ISSN: 2229-6959 (ONLINE)
DOI: 10.21917/ijsc.2017.0198

RELIABLE COGNITIVE DIMENSIONAL DOCUMENT RANKING BY WEIGHTED STANDARD CAUCHY DISTRIBUTION

S. Florence Vijila[1] and K. Nirmala[2]
[1]*Deparment of Computer Science, Manonmaniam Sundaranar University, India*
[2]*Department of Computer Science, Quaid-e-Millath Government College for Women, India*

Abstract

Categorization of cognitively uniform and consistent documents such as University question papers are in demand by e-learners. Literature indicates that Standard Cauchy distribution and the derived values are extensively used for checking uniformity and consistency of documents. The paper attempts to apply this technique for categorizing question papers according to four selective cognitive dimensions. For this purpose cognitive dimensional keyword sets of these four categories (also termed as portrayal concepts) are assumed and an automatic procedure is developed to quantify these dimensions in question papers. The categorization is relatively accurate when checked with manual methods. Hence simple and well established term frequency / inverse document frequency 'tf/ IDF' technique is considered for automating the categorization process. After the documents categorization, standard Cauchy formula is applied to rank order the documents that have the least differences among Cauchy value, (according to Cauchy theorem) so as obtain consistent and uniform documents in an order or ranked. For the purpose of experiments and social survey, seven question papers (documents) have been designed with various consistencies. To validate this proposed technique social survey is administered on selective samples of e-learners of Tamil Nadu, India. Results are encouraging and conclusions drawn out of the experiments will be useful to researchers of concept mining and categorizing documents according to concepts. Findings have also contributed utility value to e-learning system designers.

Keywords:

Standard Cauchy distribution; Document Categorization; Concept extraction; Cognitive Dimensions; Term frequencies.

1. INTRODUCTION AND BACKGROUND

Among various Cauchy probability distribution techniques, the Standard Cauchy method is relatively simpler in expression and easy to compute, beside, the derived values are also extensively used for checking uniformity and consistency of various applications [14]. Most of the University students, particularly of Computer Science and Applications, usually search for question papers from question banks that are available in respective web sites of many Universities. These students, in reality, look mostly for question papers which are uniform and consistent in their presentation (or patterns of questions) for better examination performance point of view, more specifically from the cognitive dimensional angles. Generally each University follows its own pattern and style, but consistently maintains uniformity. The cognitive dimensions here refer to dimensions (or categories) that groups the contents into four parts for answering, namely, simple facts, termed 'factual'; complex questions that expect short answers, termed as 'critical/conceptual'; questions that demand for elaborative long answers termed as 'explaining'; and problem solving or programme coding (all within the case study of Computer Science and Applications). It is presumed, what the students look for is consistency in representations of these four cognitive dimensions. In fact even though many of these Universities when they try to follow a particular pattern as far as possible to maintain uniformity in their own traditions, deviations do occur. This is true particularly in maintaining cognitive dimensions in the same ratios that may not be possible always.

To test the documents for uniformity, for the sake of automating which applies concept mining technique, appropriate concept words are required for each cognitive dimension, so that categorization could be done either through the simple *tf/IDF* (Term Frequency/Inverse Document Frequency) technique or through more complex probability algorithms, such as Naive Bayes algorithm [5]. When such documents are numerous in number, how to rank order those for the sake of acceptability to user students who look for question papers that are consistent with individual University question patterns? Term frequencies and Naive Baye's conditional probability techniques for efficiently quantifying cognitive dimensional values of documents have been proposed and tested to an acceptable degree [4]. Instead of the complex probability theories, Cauchy dense functions have been successfully used for proper consistency checking in distributions [2] and Cauchy distribution method even provides rough approximates, but obtains better solutions [3]. Term frequency can be quickly estimated, while standard Cauchy distribution can provide clue for consistency in these values. Both these extraction and consistency testing techniques in an integrated scenario has not to be tried out or seen in literature, particularly for extracting assessment tools that are relatively consistent with each other, of e-learning environment.

With this issue in the above background, the paper presents experimental results with strong literature support on the research methodologies, through seven documents (case studies) that are cross validated with social survey which was administered with user respondents of Tamil Nadu, India. Conclusions have been drawn from these experiments, which will yield both research values as well as utility values that would be useful to concept extraction researchers and e-learning user-evaluation designers.

2. LITERATURE SURVEY

Snippets or small fragments of textual documents are as good as clusters of full documents for reliability in searching of documents [6]. Short snippets consist of words of domain specific as well as portrayal specific (or cognitive dimension) such as: full explanation of the domain content or numerical example/worked out problems/programs or interrogative sentences etc. It is observed by the authors that within a few seconds, clustering was able to achieve up to one thousand snippets. For an example, in a question paper, snippets mostly would have interrogative words, like 'what', 'how', 'why' etc. This phenomena encourage using some proven techniques in identifying and clustering interrogative documents according percentiles of pedagogical types existing in a set of such documents. The *tf/IDF* weighting (term frequency-inverse document frequency) is a numerical statistic techniques which reflects how important a word is to a document in a collection or corpus [1]. By convention, the *tf/IDF* value increases proportionally to the number of times a word appears in a document, but is offset by the frequency of the word in the corpus, which helps to control for the fact that some words are generally more common than others. This technique (even though simple) has been tried out in applying concept keywords and suitable e-documents have been extracted and provided to e-learners according to their styles and interests, thus enhancing their learning performances [7]. For simple and straight forward concept extraction application (*like* cognitive dimension extraction), concept keywords have been recommended to use in computing term frequency (*tf*) & Inverse Document Frequency (*IDF*), for efficiently extracting documents [8]. Online tests are conducted to assess learner's achievements and to maintain such online items, it is a big task [9]. The knowledge types (cognitive dimensions) of such items can be differentiated and classified as 'factual', 'conceptual' and 'procedural' and their cognitive levels might include conceptual words of learning abilities like simple definition, deeper explanations, conceptual elaboration of subject matter etc., according to the authors. Standard Cauchy distribution has helped in determining consistent performances of Grid computation, specifically in load balancing, by using trust values [14]. Thus, with the support of the above literature, Cognitive dimensions of documents can be clustered (or grouped with quantified results) and standard Cauchy values could be computed from those results, for determining (or rank ordering the documents) the acceptance levels of the ranked documents by users, who were looking (searching) for similar contents in styles and presentations (consistency).

3. RESEARCH METHODOLOGIES

University students' assessment instruments or question papers are arranged in parts of short answers, long and elaborative answers and the questions on numerical problems or program coding are also formed in another part of the question paper. Even though this is not generally the case with real world pure textual documents, for the sake for experimental objectives, it was necessary to explore selecting such specific nature of University question papers (documents) for the proposed experiments. In other words, a student who answers these question papers comprehends the parts of the question paper in the following four cognitive dimensions namely, 'Factual (*F*)', 'Explanatory/elaborative (*E*)', 'Problem solving (*P*)' and 'Conceptual/critical thinking (*C*)'. Such split up parts could be quantified using pre-specified (particular) cognitive keywords, such as Bloom's taxonomy [11]. The research methodology would thus apply pre-defined concept keywords for the chosen four cognitive dimensions for extracting (or quantifying) the documents. The Table.1 presents a few sample chosen concept keywords for the four cognitive dimensions.

Probability application is valid where association rules prevail [12]. The dependability of association rules with any probability classifier has been proved by research on text classification for data mining. But this method ignores negative example for any specific class, the accuracy may fall in some cases. Besides, probability technique consumes more computational time than simple *tf/IDF* technique. The negative representation in our selected categories may also be minimal, as the cognitive keywords generally do not repeat in the selected four categories. Therefore simple *tf/IDF* technique is suggested.

The Table.1 presents the cognitive dimensional and the pre defined concept keywords (assumed by the authors, in addition to using those available from literature [11], for the chosen Computer Science subject area, namely 'Programming in C++ and Data Structures'. The use of taxonomy of concept words for defining learning objectives in instructional documents has been well established [10].

Table.1. Cognitive Dimensional Categories and Samples of Extracting Keywords

ID	**Cognitive Dimension**	**Input**	**Sample Keywords (Drawn from interrogative documents)**
F	Factual	Information of mere facts; Short answers	*list, what, note, define, tell, name, locate, identify, distinguish, acquire, write, underline, relate, state, recall, select, repeat, recognize, reproduce, measure, memorize.*
E	Procedural	Elaborative procedures; Explanatory processes; Algorithms	*demonstrate, explain, how, write, detail, summarize, illustrate, interpret, contrast, predict, associate, distinguish, identify, show, label, collect, experiment, classify, stress, discuss, select, compare, prepare, change, rephrase, differentiate, draw, estimate, fill in, choose, operate, perform, organize.*
P	Problematic	Heuristics; Methods; Techniques	*apply, write, code, calculate, illustrate, solve, make use of, predict, how, construct, assess, practice, restructure, find.*
C	Conceptual	Concepts; Schemas; Models	*analyze, resolve, justify, infer, combine, integrate, why, plan, create, design, generalize, assess, decide, rank, grade, test, recommend, select, explain, judge, contrast, survey, examine, differentiate, investigate, compose, invent, improve, imagine, hypothesize, prove, predict, evaluate, rate.*

Documents (students' assessment question papers) are located with the above cognitive dimensional concept words of Table.1. Factual/short answers: '*F*'; Elaborative/Explanatory/ Procedural: '*E*'; Problematic/Solutions/worked out examples '*P*'; Conceptual/synthesizing 'C' are the parts of the question papers. Standard Cauchy formula is used for viewing the distribution of the cognitive dimension values in seven chosen documents for case studies. For validating the proposed technique, social survey techniques has been proposed, as validation of similarity or pattern of documents is more a matter of user dependent subjective in nature.

3.1 COMPUTATIONAL PARAMETERS

The parameters considered for the study are Term Frequencies and Standard Cauchy distributions. Both these parameters are computed using the following procedures.

- For term frequency, the inverse document frequency is computed as:

$$IDF = \log(N/K)$$

For every cognitive dimension weighted coefficients parameter is computed = $tf. \left|\log(N/K)\right|$

where, K - Number of occurrences of terms in all the documents considered. i = No. of CD (Cognitive Dimensional) term occurring in one document; n = No. of most frequently occurred CD in one document; $tf = i/n$ is the normalized term frequency and N = Total number of documents considered

- The standard Cauchy dense function is computed for every functional variable $f(x;0,1) = 1/(\pi(1+x^2))$. The normality is computed for every distribution of 'x' and $f(x)$.

4. EXPERIMENTAL SETUP

The documents selected for the experiments are based on University question papers (B.Sc. Computer Science) of Madras University, Chennai, India. Four documents were extracted from University question papers (the University of Madras, Chennai's public domain), namely $D(i)$, i=1,4 and the rest of the documents, $D(i)$, i= 5,7 were purportedly edited by the authors, so as to represent inconsistencies on three question papers (documents of the case studies) for comparative studies.

The subject belonging to these textual documents (question papers) is 'Programming in C++ and Data Structures' of the B.Sc degree programme of the University. Information about the documents for case studies is presented in Table.2. Stemming the unwanted parts of the words and removal of unwanted stop words have been done by authors algorithm coded in Java.

Table.2. Documents for Experimental Setup and Cognitive Dimensional Values

Doc. Id '*i*'	**Source (University Reference)**	**No. of words after stemmed & stopped**	**Cognitive Dimensional Values in %**				**Remarks**
			F	***E***	***P***	***C***	
1	PC3A, Nov. 2008	218	48	22	26	4	Except for one year, the rest are consistent and more or less uniform in each cognitive dimension.
3	SAZ3A, Nov. 2009	116	30	42	21	7	
5	SAZ3A, Nov. 2010	174	47	19	11	11	
6	SAZ3A, Nov. 2011	202	52	26	2	2	
2	Researcher	432	27	28	36	9	Purportedly designed to be abnormal, for testing with experiments.
4	Researcher	339	31	33	29	7	
7	Researcher	306	18	43	36	3	
-	Long term rounded average	386	50	25	20	5	

Using the *tf*/*IDF* value that increases proportionally to the frequency of word occurring in a document, but is offset by the frequency of the same word in the whole group of documents, which indicates the fact that some words are generally more common than others. This was tested by the authors program and results published earlier [4].

Standard Cauchy distribution is a continuous probability distribution represented through a graph (for want of space, only the computed values are presented in Table.3 and no graphical representation is shown) which is used to view the levels of acceptance of any functional distribution. It is found from the literature that this standard Cauchy distribution has been successfully adopted in many Computer Science areas [14]. Instead of complex probability values, the Cauchy dense function values are themselves used for the distribution for the study of consistency in the distribution. In statistical analysis, standard deviation and mean are not determined for Cauchy distributions [2]. Literature strongly demonstrates that the predictions are performed in acceptable levels, using Cauchy distributions. Load balancing in Grid computing is an example to support this claim, as the reliability can be demonstrated using Cauchy and normal distribution methods [14]. They even provide rough approximate but better solutions in many other applications too [3]. The Cauchy standard distribution computation is briefly explained below.

The Cauchy standard distribution formula is provided in Eq.(1).

$$f(x;0,1) = 1/(\pi(1+x^2)) \tag{1}$$

where, x represents the ratio of a particular cognitive dimension to the total cognitive capacity of that dimension in the selected document (or in other words resulted in %). If x either becomes 0 or 1 (as extreme cases), the distribution would become $1/\pi$ and $1/2\pi$, respectively. The proposed experiment aims at determining only the standard Cauchy values that are distributed (the presence of) in cognitive dimensions of 7 chosen documents and it is not a normal distribution. The proposed experiment is aimed at determining the performance based only on the distribution of Cauchy function in Eq.(1). The Table.2 presents the distribution of standard Cauchy values for the chosen seven documents.

5. RESULTS AND DISCUSSIONS

The novelty of the research is the procedure described in arriving at the weight order ranking of the documents (seven case study documents). The proposed Standard Cauchy Weight order sorting algorithm (SCW): The algorithm computes the weight order in a linear fashion that considers the input values of 'x' which is the ratio (percentage) of each quantity of the four cognitive dimensions (‘*F*’, ‘*E*’, ‘*P*’ and ‘*C*’) of each document, totalling 100% or 1.0 per document (each document’s (i) each dimension is considered for the computation). Seven documents are considered and presented for each dimension in the second column of Table.3.

The Cauchy value is computed for each document and presented in column (3) of the table. The document id is shown in brackets in column (4) along with the Cauchy values. These values are sorted according to descending order (column 4) of respective document id. For each dimension, the sorted values are then paired up with each adjacent document and the difference (deviation) between the adjacent is documented (column 5). The lowest valued pair is identified and arranged accordingly from lower to higher values (shown in column 6). The first repeated lower document is identified and presented in the last column of Table.3. These documents are ready to be presented in serial order (ranked order) to the user as each one in series represent more consistent in order, as per Cauchy’s theorem.

The procedure adopted in the proposed algorithm SCW, is actually the differences between the Cauchy values of adjacent sorted values for determining the slope (or shallowness) of the line (distribution) which has been computed by the algorithm. Similar approach has been demonstrated in determining trust worthiness of grid computing [14].

Table.3. Distribution of Cauchy Values and SCW of Experimental Documents

Cognitive Dimension	**$x(i)$ i = (1) to (7)**	**$\frac{1}{\pi\left(1+x(i)^2\right)}$ i = (1) to (7)**	**Sorted Cauchy Values CV(i)**	**Deviation between adjacent**	**Lowest to Highest Pairs**	**SCW (Lower presence in the sequence)**
F(*i*)	0.48;	0.2587;	(2):0.2967;	(2-3): .0047;	(5-6);	(6);
	0.27;	0.2967;	(3):0.2920;	(3-4):.0016;	(3-4);	(5);
	0.30;	0.2920;	(4):0.2904;	(4-5):.0317;	(6-1);	(3);
	0.31;	0.2904;	(5):0.2607;	(5-6):.0001;	(2-3);	(4);
	0.47;	0.2607;	(6):0.2606;	(6-1): .0019;	(4-5);	(1);
	0.52;	0.2606;	(1):0.2587;	(1-7):.0646;	(1-7);	(2);
	0.18	0.1941	(7):0.1941	(7-2): .1026.	(7-2).	(7).
E(*i*)	0.22;	0.3036;	(5):0.3072;	(5-1): .0036;	(3-7);	(1);
	0.28;	0.2952;	(1):0.3036;	(1-6): .0054;	(6-2);	(6);
	0.42;	0.2706;	(6):0.2982;	(6-2): .0030;	(5-1);	(2);
	0.33;	0.2871;	(2):0.2952;	(2-4):.0081;	(1-6);	(3);
	0.19;	0.3072;	(4):0.2871;	(4-3):.0165;	(2-4);	(4);
	0.26;	0.2982;	(3):0.2706;	(3-7):.0020;	(4-3);	(7);
	0.43	0.2686	(7):0.2686	(7-5):.0386	(7-5)	(5)
P(*i*)	0.26;	0.2982;	(6):0.3061;	(6-3): .0012;	(2-7);	(3);
	0.36;	0.2818;	(3):0.3049;	(3-5): .0026;	(6-3);	(5);
	0.21;	0.3049;	(5):0.3023;	(5-1): .0041;	(3-5);	(1);
	0.29;	0.2936;	(1):0.2982;	(1-4): .0046;	(5-1);	(4);
	0.23;	0.3023;	(4):0.2936;	(4-2): .0118;	(1-4);	(2);
	0.20;	0.3061;	(2):0.2818;	(2-7): .0000;	(4-2);	(7);
	0.36	0.2818	(7):0.2818	(7-6): .0243	(7-6)	(6)
C(*i*)	0.04;	0.3178;	(6):0.3182;	(6-7):.0002;	(3-4);	(7);
	0.09;	0.3158;	(7):0.3180;	(7-1):.0002;	(6-7);	(4);
	0.07;	0.3168;	(1):0.3178;	(1-3):.0033;	(7-1):	(2);
	0.07;	0.3168;	(3):0.3145;	(3-4):.0000;	(4-2);	(1);
	0.11;	0.3145;	(4):0.3145;	(4-2):.0013;	(2-5);	(3);
	0.02;	0.3182;	(2):0.3158;	(2-5):.0013;	(1-3);	(6);
	0.03	0.3180	(5):0.3145	(5-6):.0037	(5-6)	(5)

6. VALIDATION

As the acceptance of rank ordered documents (according to similarity or consistence) is a subjective matter of users' view point, and hence social survey is preferred for validating the proposed application of Cauchy's standard distribution and also to test the proposed SWC algorithm. The details of respondents, sampling technique and statistical procedures adopted for validating the proposed technique are presented below.

Questionnaire: (1) Consistent with conventional practice followed in the assessments (for reliability); (2) Representation of requested cognitive dimension without much deviation from previous assessment tools (for validity and usefulness). Scale: (1) Yes, high; (2) More or less; (3) No, low. Sample: No. of respondents, 53; Demography: B.Sc students of second and final years of an affiliated college in Chennai (where one of the authors is employed as an Asst. Professor. For well matched composition of the sample like male/female/urban/rural/ talented/average students, appropriate demography has been selected. Sampling technique: Purposive sampling method [13]. Computations are performed by well known s/w package SPSS 17.0 (Statistical Program for Social Science, Ver. 17.0). Validity of the question paper was achieved through opinions drawn from expert committee and the reliability of feedbacks was tested using Chronbach's alpha (pilot study results, modification of questionnaire and reliability analytical results are not presented for want of space). Chronbach's alpha was found to be greater than 0.7 which may be accepted. Only two sample results are shown below to save space.

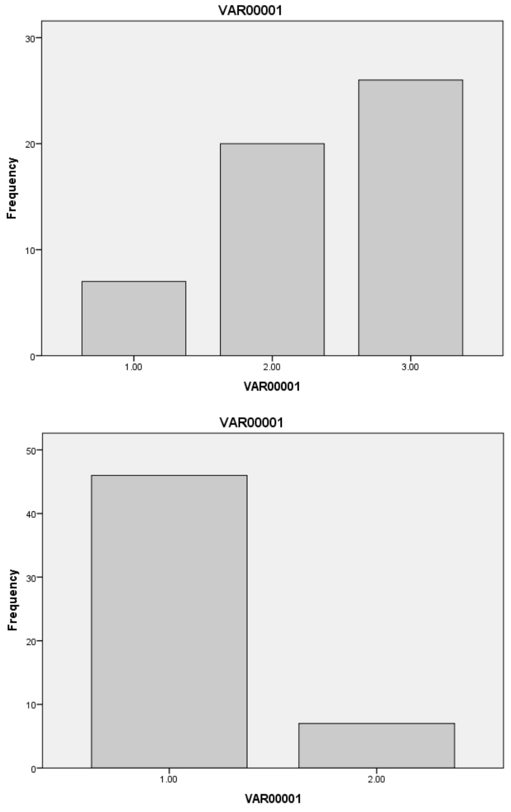

Fig.1. Responses on Document 1 (LHS) and Document 6 (RHS) on 'Factual' Cognitive Dimension for first variable

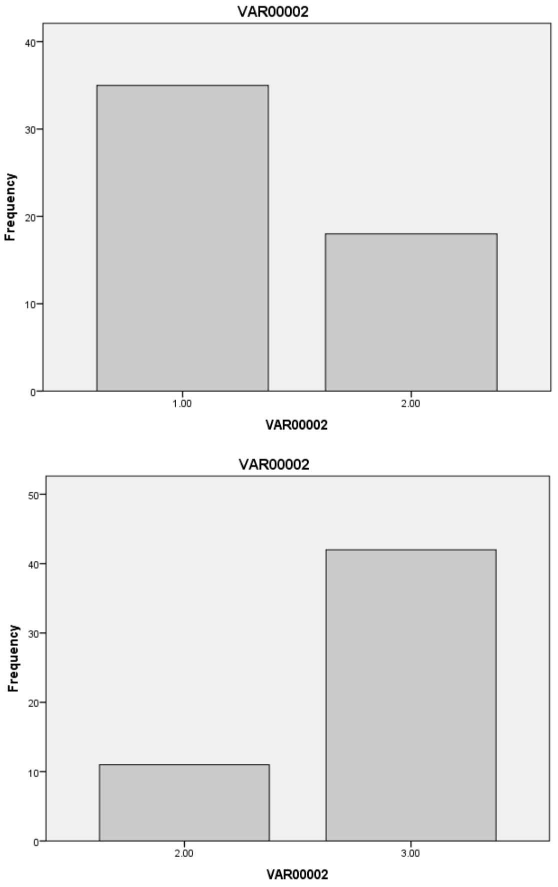

Fig.2. Responses on Document 1 (LHS) and Document 5 (RHS) on 'Procedural' Cognitive Dimension for second variable

7. OBSERVATION

For the purpose of validating the proposed technique, the normal consistent and average representative values of '*F*', '*E*', '*P*' and '*C*' have been considered as about 50%, 25%, 20% and 5% respectively. These values were determined manually by averaging out the cognitive dimensional questions of 10 years previous question papers. Sample feedbacks received from respondents after viewing documents 1 and 5. The results on question 1 (variable VAR00001) on consistency of cognitive dimension '*F*' are shown in Fig.1. The LHS is for document (1) and the RHS is for document (6). The feedbacks shown are obviously negative for document 1 and highly positive for document 6 on the '*F*' cognitive dimension. The last column of Table.3 of '*F*' clearly matches with these results for document (1), and document (6). Document (6) superseding document (1) is shown in the rank order of the table.

The results on question 2 (variable VAR00002) on consistency of cognitive dimension 'E' are shown in Fig.2. The LHS is for document 1 and the RHS is for document 5. The feedbacks shown are obviously positive for document 1 and negative for document 5 on the '*E*' cognitive dimension. The last column of Table.3 of '*E*' clearly matches with these results for document (1), and document (5). Document (6) superseding document (1) is shown in the rank order of the table. These results clearly demonstrate the validity of the proposed technique.

8. CONCLUSIONS

The surveyed results have proved clearly that standard Cauchy distribution values will aid in rank ordering the documents according consistency and uniformity of cognitive dimensions (concepts) of the contents. It is demonstrated clearly through the proposed methodology that the deviations in slope values of the standard Cauchy curve, can help in rank ordering the nodes according to consistency. It is concluded that while simple *tf/IDF* values are used to extract desired documents, standard Cauchy distribution can be adopted for rank ordering the documents as per desired norms and rules, in an efficient reliable manner.

REFERENCES

[1] G. Salton and C. Buckley, "Term-Weighting Approaches in Automatic Text Retrieval", *Journal of Information Processing and Management*, Vol. 24, No. 5, pp. 513-523, 1988.

[2] M. Jelasity, Alberto Montresor and Ozalp Babaoglu, "Gossip-based Aggregation in Large Dynamic Networks", *ACM Transactions on Computer Systems*, Vol. 23, No. 3, pp. 219-252, 2005.

[3] S. Rajarajeswari, "A Novel Exhaustive Criterion Based Load Balancing Algorithm for e-Learning Platform by Data Grid Technologies", *International Journal of Advanced Networking and Applications*, Vol. 4, No. 6, pp. 1786-1792, 2013.

[4] S. Florence Vijila and K. Nirmala, "Quantification of Portrayal Concepts using tf-IDF Weighting", *International Journal of Information Sciences and Techniques*, Vol. 3, No. 5, pp. 1-6, 2013.

[5] S. Florence Vijila and K. Nirmala, "Structural Effectiveness for Concept Extraction through Conditional Probability", *Advances in Natural and Applied Sciences*, Vol. 9, No. 7, pp. 39-47, 2015.

[6] Oren Zamir and Oren Etzioni, "Web Document Clustering: A Feasibility Demonstration", *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 46-54, 1998.

[7] T.C. Tseng, H.C. Chu, G.J. Hwang and C.C. Tsai, "Development of an Adaptive Learning System with Two Sources of Personalization Information", *International Journal of Computers and Education*, Vol. 51, No. 2, pp. 776-789, 2008.

[8] Masaru Ohba and Katsuhiko Gondow, "Toward Mining 'Concept Keywords' from Identifiers in Large Software Projects", *Proceedings of International Workshop on Mining Software Repositories*, pp. 1-5, 2005.

[9] Ming-Hsiung Ying and Heng-Li Yang, "Computer Aided Generation of Item Banks based on Ontology and Bloom's Taxonomy", *Proceedings of International Conference on Web-Based Learning*, pp. 157-166, 2008.

[10] Robert M Gagne, "*The Conditions of Learning and Theory of Instructions*", 4th Edition, Wadsworth Publishing Co Inc, 1985.

[11] M. Suriakala and T.G. Sambanthan, "Problem Centric Objectives for Conflicting Technical Courses", *The Indian Journal of Technical Education*, Vol. 31, No. 2, pp. 87-90, 2008.

[12] S.M. Kamruzzaman, Farhana Haider and Ahmed Ryadh Hasan, "Text Classification using Association Rule with a Hybrid Concept of Naive Bayes Classifier and Genetic Algorithm", *Proceedings of 7th International Conference on Computer and Information Technology*, pp. 682-687, 2004.

[13] B.A.V. Sharma, "*Research Methods in Social Sciences*", Sultan Chand and Sons Publications, 1988,

[14] S. 14, G. Manoharan and K. Nirmala, "Trust Worthiness on Load Balance in Grids using Standard Cauchy Distribution", *Research Journal of Applied Sciences, Engineering and Technology*, Vol. 8, No. 16, pp. 1833-1837, 2014.