Journal of Advance Computational Research
Vol. 3 No. 2 (July-December, 2018)

A Fuzzy Document Ranking System using Confusion Matrices

Arpana Rawal[1], Manoj Kumar Kowar[2] and Sanjay Sharma[3]
[1]Reader, Department of Computer Science
[2]Dept. of Electronics & Telecommunications
[3]Professor, Department of Mathematics
Bhilai Institute of Technology, Bhilai House, Durg (C.G.) – 491001 India
E-mail: arpana_rawal@rediffmail.com , kowar_bit@Rediffmail.com , ssharma_bit@yahoo.co.in

RECEIVED: January 06,2017. Revised April 17, 2017

Abstract: *Till date, voluminous text data have been put to knowledge engineering based on machine learning approaches by the text miners. The context oriented information retrieval has always been based on some or the other explicit ontologies viz. Hierarchical thesaurus or subject oriented control dictionaries or vocabularies. Unlike the above ,the authors in this paper, emphasize that the key concepts selected and aggregated, contributing to the background knowledge are extracted from self-acquired Ontologies. The proposal incorporates a system (tool) to rank text documents available in machine-readable format by analyzing them upon softcopies of the syllabus content, through congenial content filtering techniques. This in turn, is implemented with a hybrid machine learning approach, encompassing Naïve Bayesian classifier to predict syllabus-relevant text material upon which suitable fuzzy ordering technique is applied which ranks over the relevance degree of extracted text. The above task is presumed to be helpful in organizing the prescribed text material into a finite sequence of topics of one's own interest or sorting them in order of their utility value as courseware material.*

Keywords : *Self-acquired back ground knowledge, Bootstrapping Semantics, Naïve Bayesian Classifier, Confusion matrices, Fuzzy Ordering.*

1. INTRODUCTION

Current research on knowledge management techniques show that the knowledge miners largely depend on the explicitly available Ontologies, specially from domain-specific subject dictionaries, controlled vocabularies, thesauri, etc., for designing an effective text retrieval environment. Unlike the above, the authors here, emphasize on preferring the Ontologies in Implicit form, if suitably extracted from the given text documents, upon which the organizing and retrieval tasks are to be carried through. This could be envisioned towards the goal of extracting the best out of the text matter, put for processing in any subject domain.

In the present communication, the authors introduce the fuzzy-ordering techniques to rank the already identified the text in order of their topical relevance. Some of the related works and document relevance issues are highlighted in section 2. Section 3 explains the exemplary domain of case study in which the syllabus strings of a particular course enact as search topics to explore. Section 4 proposes the steps of content filtering from the vast text corpora by unfolding text semantics that yield the training and test documents. Section 5 describes, how naïve Bayesian classification predicts the document beds against the observed documents, taking classification parameter as document-category of relevance. Section 6 and 7 justifies the selecting of classification hits and classification miss values of the confusion matrix as the comparison membership values to begin evaluating the document ranks by formulating relativity functions [11].

2. THE STATE-OF-ART

As an illustrative case study in technical realm, if a humane approach is taken up to draw out appropriate learning material out of a list of prescribed text, there may arise situations, when the teaching instructor finds himself in utter confusion in precisely deciding the most appropriate to-be-taught topics and their chronological order. This might happen in cases of crudely designed syllabus, may be newly introduced in a courseware, where strict topic-sequencing and co-relation among prescribed syllabus terms are not cared in the drafting step. This can only be determined if in-depth exhaustive reading process is performed on the prescribed learning material.

Atlam, *et al.* (2002, 2003) have introduced the use of five levels of Field Association terms confined to the specific domain for recognizing content similarities in

large heterogeneous texts [3] [4]. Inspired by the idea, the paper focuses on finding term-to-term co-occurrences, but not limited to finite levels, instead continuing till redundancies are obtained. Another measuring metric, page-vicinity levels too are introduced that help in finding category levels of document relevance. Another significant work by Mayr, *et al.* (2007) can be viewed, where document relevance is achieved by creating meta-data repositories to begin the conceptual searches within document titles, abstracts, headings, sub-headings and using author profiles on the extreme [10]. This conformed the idea of formulating implicit ontology as true search domain.

As a matter of fact, the authors have already exploited the document classification performance metric i.e. confusion matrix in organizing so obtained relevant documents, so that a hierarchical topic and sub-topic sequencing can be looked into, while teaching in that courseware domain [9]. The paper thus aims at another text mining task where page-ranges identified from the courseware-corpus get ranked in order of their topic-relevancies.

3. SCENARIO AND USER REQUIREMENTS

To begin with machine learning for the above problem, the set of search terms lying in the mentioned syllabus, now can be equivalently thought for being comfortably available either at front-page 'table-of-contents' or at the 'back-of-the-book-index'. These could be presumed to be implicitly available as portions of available text in soft format say in form of scanned 'read-only' .pdf documents and e-books for a given set of topics as focused terms.

Given the syllabus snapshot as shown in fig. 1, with a prescribed text learning material, the book of technical viz. engineering domain, entitled "*Neural Networks : Algorithms, Applications and Programming techniques*" authored by *James Freeman and David M. Strapetus*, the authors hereby put forth an inspired proposal for formulating a concept space of a more concise dimensionality criterion, keeping two issues in mind – not all the keywords and term to term co-occurrences shall be found in the vicinity of one chapter [1] [2]. Further, it is quite evident that a syllabus keyword / concept may rarely strike a string-match into all the chapter content of the prescribed material.

The governing parameters for such a concept space begins with the extraction of :

- Noun / verb phrases parsed from natural language syllabus text-strings – *the driving lexicon*.
- Chapter header nodes, section / subsection header nodes, paragraph header nodes – *the chapter header tree*, formed as an extended portion of implicitly available front index.

The formulation of Chapter Header Tree along with simultaneous tagging up of the paragraphs can be perceived as an already accomplished logic in the step towards extracting targeted paragraphs containing the relevant content from the huge analyzing text corpus available [7].

Table 1
Syllabus Snapshot

Unit-I Elementary neurophysiology of neuron model, processing element, neural Network architecture, single layered feed forward network networks, recurrent networks.
Unit-II Neocognitron, neocognitron architectures, neocognitron data processing, neocognitron character recognition, neocognitron handwritten digit recognition, Neural phonetic typewriter.
Unit-III Neural network survey, Neural Network models, single layered preceptron, Multi layered perceptrons, XOR problem

4. STEPS OF INFORMATION RETRIEVAL

The tagged paragraphs are matched against the syllabus strings using page overlap and content filtering methods to segregate the document collections needed for analysis.

4.1. Page overlap – Content Extraction

Now the syllabus strings are ready for at least single-occurrence match from the self-acquired ontologies – the initial back ground knowledge being 'Table of Contents' or 'Back-of-the-book-index'. It is at this juncture, page overlap operation is performed over extracted pages of front and back indexes, resulting in the formation of concept space baseline. Here, the tagged paragraphs, as a result of pre-processing step, can be viewed as a significant parameter to determine page-range limits of the extracted pages tracking till the end of the concerned section header / sub-section header strings.

4.2. Semantic Content Filtering

The obvious storage structures for precise text content representation of the document content were selected to be n-grams. This usage of n-grams follows a recent successful implementations of machine generated back-of-the-book indexes [7] and machine generated subjective model-answers [8].

The authors are motivated to generate semantic concept spaces relevant to the topic terms, comprising of n-gram pools, depicting syllabus term–term co-occurrences. This is made possible by generating dependency relations (triples), as the text extracted in

Table 2
Page Overlap for Deciding the Baseline Documents from Syllabus Terms

Name	*Syllabus terms*	*Page nos. extracted from front / back index*
t_1	Elementary neurophysiology	8,293
t_2	Neuron model	NA(Not Available)
t_3	Processing Element	4,17,18
t_4	Neural network architecture	NA(Not Available)
t_5	Single layered feed forward network	NA(Not Available)
t_6	Multi layered feed forward network	NA(Not Available)
t_7	Recurrent networks	NA
t_8	Neocognitron	373-393
t_9	Neocognitron architectecture	376
t_{10}	Neocognitron data processing	381
t_{11}	Neocognitron character recognition	5
t_{12}	Neocognitron handwritten digital recognition	7
t_{13}	Neural phonetic typewriter	274,283
t_{14}	Neural Network survey	3, 41
t_{15}	Neural Network Models	3, 41
t_{16}	Single layered perceptron	17, 21, 24, 28
t_{17}	Multi layered perceptron	17, 21, 24, 28
t_{18}	XOR problem.	25, 26, 27

table 3 column 3, is put to dependency parser for learning semantics.

The dependencies among the paragraph level n-grams (may or may not be clustered into section/sub-section level n-grams), that lie as a part of filtered pages, can now be put to semantic content filtering process by computing statistical measures of term frequencies and term-to-term co-occurrence counts from extracted dependencies. These yield a set of page-ranges that encompass the full length explanation of the topic searched in target pages. It may also be noteworthy that these measures do contribute in the weighting process, i.e. determining the category levels of topic-significance for the extracted text that follows in *table 3, column 5*.

Moreover, it was observed from *table 3 column 4* that semantically-filtered page vicinities were found overlapping for some of the topics and so could be grouped together to get identified distinctly from other sets of overlapping page ranges. In this way, distinct document beds could be prepared so as to include all levels of co-occurrences among the closely related syllabus terms (topics) in each document bed.

As observed from the tabulations in table 3, column 6, such distinct page-ranges of text, nomenclated from d_1 to d_7 can be assigned to varied vicinities of filtered content, for the mentioned syllabus, all collectively shown in table 4. It may also be noted that, as the terms t_{14} and t_{15} exhibit one of their occurrences on page 41 that appears as a portion after the completion of a chapter in the form '*Suggested Reading & Bibliography*' and hence not considered as a search area for document ranking procedures.

Table 3
Category Levels of Relevance for the Filtered Content

tuple-id	*syllabus terms*	*relevant target pages*	*semantic-ally filtered page ranges*	*category level of relevance*	*observed relevant documents*
Unit-I					
u_1	t_1	8	8-17	c_1	d_1
u_2	t_1	293	291-293	c_3	d_2
u_3	t_3	4	4-7	c_2	d_3
u_4	t_3	17-18	17-30	c_2	d_4
u_5	t_8	373-393	373-393	c_1	d_5
u_6	t_9	376	376-393	c_1	d_5
u_7	t_{10}	381	381-393	c_1	d_5
Unit-II					
u_8	t_{11}	5	5-7	c_3	d_3
u_9	t_{12}	7	6-7	c_3	d_3
u_{10}	t_{13}	274	274-275	c_2	d_6
u_{11}	t_{13}	283	283-286	c_3	d_7
u_{12}	t_{14}, t_{15}	3	3-7	c_3	d_3
u_{13}	t_{14}, t_{15}	41	NULL	—	—-
Unit-III					
u_{14}	t_{16}, t_{17}	17	17-30	c_3	d_4
u_{15}	t_{16}, t_{17}	21	21-30	c_2	d_4
u_{16}	t_{16}, t_{17}	28	28-30	c_3	d_4
u_{17}	t_{16}, t_{17}	24	24-30	c_2	d_4
u_{18}	t_{18}	25 -27	25-30	c_3	d_4

Now, computing the category level of topical relevance from the semantically filtered page ranges becomes an easy task, if presumed to be categorized into c_1, c_2 and c_3 each interpreted as : c_1 , chapter/section header level category of selected text, c_2, page/sub-section level category and c_3 , paragraph level category. In this way, they are weighted according to the levels of useful content coverage either in the vicinity of within few paragraphs or within pages or entire sections / sub-sections.

Table 4
Distinct Document Beds for Relevance Measures

Document bed	*Section*	*Page Range*
d_1	1.1	8-17
d_2	8.0	291-293
d_3	1.0	1-7
d_4	1.2	17-30
d_5	chapter 10	373-393
d_6	7.2.1	274-275
d_7	7.3.2	281-286

5. TRAINING AND TEST COLLECTIONS

After the content extraction process through semantic filtering technique, the carefully cleansed and trained documents are tested upon with a naïve bayes classifier, a frequently chosen one, among several popular statistical machine learning techniques. Since, document classification can be viewed as the calculation of the statistical distribution of topic terms into specific document beds, a Bayesian classifier first trains the model by calculating a generative document distribution $P(d_j \mid u_i)$ for the observed relevance of each syllabus tuple 'u_i' in the document 'd_j' and then tests into which document does the term 'u_i' finds the predicted relevance. Since the above method handles high dimensional data sets as huge text corpora, they can be used for effective statistical inferencing [11].

5.1. The Prior Probability Measures

To predict the probability of most relevant document bed for a term 'u_i' provided it belongs to category 'c_k' and observed document 'd_j' , initially the syllabus term counts are framed belonging to category c_k and document bed d_j. The count values for seventeen such terms are distributed among seven document beds, as shown in table 5.

Table 5
Term-count Matrix with Respect to Two dimensions, Categories and Documents

c_k \ d_7	d_j	d_1	d_2	d_3	d_4	d_5	d_6
c_1	1	0	0	0	3	0	0
c_2	0	0	1	3	0	1	0
c_3	0	1	3	3	0	0	1
$\Sigma c_k \backslash d_j$	1	1	4	6	3	1	1
$P(d_j)$	1/17	1/17	4/17	6/17	3/17	1/17	1/17

From the above, the term-probability matrix can be formulated, comprising of conditional probabilities, $P(u_i \mid d_j)$, upon the stated conditions, revealing the extent up to which they belong to respective category levels of content usage. The computed values are shown in table 6.

Table 6
Term-probability Matrix for Documents d_1 to d_7

$P(u_i \mid d_j)$	d_1	d_2	d_3	d_4	d_5	d_6	d_7
c_1	1/1	0	0	0	3/3	0	0
c_2	0	0	1/4	3/6	0	1/1	0
c_3	0	1/1	3/4	3/6	0	0	1

5.2. The Posterior Probability Measures

One domain with potentially large number of classes and having very high dimensional data is text. One dimension to large scale classification which has not been explored in sufficient depth in the text mining literature, is the scaling up of the classification systems with respect to a very large number of classes. In text classification, except Naïve Bayesian systems, most of the other systems, implementing SVMs and Neural Nets, involve solving complex sub-problems often exponential in time complexity. So, Naïve Bayesian classifiers are presumed to be well known to perform fairly well towards solving multi-class classification problems. The Bayesian conditional probabilities for classifying of syllabus terms into one of the assigned document classes is expressed as:

$$P\left(d_k / t_i\right) = \frac{P\left(t_i / d_k\right) * P\left(d_k\right)}{\sum P\left(t_i / d_j\right) * P\left(d_j\right)}$$

where d_k represents k^{th} document bed and t_i the ith term for which the predicted document bed is to be calculated among 'm' number of documents i.e : holding maximum value of $P(d_k|t_i)$.

The seven conditional probability values for predicting the content coverage of term u_1 among seven distinct documents were computed as shown in the table 7. It was observed that the conditional probabilities $P(d_1 \mid u_1)$ and $P(d_5 \mid u_1)$ attain non-zero values indicating some what coverage, while $P(d_2 \mid u_1)$, $P(d_3 \mid u_1)$, $P(d_4 \mid u_1)$, $P(d_6 \mid u_1)$ and $P(d_7 \mid u_1)$ take over zero values depicting *nil* coverage.

Of the two documents, d_1 and d_5, the coverage was maximally observed in document d_5 and hence, document of relevance for syllabus term u_1 was predicted as d_5.

Similarly, the non-zero content coverage for the term u_2 were computed as $P(d_2 \mid t_2)$, $P(d_3 \mid t_2)$, $P(d_4 \mid t_2)$ and $P(d_7 \mid t_2)$ of which the conditional probability measures in context with documents d_3 and d_4 obtain the highest

Table 7
Computed Posterior Probability Measures for Terms u_1 and u_2

i	j	$P(u_i\|d_j)$	$P(d_j)$	$P(u_i\|d_i)^* P(d_i)$	$\Sigma P(u_i\|d_i)^* P(d_i)$	$P(d_j\|u_i)$
1	1	1	1/17	1/17	4/17	1/4
1	2	0	1/17	0		0
1	3	0	4/17	0		0
1	4	0	6/17	0		0
1	5	1	3/17	3/17		3/4
1	6	0	1/17	0		0
1	7	0	1/17	0		0
2	1	0	1/17	0	8/17	0
2	2	1	1/17	1/17		1/8
2	3	3/4	4/17	3/17		3/8
2	4	1/2	6/17	3/17		3/8
2	5	0	3/17	0		0
2	6	0	1/17	0		0
2	7	1	1/17	1/17		1/8

values, predicting the maximum content coverage. Hence, the document of relevance for syllabus term u_2 was predicted as d_3 and d_4.

Continuing with this approach, given the hypothesis that every term lies in the vicinity of finitely extracted page ranges of the observed document beds 'd_j s', the predicted document beds were calculated for each of the terms u_3 till u_{15} as shown in table 8, column 4.

6. CONFUSION MATRICES: MATHEMATICAL COMPOSITION TO DOCUMENT RANKING

The Confusion Matrix is a standard output representation of classification problems. It gives the predicted distribution of the test instances into each of the trained classes [2] [11]. If attained 100% Bayesian classification accuracy, the n X n matrix only has diagonal elements corresponding to all the test terms being correctly predicted to their true class.

But, in reality, the matrix smudged with small values, distributed all over, gives much more information about the nature of the classification problem, here in our case, interpreting the degree of semantic overlap among the fuzzy boundaries of the selected document beds. Thus, the observed relevant documents are compared with the predicted relevant ones showing the actual membership degree of one with respect to the other as shown in the 7 X 7 document confusion matrix depicted in table 9.

Table 8
The Predicted Documents of Relevance for the Unique Tuple-ids

Tuple-id	*Syllabus terms*	*Observed Relevant documents*	*Predicted Relevant documents (with Bayesian classifier)*
u_1	t_1	d_1	d_5
u_2	t_1	d_2	d_3, d_4
u_3	t_3	d_3	d_4
u_4	t_3	d_4	d_4
u_5	t_8	d_5	d_5
u_6	t_9	d_5	d_5
u_7	t_{10}	d_5	d_5
u_8	t_{11}	d_3	d_3, d_4
u_9	t_{12}	d_3	d_3, d_4
u_{10}	t_{13}	d_6	d_4
u_{11}	t_{13}	d_7	d_3, d_4
u_{12}	t_{14}, t_{15}	d_3	d_3, d_4
u_{13}	t_{14}, t_{15}	—	—
u_{14}	t_{16}, t_{17}	d_4	d_3, d_4
u_{15}	t_{16}, t_{17}	d_4	d_4
u_{16}	t_{16}, t_{17}	d_4	d_3, d_4
u_{17}	t_{16}, t_{17}	d_4	d_4
u_{18}	t_{18}	d_4	d_3, d_4

Table 9
Confusion Matrix for Assigned and Predicted Relevant Documents

Observed document bed \ *Predicted document bed*	d_1	d_2	d_3	d_4	d_5	d_6	d_7
d_1	cf_{11} =0	cf_{12} =0	0	0	1	0	0
d_2	0	0	1	1	0	0	0
d_3	0	0	3	4	0	0	0
d_4	0	0	3	6	0	0	0
d_5	0	0	0	0	3	0	0
d_6	0	0	0	1	0	0	0
d_7	0	0	1	1	0	0	cf_{77} =0

[where $cf_{11} = fd_1(d_1)$, $cf_{17} = fd_1(d_7)$ etc.]

7. RANKING CRITERIA

Decision making for document relevancy could have been done on the basis of crisp ordinal raking, if relevance measures had been deterministic with no ambiguities. However, in the situations, where the syllabus coverage is not confined to any specific chapter of the learning material, the authors feel more determined to state that the syllabus topics of a unit might be found overlapping

across the section-header or chapter boundaries. Hence, the nature of the content-finding process (which precisely defines the problem statement) being fuzzy in nature, fuzzy ordering seems to be an appealing technique of ranking, for there do exist non-cardinal type term-to-term relationships lying in the enormous chapter text corpora.

7.1. The Document Preferences : Fuzzy Ordering

On comparing the document beds for relevancy ranking to the terms of syllabus corpus, the pair-wise membership functions are formalized to accommodate this form of non-transitive ranking. These represent the subjective measurement of the appropriateness of each predicted document bed when compared only to the other observed document bed for a particular term. This can be substituted for the actual membership degree of one with respect to the other, collectively forming the so-called fuzzy preference relation matrix 'R' for the collection of syllabus terms [12].

Hence for the syllabus terms u_1 to u_{18}, the fuzzy preference matrix '*R*' becomes :

$$\begin{array}{c} \\ d_1 \\ d_2 \\ | \\ | \\ d_5 \end{array} \begin{array}{c} \begin{array}{ccccc} d_1 & d_2 & d_3 & d_4 & d_5 \end{array} \\ \begin{pmatrix} f_{d1}(d_1) & f_{d1}(d_2) & f_{d1}(d_3) & f_{d1}(d_4) & f_{d1}(d_5) \\ | & f_{d2}(d_2) & | & | & | \\ | & | & f_{d3}(d_3) & | & | \\ | & | & | & f_{d4}(d_4) & | \\ f_{d5}(d_1) & f_{d5}(d_2) & f_{d5}(d_3) & f_{d5}(d_4) & f_{d5}(d_5) \end{pmatrix} \end{array}$$

where the confusion degree values of the confusion matrix (table 9) is visualized as an intermediary platform that can be appropriately exploited to determine the document class-pair-wise fuzzy membership degrees as illustrated in matrix 'R' below, depicting subjective measures of comparisons for preferring a column entity, d_j over a row entity, d_i symbolized by $f_{di}(d_j)$ [11] [12].

Consequently, the elements of the confusion matrix derived from *table 10*, when substituted for these pair-wise fuzzy membership values, resulted in framing up of fuzzy preference matrix as shown below.

$$R = \begin{array}{c} \\ d_1 \\ d_2 \\ d_3 \\ d_4 \\ d_5 \\ d_6 \\ d_7 \end{array} \begin{array}{c} \begin{array}{ccccccc} d_1 & d_2 & d_3 & d_4 & d_5 & d_6 & d_7 \end{array} \\ \begin{bmatrix} 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 3 & 4 & 0 & 0 & 0 \\ 0 & 0 & 3 & 6 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 3 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 & 0 & 0 \end{bmatrix} \end{array}$$

7.2. Relativity Functions : Document Comparison

The above mentioned pair-wise membership values of the fuzzy preference matrix featuring the classification accuracy, were further used to measure relative membership values of classifying document d_j over d_i in the form of relativity function as denoted in the expression below [12].

$$f(d_j \mid d_i) = \frac{f_{di}(dj)}{\max[\, f_{di}(dj),\ f_{dj}(d_i)]}$$

where $f\,(d_j \mid d_i)$ is the relativity function for choosing d_j over d_i. On employing the above relativity function for any two comparable fuzzy parameters of matrix 'R', the comparison matrix 'C' can thus be obtained as:

$$C = \begin{pmatrix} 1 & f(d_1 \mid d_2) & -- & ---- & f(d_1 \mid d_5) \\ -- & 1 & -- & ---- & ------ \\ -- & & & & \\ -- & -- & -- & ---- & \\ f(d_5 \mid d_1) & f(d_5 \mid d_2) & -- & -- & 1 \end{pmatrix}$$

Hence, for the put up observations, the values subsequently computed as relativity function measures can be shown below as in matrix 'C':

$$C = \begin{array}{c} \\ d_1 \\ d_2 \\ d_3 \\ d_4 \\ d_5 \\ d_6 \\ d_7 \end{array} \begin{array}{c} \begin{array}{ccccccc} d_1 & d_2 & d_3 & d_4 & d_5 & d_6 & d_7 \end{array} \\ \begin{bmatrix} 1 & 1 & 1 & 1 & 0 & 1 & 1 \\ 1 & 1 & 0 & 0 & 1 & 1 & 1 \\ 1 & 1 & 1 & 0.75 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 0 & 1 & 1 & 1 \\ 1 & 1 & 0 & 0 & 1 & 1 & 1 \end{bmatrix} \end{array}$$

To determine the overall ranking, we select the minimum of the membership values among the seven documents, as this value indicates the minimum ensured chances of selecting document 'd_i' over documents 'd_j s'.

$$C' = \min [C] = \begin{array}{c} \\ d_1 \\ d_2 \\ d_3 \\ d_4 \\ d_5 \\ d_6 \\ d_7 \end{array} \begin{bmatrix} 0 \\ 0 \\ 0.75 \\ 1 \\ 1 \\ 0 \\ 0 \end{bmatrix}$$

8. RESULTS AND DISCUSSIONS

The statistics, as analyzed by Subject Experts reveal that *seven* of the search terms were found content-relevant in the document d_4 and d_5, and *four* terms were found relevant in d_3 and d_1 each.

The similar results were found in the columnar matrix, *C*', depicting highly ranked text document beds, d_4 and d_5 carrying maximum topic-relevance with the given set of syllabus terms, followed by the document, d_3 among the seven analyzed ones. Further, some of these relevantly ranked documents were, in actual, found embedded under the respective chapter headers or frequently occurring page ranges, in which the search string patterns were successfully matched. The results were also found congenial to that obtained by one of the conventional approaches, that computed the relevance numbers, i.e. degree of document usages uponUsing this notion of relevance number, the documents d_4 and d_5 are ranked at the top, followed by d_3 and d_1, followed by d_6 and then d_2 and d_7 ranked last in order, (summarized in *table 10)*. These interpret well with the rank-values reflected by each document in the above columnar matrix, C', pertaining to the mentioned text-material put up in the example.

9. CONCLUSIONS

The two ranking techniques need further in-depth analysis of relevantly identified documents, if compared for the same syllabus string across chapter boundaries and book boundaries, pertaining to same subject domain. Thus, one can obviously visualize this work, in extended scope, as drafting a tool in order to rank courseware books for fast and accurate browsing in Academic Digital Libraries. This may further help in effective Information Retrieval by user-communities, sharing common interests in subject specificity.

Table 10
Relevance Numbers Due to Search Content for the Syllabus Terms

di / Ij	*P(A, d_1, Ij) · P(A, d_1)*	*P(A, d2, Ij) · P(A, d_2)*	*P(A, d3, Ij) · P(A, d_3)*	*P(A, d4, Ij) · P(A, d_4)*	*P(A, d5, Ij) · P(A, d_5)*	*P(A, d6, Ij) · P(A, d_6)*	*P(A, d7, Ij) · P(A, d_7)*	*Most topic-relevant document*	*Relevance Number (Maximum)*
I_1	5· 4/18	0	0	0	0	0	0	d_1	20/18
I_2	0· 4/18	0	0	0	0	0	0	—-	——
I_3	5· 4/18	1· 1/18	1· 4/18	8· 7/18	2· 7/18	0	0	d_4	56/18
I_4	0	0	0	6· 7/18	7· 7/18	0	0	d_5	49/18
I_5	0	0	0	0	0	0	0	—-	
I_6	0	0	0	0	0	0	0	—-	
I_7	0	0	0	0	0	0	0	—-	
I_8	0	0	0	0	59· 7/18	0	0	d_5	413/18
I_9	0	0	0	0	59· 7/18	0	0	d_5	413/18
I_{10}	0	0	0	0	59· 7/18	0	0	d_5	413/18
I_{11}	0	0	1· 4/18	0	0	0	0	—-	——
I_{12}	0	0	0	0	0	0	0	—-	——
I_{13}	0	0	0	0	0	3· 3/18	1· 1/18	d_6	9/18
I_{14}	2· 4/18	0	5· 4/18	6· 7/18	7· 7/18	1· 3/18	0	d_5	49/18
I_{15}	2· 4/18	0	5· 4/18	6· 7/18	7· 7/18	1· 3/18	0	d_5	49/18
I_{16}	0	0	0	33· 7/18	0	0	0	d_4	231/18
I_{17}	0	0	0	33· 4/18	0	0	0	d_4	231/18
I_{18}	0	0	0	6· 4/18	0	0	0	d_4	42/18

10. ACKNOWLEDGEMENTS

The work is being carried in Research and Development Cell, Bhilai Institute of Technology, Durg, India. The authors wish to express their heart felt gratitude to the Management, for their inspiring encouragement and support towards the completion of the work.

REFERENCES

[1] Arpana Rawal, H. R. Sharma, M. K. Kowar, Sanjay Sharma (2009), *"Ranking Text Relevance Measures using Probabilistic Techniques" International Journal on Computer Engineering and Information Technology, IJCEIT*, ISSN 0974-2034, **1**(1), 2009, 17-22.

[2] Arpana Rawal, M. K. Kowar and Sanjay Sharma Exploiting Confusion Matrices for Relevance Ranking of Text Documents, Proceedings of 1st International Conference on Emerging Trends in Engineering and Technology, ICETET' 08, India, Online IEEE Explore System, pp. 492-497, ISBN: 978-0-7695-3267-7, 2008.

[3] El-Sayed Atlam, K. Morita, M. Fuketa, Jun-ichi Aoe (2002), "A New Method for Selecting English Field Association Terms of Compound Words and its Knowledge Representation" *International Journal of Information Processing and Management*, 38, © 2002 Elsevier Science Ltd, pp. 807-821.

[4] El-Sayed Atlam, M. Fuketa, K. Morita, Jun-ichi Aoe (2003), "Document Similarity Measurement using Field Association Terms" *International Journal of Information Processing and Management*, 38, © 2003 Elsevier Science Ltd., pp. 809-824.

[5] Gred Stumme, Julien Tane, Christoph Schmitz, Steffen Staab, Rudi Studer (2003), "The Courseware Watchdog – an Ontology-based Tool for Finding and Organizing Learning Material", Learning Lab Lower Saxony (L3S), Hannover, Germany; *www.learninglab.de* and Institute for Applied Informatics and Formal Description Methods (AIFB), University of Karlsruhe, Karlsruhe, Germany; *www.aifb.uni-karlsruhe.de/WBS/*.

[6] M. E Maron. The RAND Corporation, Santa Monica, California, Kuhns J. L. Ramo-Wooldridge, Canoga Park, California (1959), "On Relevance, Probabilistic Indexing and Information Retrieval".

[7] M. K. Kowar, Arpana Rawal, Ani Thomas (2006), "Automatic Generation of Back-of-the-Book-Index : An Integrated Approach through Text Mining Operations", Proceedings of IRIS 06 - International Conference on Recent Trends in Information Systems, NEC- Kovilpatti, India, pp. 416-423.

[8] M. K. Kowar, Arpana Rawal, Ani Thomas, Sanjay Sharma (2007), "Fuzzy Decision Making for Automatic Answer Evaluation in Restricted Domains", *Journal Reflections' des ERA, Modinagar*, India, **3**(1), 15-26, (2008).

[9] M. K., Kowar, S. Sharma, A. Rawal, A. Thomas (2009), "An Intelligent Hybrid Tool for Finding and Organizing Relevant Text", *International Journal of Computer Science and Applications*, **2**(1), 34-37.

[10] Rohini U[1] and Vamshi Ambati[2] , "A Colllaborative Filtering based Re-ranking Strategy for Search in Digital Libraries", In the Proceedings of 8th International Conference ICADL, 2005.

[11] Shantanu Godbole (2002), "Exploiting Confusion Matrices for Automatic Generation of Topic Hierarchies and Scaling up Multi-way Classifiers", Annual Progress Report, Indian Institute of Technology – Bombay, India.

[12] Timothy J. Ross (1997), "Fuzzy Logic with Engineering Applications" Mc . Graw Hill, Inc. International Edition, pages 102-105, 317-322.