}} * \log\left(\frac{\mathbf{N}}{\mathbf{D_{dom}}}\right) \quad (1)$$

From the above equation (1), the domain frequent inverse domain frequency '$\mathbf{DF - IDF}$' is arrived at based on the product of the number of domain occurrences in the document '$\mathbf{N_{dom\_occ}}$' and the logarithmic ratio of the number of documents '$\mathbf{N}$' to the number of documents that contains this domain '$\mathbf{D_{dom}}$' respectively. The pseudo code representation of Domain Frequency Pre-processing is as given below.

**Input**: raw text web content
**Output**: Optimal features selected

```
Begin
    Repeat
        If raw text web content contains delimiter character ',' or '()' then
            Split and store them as tokens
            If raw text web content has resource with extension 'gif' or 'jpeg' or 'css' then
                Delete the resource
                Measure domain frequency using equation (1)
            End if
        End if
    Until (end of raw text web content occurred)
End
```

**Algorithm 1 Domain Frequency Pre-processing**

As given in the above algorithm, the Domain Frequency Pre-processing algorithm, initially, obtains raw web content data as input, where the raw web content data includes several domains like '**Sports**', '**Business**', '**Banks**' and so on. With the objective of extracting or selecting optimal features, Domain Frequency Pre-processing is applied in the proposed method. Here, data cleansing and parsing is performed, followed by which domain frequency formulation is applied to extract the domain relevant document. Finally, optimal features related to various domains are extracted that in turn makes the mining faster.

### 1.4 Analytic Hierarchy Tree Construction

Related works have largely focused on browsing community question-answering archives [2] and web content prediction using the content credibility corpus [1], rather than the discovery of the semantic relationships within specific domains or patterns. Hence, a technique is required to detect the domains or patterns using semantic relations that provide certain relation in text, across domains.

With this objective in mind, a machine learning technique is applied to discover the pattern using semantic relations for textual web content mining. In this respect, Domain Sense Disambiguation (DSD) has been applied to mine only the correct semantic relation based on domain usage. DSD involves the assignment of appropriate tags to domains based on the context in which the domains occur. Figure 3 shows the sample documents each including three different domains, sports, business and research, present in '$\mathbf{n}$' different documents.

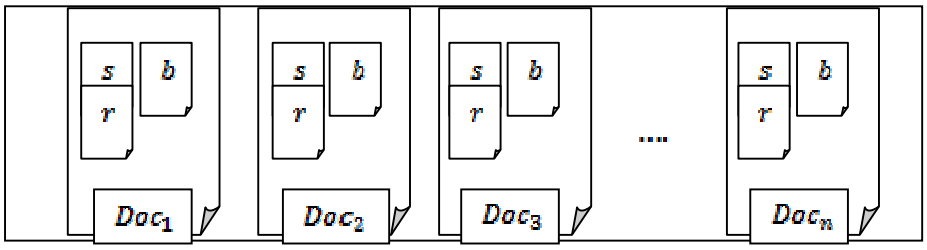


**Figure 3 Sample web log files involving '$\mathbf{n}$' documents with three different domains**

As illustrated in figure 3, for example with several documents in the web content represented as '$\mathbf{Doc = Doc_1, Doc_2, \ldots, Doc_n}$', domain represented by '$\mathbf{D = d_1, d_2, d_3}$'. Here, the documents are presented in the form of rows and domains are presented in the form of columns, therefore forming a matrix representation. Let us assume that the domain is presented as '$\mathbf{d_1}$' for sports, '$\mathbf{d_2}$' for business, '$\mathbf{d_3}$' for research and so on and is formulated as given below.

$$\mathbf{Doc_1 d_1 \rightarrow Domain\ Sport\ in\ Document\ 1} \quad (2)$$

$$\mathbf{Doc_1 d_2 \rightarrow Domain\ Business\ in\ Document\ 1} \quad (3)$$

$$\mathbf{Doc_1 d_3 \rightarrow Domain\ Research\ in\ Document\ 1} \quad (4)$$

With the domain '$\mathbf{d_1}$' in '$\mathbf{Doc_1}$' represented by several attributes '$\mathbf{d_1 s_1, d_1 s_2, \ldots d_1 s_3}$' related to sports, domain '$\mathbf{d_2}$' in '$\mathbf{Doc_1}$' represented by several attributes '$\mathbf{d_2 b_1, d_2 b_2, \ldots d_2 b_n}$' related to business and '$\mathbf{d_3}$' in '$\mathbf{Doc_1}$' represented by several attributes '$\mathbf{d_3 r_1, d_3 r_2, \ldots d_3 r_n}$' related to business and so on, the proposed work applies Analytic Hierarchy Tree to enhance, the indexing considering the semantic

relation between domains found on the page. The advantage of Analytic Hierarchy Tree index remains in web content sorting that allows searches and sequential access at the same time. As a result, higher amount of textual web content retrieval is said to be achieved and therefore enhances the current mining technology on the web.

While constructing Analytic Hierarchy Tree, frequency is used to evaluate how documents are related to the web user requirements, by searching in the document for the number of occurrences of the required domains. The higher the frequency the more important the corresponding criterion is said to be. Next, based on the semantic distance between the domains, the proposed work assigns a score. This is performed by calculating the semantic distance. The higher the score, the better the performance of the optimal for the considered criterion is said to be. Finally, the semantic relation determines the global score for each domain and consequent indexing.

In the previous works, they took into account the explicit mapping between search queries and questions [1]. But in DF-AHR method, to address sparseness, the semantic relations using Analytic Hierarchy Tree is constructed to ensure that the document is highly related to the domain selected. Each frequency is weighted to indicate their priority. If '$fd_i$' represents the frequency of domain '$d_i$', then the total frequency relevance is mathematically evaluated as given below.

$$F_{rel}(D) = \frac{\sum_{i=1}^{n} w_i fd_i}{|D|} \quad (5)$$

From the above equation (5), the frequency relevance '$F_{rel}$' domain for a specific document '$D$' is the ratio of the frequency '$f$' of domain '$i$' with its corresponding weight to the total documents considered. With the objective of addressing data sparseness, semantic relation between two domains that considers the weight of their hierarchical structure in domains is considered. The main objective of the work is to consider the position between two domains based on the semantic relation which is obtained by evaluating the distance. The distance between two domains is measured by the distance between domains in a document. It is mathematically formulated as given below.

$$\sum_{i,j=1}^{n} SD(d_i, d_j) = \sum_{i,j=1}^{n} \frac{1}{Distance(d_i, d_j)} / Doc_{size} \quad (6)$$

From the above equation (6), '$SD(d_i, d_j)$', represents the semantic distance between domain '$d_i$' and '$d_j$'. Followed by this, finally, the semantic relation between domains '$SR(d)$' in web page tagged with title, head and body '$p = 3$' is calculated and as given below.

$$SR(d) = \sum_{p=1}^{3} w_p * \sum_{i=1}^{n} w_i fd_i * \sum_{i,j=1}^{n} SD(d_i, d_j) \quad (7)$$

From the above equation, with the aid of for each domain, the semantic relation between this domain and all other searched domains and their synonyms and hyponyms for each part in documents are identified. If the domain occurred several times in the sentence, each sentence is dealt separately. On the other hand, if no domain occurred, the paragraph is considered as a whole. The pseudo code representation of Analytic Hierarchy Tree is given below.

| **Input**: Web document '$\mathbf{Doc} = \mathbf{Doc_1, Doc_2, \ldots, Doc_n}$', Domain '$\mathbf{D} = \mathbf{d_1, d_2, d_3}$' |
|---|
| **Output**: optimal textual web content mining |
| 1: **Begin**<br>2: **For** each Web document '$\mathbf{Doc}$'<br>3: **For** each Domain '$\mathbf{D}$'<br>4: Measure total frequency relevance using equation (5)<br>5: Measure length between domains in a document using equation (6)<br>6: Measure semantic relation between domains using equation (7)<br>7: **End for**<br>8: **End for**<br>9: **End** |

**Algorithm 2 Analytic Hierarchy Tree Construction algorithm**

As given in the above analytic hierarchy tree construction algorithm, the textual web content mining using semantic relations is performed. The above algorithm involves a mining process for documents based on the semantic relation between domains and the query terms. For this, three steps are performed. To start with, initially, the frequency relevance is measured. With the measured frequency relevance, next the distance between the domains in a document are measured to identify the occurrence of similar domain in the document and/or the occurrence of similar domain in the other domain. Finally, with this measured domain relevance between documents, semantic relation between domains is measured to ensure that the document is related to the domain selected, that in turn enhances the current search technology on the web.

## II. Experiments

The experiments were performed on a desktop PC running Windows XP with a P4-2.6G CPU and 1 G RAM. The Analytic Hierarchy Tree Construction algorithm was implemented using Java (JDK 1.4.2). Three domain news data are selected from Freebase Data Bump. Freebase is an open database of the world's information. It comprises of millions of topics in several hundreds of categories. It has been collected from large open data sets like Wikipedia, MusicBrainz, and the SECarchives. Besides, it also contains structured information covering several popular topics, like, movies, music, people and locations.

For the evaluation of DF-AHR, the Freebase Data Bump is randomly divided into 60% training and 50% testing documents, so that both training and testing dataset are disjoint. The experiments in the DF-AHR are repeated for 10 times, and final performance is reported by averaging the results.

Accuracy is used as an evaluation measure. The feature pre-processing accuracy is a measure to evaluate the significance of optimization. The accuracy is measured using the ratio of number of correct pre-processed domains to the total number of domains in target document and is as given below.

$$\mathbf{A} = \sum_{i=1}^{n} \frac{\mathbf{CPre-Processed_d}}{\mathbf{Doc_i}} \quad (8)$$

From (8), the accuracy factor '$\mathbf{A}$' is measured using the total document '$\mathbf{Doc_i}$' to the correct pre-processed domains '$\mathbf{CPre - Processed_d}$' respectively. It is measured in terms of percentage (%). The second evaluation measure used in the DF-AHR method is computational time or the time consumed for textual web content mining. It is mathematically formulated as given below.

$$\mathbf{CT} = \mathbf{n} * \mathbf{Doc_{size}} * \mathbf{Time}\,[\mathbf{SR}\,(\mathbf{d})] \quad (9)$$

From the above equation (9), the computational time for textual web content mining '$CT$' is formulated based on the number of documents '$n$', the document size '$Doc_{size}$' and the time consumed for arriving at semantic relations between domains '$Time\ [SR\ (d)]$'. It is measured in terms of milliseconds (ms). The final and the third evaluation measure is the computational complexity. The computational complexity measures the efficiency of the algorithm. In other words, the computational complexity measures the required to execute the algorithm.

$$CC = n * MEM\ [SR(d)] \quad (10)$$

From the above equation (10), the computational complexity for textual web content mining '$CC$' is arrived at using the number of documents in the web '$n$' and the memory consumed during semantic relations between domains '$MEM\ [SR(d)]$'. It is measured in terms of Kilo Bytes (KB).

## III. Discussion

The first set of experiments is intended to evaluate the accuracy performance of the proposed Domain Frequent-Analytic Hierarchy Tree (DF-AHR). In the experiment, to perform a precise comparison of the features of both Domain Frequent-Analytic Hierarchy Tree (DF-AHR) and existing Quantitative Predictive model for Web Content Credibility (QP-WCC) [1] Community Question Answering Knowledge Bases (CQA-KB) [2], the accuracy involved for textual web content mining is simplified in equation (8).

The accuracy involved during textual web content mining measure the accuracy with respect to the documents and user request provided as input. For this experiment, test set of 100 user requests with five different documents each possessing different set of domains were collected to verify the rate of accuracy. Sample calculation for arriving at accuracy is given below.

**Sample calculation**

- **QP-WCC**: With '$10$' user requests being made and a document size of '$10$', the accuracy is evaluated as given below.

$$A = \frac{7}{10} * 100 = 70\%$$

- **CQA-KB**: With '$10$' user requests being made and a document size of '$10$', the accuracy is evaluated as given below.

$$A = \frac{6}{10} * 100 = 60\%$$

- **Proposed DF-AHR**: With '$10$' user requests being made and a document size of '$10$', the accuracy is evaluated as given below.

$$A = \frac{8}{10} * 100 = 80\%$$

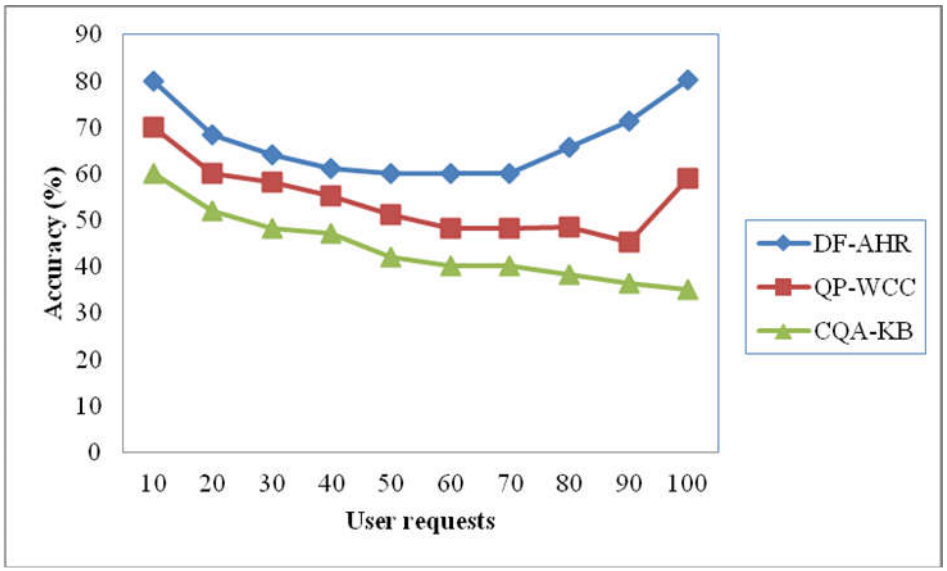


**Figure 4 Measure of accuracy**

Figure 4 shows the measure of accuracy with user requests in the range of 10 to 100 with different number of documents involving several domains collected from Freebase Data Dump. The decision point of ten different user requests (music, people and locations) was selected in a random manner that achieved a substantial improvement in ratings from the previous decision. The results show better performance of the proposed DF-AHR, but however not seen to be linear due to the presence of certain irrelevant information, not removed during the pre-processing stage. The final values plotted in the graph confirm the working hypothesis that the rate of accuracy for textual web content mining using Analytic Hierarchy Tree increases with the increase in the user requests.

As illustrated when compared to two other methods QP-WCC [1] and CQA-KB [2], the DF-AHR method substantially improved the rate of accuracy for textual web content mining using the Domain Frequency Pre-processing algorithm. This is because the DF-AHR method adapted a Domain Frequent Inverse Domain Frequency to perform the pre-processing. Here, based on the domain frequency, pre-processing is said to be performed separately for each domains (music, people and locations) to decide upon the mining process. Furthermore based on the two delimiter characters, splitting of the tokens for domain characterization was formed. This in turn improved the rate of accuracy using DF-AHR method by 25% compared to QP-WCC and 57% compared to CQA-KB.

Next, we address the second goal of the experiments with respect to computational time showing the comparison between DF-AHR method, QP-WCC [1] and CQA-KB [2] respectively. The computational time involves the time consumed for textual web content mining. It is evaluated from the equation (9). Covering millions of topics in hundreds of categories from Freebase Data Dump, the training dataset includes number of user requests in the range of 10 to 100. The results of 10 different documents obtained from Freebase Data Dump for experimental setup is listed in figure 5. Sample calculation for arriving at computational time is given below.

**Sample calculation**

- **QP-WCC**: With '10' user requests being madse and a document size of '3', the time consumed for semantic relations between domain being '0.035ms', the computational time for textual web content mining is evaluated as given below.

$$CT = 10 * 3 * 0.035 = 1.05ms$$

- **CQA-KB**: With '10' user requests being made and a document size of '3', the time consumed for semantic relations between domain being '0.059ms', the computational time for textual web content mining is evaluated as given below.

$$CT = 10 * 3 * 0.059 = 1.77ms$$

- **Proposed DF-AHR**: With '10' user requests being made and a document size of '3', the time consumed for semantic relations between domain being '0.023ms', the computational time for textual web content mining is evaluated as given below.

$$CT = 10 * 3 * 0.023 = 0.69ms$$

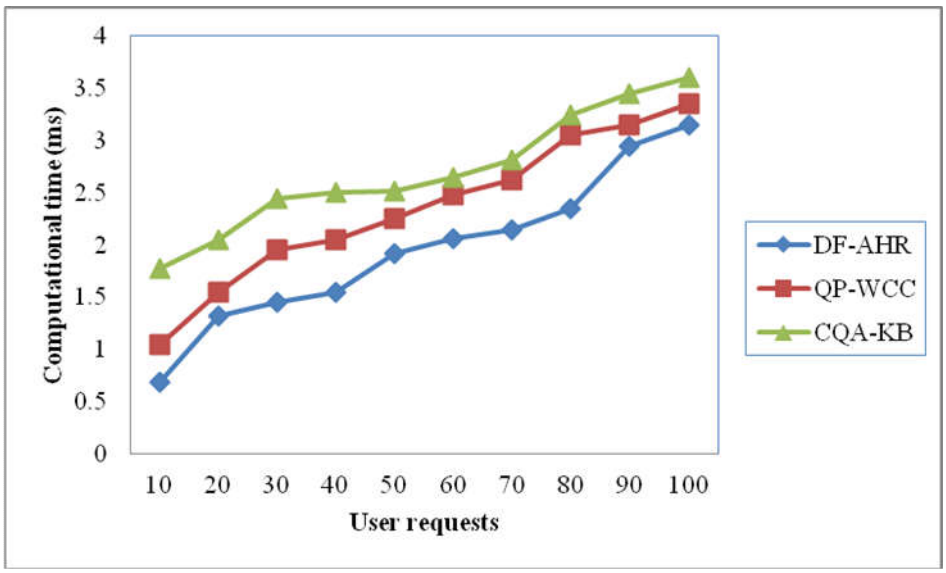


**Figure 5 Measure of computational time**

Results are presented for different user requests and differing document sizes including different domains. Higher, the number of user requests, higher the document size and therefore higher the time consumed for textual web content mining. This is because with higher user requests, the size of the document also grows exponentially and therefore the computational time for textual web content mining. Though increase in computational time is observed, but the increase is not said to be directly proportion with respect to user requests. This is because, with 10 user requests being processed, the resultant mining is used for further mining of next set of 10 user requests. However, from the figure it is evident that the computational time for textual web content mining is comparatively observed to be lower using the proposed DF-AHR method.

By applying, Domain Sense Disambiguation (DSD) in DF-AHR method, several domains present in the similar document is mined efficiently, comparing their domain frequency value. This in turn removes the noise or irrelevant information present in the document resulting in minimizing the computational time. The process is repeated with user requests of 10 to 100 for conducting experiments. The results reported here confirm that with the increase in the number of user requests, the computational time also increases, though betterment achieved using DF-AHR method.

As shown in the figure, when compared to two other methods QP-WCC [1] and CQA-KB [2], the DF-AHR method had better changes using the extensive Domain Frequency Pre-processing algorithm. This is because the Domain Frequency Pre-processing algorithm applied in DF-AHR method enhances the domain discrimination by selecting logarithmic ratio of number of documents to the number of documents that contains this domain. This in turn only retrieves domain related information present in the document and therefore reduces the computational time using DF-AHR by 18% compared to QP-WCC and 30% compared to CQA-KB.

Finally we address the third goal of the experiments with respect to computational complexity. Computational complexity is a measure of the amount of working storage required to perform Analytic Hierarchy Tree Construction algorithm evaluated from the equation (10). Sample calculation is as given below.

**Sample calculation**

- **QP-WCC**: With the number of user requests being '10', the memory consumed while performing Analytic Hierarchy Tree Construction between domains is '31KB', then the computational complexity is obtained as given below.
  $CC = 10 * 31KB = 310KB$
- **CQA-KB**: With the number of user requests being '10', the memory consumed while performing Analytic Hierarchy Tree Construction between domains is '45KB', then the computational complexity is obtained as given below.
  $CC = 10 * 45KB = 450KB$
- **Proposed DF-AHR**: With the number of user requests being '10', the memory consumed while performing Analytic Hierarchy Tree Construction between domains is '24KB', then the computational complexity is obtained as given below.
  $CC = 10 * 24KB = 240KB$

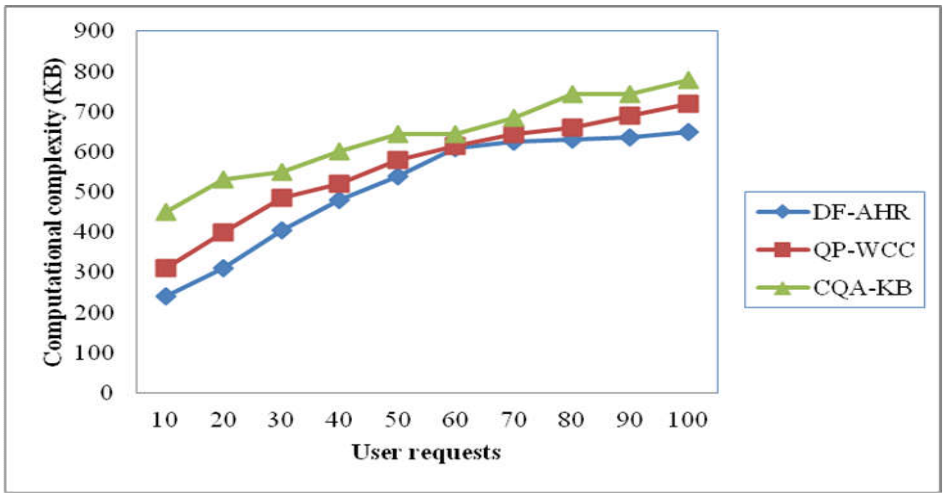


**Figure 6 Measure of computational complexity**

For all cases as shown in figure 6, the computational complexity is increasing with the user requests considered from different users. The targeting results of computational complexity using DF-AHR method is compared with two state-of-the-art methods [1], [2] in figure 6 is presented for visual comparison. Our method differs from the QP-ECC [1] and CAQ-KB [2] in that we have incorporated semantic distance and semantic relation between domains for text web content mining.

Here the weight of the hierarchical structures is considered while constructing the tree using positions, obtained by evaluating the semantic distance that in turn selects the domain frequent documents and therefore reduces the dimensionality factor. Besides, dominant domains within a document and presence of similar domains in other documents are extracted by applying the semantic relation. With the resultant values obtained through semantic distance and semantic relation, further reduces the computational complexity during textual web content mining. Therefore the computational complexity for

textual web content mining using DF-AHR method is reduced by 10% compared to QP-ECC [1] and 21% compared to CAQ-KB [2] respectively.

## IV. Conclusion

Textual web content mining is a complex task due to the explosive growth of World Wide Web and increasing amount of contextual information. Therefore, there arises an urgent need to provide method for this contextual information and content credibility and to propose an efficient method to provide high rate of accuracy for textual web content mining. In this article we provide Domain Frequent-Analytic Hierarchy Tree (DF-AHR) method that can be employed as a textual web content mining method. Initially, pre-processing was performed using Domain Frequent Pre-processing model. With the pre-processed content, Domain Sense Disambiguation was applied to Analytic Hierarchy Tree construction. This in turn searches the presence of domains in documents based on hierarchy model. Finally, semantic distance and semantic relations between domains were evolved to mine textual web content. Through the experiments using real traces, we observed that our textual web content mining method reduced computational time and computational complexity compared to the existing methods.

## References

*[1] Michal Kakol, Radoslaw Nielek, Adam Wierzbicki, "Understanding and predicting Web content credibility using the Content Credibility Corpus", Information Processing and Management, Elsevier, May 2017*

*[2] Alejandro Figueroa, Günter Neumann, "Context aware semantic classification of search queries for browsing community question–answering archives", Knowledge-Based Systems, Elsevier, Jan 2016*

*[3] Dimitrios Kotzias, Moshe Lichman, and Padhraic Smyth, "Predicting Consumption Patterns with Repeated and Novel Events", IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 30, NO. 8, AUGUST 2018*

*[4] Yifan Chen, Xiang Zhao, Xuemin Lin, Yang Wang, Deke Guo," Efficient Mining of Frequent Patterns on Uncertain Graphs", IEEE Transactions on Knowledge and Data Engineering , Apr 2018*

*[5] Eleonora Ciceri, Piero Fraternali, Davide Martinenghi and Marco Tagliasacchi, "Crowdsourcing for Top-K Query Processing over Uncertain Data", IEEE Transactions on Knowledge and Data Engineering (Volume: 28, Issue: 1, Jan. 1 2016)*

*[6] Vincent S. Tseng, Senior Member, Cheng-Wei Wu, Philippe Fournier-Viger, Philip S. Yu," Efficient Algorithms for Mining Top-K High Utility Itemsets", IEEE Transactions on Knowledge and Data Engineering (Volume: 28, Issue: 1, Jan. 1 2016)*

*[7] Abdul Mohsen Almalawi, Zahir Tari, Adil Fahad, Muhammad Aamir Cheema, Ibrahim Khalil, "kNNVWC: An Efficient k-Nearest Neighbours Approach based on Various-Widths Clustering", IEEE Transactions on Knowledge and Data Engineering, Jun 2016*

*[8] Shuyao Qi, Dingming Wu, and Nikos Mamoulis, "Location Aware Keyword Query Suggestion Based on Document Proximity", IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL 13, NO 9, SEPTEMBER 2014*

*[9] Ouz Mustapaaa, Dilek Karahocaa, Adem Karahocaa, Ahmet Yücela, Huseyin Uzunboylua, "Implementation of Semantic Web Mining on E-Learning", Elsevier, Jan 2010*

*[10] Feifei Wang, Xiaohan Wang, Siluo Yang, "Mining author relationship in scholarly networks based on tripartite citation analysis", PLOS ONE | https://doi.org/10.1371/journal.pone.0187653 November 8, 2017*

*[11] Karthik Krishnamurthi, Vijayapal Reddy Panuganti, Vishnu Vardhan Bulusu," Including category information as supplements in latent semantic analysis of Hindi documents", Int. J. Computational Science and Engineering, Vol. 15, Nos. 1/2, 2017*

*[12] Sunny Sharma and Vijay Rana, "Web Personalization through Semantic Annotation System", Advances in Computational Sciences and Technology, Research India Publications, Volume 10, Number 6 (2017) pp. 1683-1690*

*[13] Yaqing Shi, Meijuan Wang, Zhenghong Qiao,Lei Mao, "Effect of Semantic Web technologies on Distance Education", Advanced in Control Engineering and Information Science, Elsevier, May 2011*

*[14] Roberta Akemi Sinoara, João Antunes and Solange Oliveira Rezende, "Text mining and semantics: a systematic mapping study", Journal of the Brazilian Computer Society, Springer, May 2017*

*[15] Halil Kilicoglu, Graciela Rosemblat, Thomas C. Rindflesch, "Assigning factuality values to semantic relations extracted from biomedical research literature", PLOS ONE | https://doi.org/10.1371/journal.pone.0179926 July 5, 2017*
*[16] Xiaomei Zou, Jing Yang, Jianpei Zhang, "Microblog sentiment analysis using social and topic context", PLOS ONE | https://doi.org/10.1371/journal.pone.0191163 February 2, 2018*
*[17] Andreas Hotho a, Robert Jäschke b, Kristina Lerman, "Mining Social Semantics on the Social Web", Semantic Web, Semantic Web – Interoperability, Usability, Applicability on IOS Press Journal, May 2017*
*[18] Andrzej Siemiński, "Fast algorithm for assessing semantic similarity of Texts", Int. J. Intelligent Information and Database Systems, Vol. 6, No. 5, 2012*
*[19] Jianhua Yang, Harsimrat Singh, Evor L. Hines, Friederike Schlaghecken, Daciana D. Iliescu , Mark S. Leeson, Nigel G. Stocks, "Channel selection and classification of electroencephalogram signals: An artificial neural network and genetic algorithm-based approach", Artificial Intelligence in Medicine, Elsevier, Jan 2012*
*[20] Siaw Ling Lo*, Raymond Chiong, David Cornforth, "Using Support Vector Machine Ensembles for Target Audience Classification on Twitter", PLOS ONE | DOI:10.1371/journal.pone.0122855 April 13, 2015*