A Fuzzy Document Ranking System using Confusion Matrices

Arpana Rawal¹, Manoj Kumar Kowar² and Sanjay Sharma³

¹Reader, Department of Computer Science
²Dept. of Electronics & Telecommunications
³Professor, Department of Mathematics
Bhilai Institute of Technology, Bhilai House, Durg (C.G.) – 491001 India
E-mail: arpana_rawal@rediffmail.com, kowar_bit@Rediffmail.com, ssarma_bit@yahoo.co.in

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 background 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 metadata repositories to begin the conceptual searches within document titles, abstracts, headings, sub-headings and using author profiles on the extreme [10]. This confirmed 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>
<thead>
<tr>
<th>Table 1</th>
<th>Syllabus Snapshot</th>
</tr>
</thead>
<tbody>
<tr>
<td>Unit-I</td>
<td>Elementary neurophysiology of neuron model, processing element, neural Network architecture, single layered feed forward network networks, recurrent networks.</td>
</tr>
<tr>
<td>Unit-II</td>
<td>Neocognitron, neocognitron architectures, neocognitron data processing, neocognitron character recognition, neocognitron handwritten digit recognition, Neural phonetic typewriter.</td>
</tr>
<tr>
<td>Unit-III</td>
<td>Neural network survey, Neural Network models, single layered perceptron, Multi layered perceptrons, XOR problem</td>
</tr>
</tbody>
</table>

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
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**

<table>
<thead>
<tr>
<th>tuple-id</th>
<th>syllabus terms</th>
<th>relevant target pages</th>
<th>semantically filtered page ranges</th>
<th>category level of relevance</th>
<th>observed relevant documents</th>
</tr>
</thead>
<tbody>
<tr>
<td>$u_1$</td>
<td>$t_1$</td>
<td>8</td>
<td>8-17</td>
<td>$c_1$</td>
<td>$d_1$</td>
</tr>
<tr>
<td>$u_2$</td>
<td>$t_1$</td>
<td>293</td>
<td>291-293</td>
<td>$c_1$</td>
<td>$d_1$</td>
</tr>
<tr>
<td>$u_3$</td>
<td>$t_3$</td>
<td>4</td>
<td>4-7</td>
<td>$c_2$</td>
<td>$d_4$</td>
</tr>
<tr>
<td>$u_4$</td>
<td>$t_3$</td>
<td>17-18</td>
<td>17-30</td>
<td>$c_3$</td>
<td>$d_4$</td>
</tr>
<tr>
<td>$u_5$</td>
<td>$t_8$</td>
<td>373-393</td>
<td>373-393</td>
<td>$c_4$</td>
<td>$d_5$</td>
</tr>
<tr>
<td>$u_6$</td>
<td>$t_9$</td>
<td>376</td>
<td>376-393</td>
<td>$c_4$</td>
<td>$d_5$</td>
</tr>
<tr>
<td>$u_7$</td>
<td>$t_{10}$</td>
<td>381</td>
<td>NA</td>
<td>NA</td>
<td></td>
</tr>
</tbody>
</table>

### Table 3 (Continued)

<table>
<thead>
<tr>
<th>Unit-II</th>
</tr>
</thead>
<tbody>
<tr>
<td>$u_{10}$</td>
</tr>
<tr>
<td>$u_{11}$</td>
</tr>
<tr>
<td>$u_{12}$</td>
</tr>
<tr>
<td>$u_{13}$</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Unit-III</th>
</tr>
</thead>
<tbody>
<tr>
<td>$u_{14}$</td>
</tr>
<tr>
<td>$u_{15}$</td>
</tr>
<tr>
<td>$u_{16}$</td>
</tr>
<tr>
<td>$u_{17}$</td>
</tr>
<tr>
<td>$u_{18}$</td>
</tr>
</tbody>
</table>

Now, computing the category level of topical relevance from the semantically filtered page ranges becomes an easy task, if presumed to be categorized into $c_i$, $c_j$ and $c_k$ each interpreted as : $c_i$, chapter/section header level category of selected text, $c_j$, page/sub-section level category and $c_k$, 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 / subsections.
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 naive 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 \( P(d_j | u) \) for the observed relevance of each syllabus tuple ‘\( u \)’ in the document ‘\( d_j \)’ and then tests into which document does the term ‘\( u \)’ 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 ‘\( u \)’ 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.

\[
\begin{array}{cccccccc}
\text{Table 4} & \text{Distinct Document Beds for Relevance Measures} \\
\hline
\text{Document bed} & \text{Section} & \text{Page Range} \\
\hline
d_1 & 1.1 & 8-17 \\
d_2 & 8.0 & 291-293 \\
d_3 & 1.0 & 1-7 \\
d_4 & 1.2 & 17-30 \\
d_5 & \text{chapter 10} & 373-393 \\
d_6 & 7.2.1 & 274-275 \\
d_7 & 7.3.2 & 281-286 \\
\end{array}
\]

\[
\begin{array}{cccccccc}
\text{Table 5} & \text{Term-count Matrix with Respect to Two dimensions, Categories and Documents} \\
\hline
\text{\( c_k \)} & d_1 & d_2 & d_3 & d_4 & d_5 & d_6 \\
\hline
\text{\( c_1 \)} & 1 & 0 & 0 & 0 & 3 & 0 & 0 \\
\text{\( c_2 \)} & 0 & 0 & 1 & 3 & 0 & 1 & 0 \\
\text{\( c_3 \)} & 0 & 1 & 3 & 0 & 0 & 1 & 0 \\
\Sigma_{\text{\( c_k \)}} & 1 & 1 & 4 & 6 & 3 & 1 & 1 \\
\end{array}
\]

From the above, the term-probability matrix can be formulated, comprising of conditional probabilities, \( P(u | d) \), 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.

\[
\begin{array}{cccccccc}
\text{Table 6} & \text{Term-probability Matrix for Documents \( d_1 \) to \( d_7 \)} \\
\hline
\text{\( P(u|d_j) \)} & d_1 & d_2 & d_3 & d_4 & d_5 & d_6 & d_7 \\
\hline
\text{\( c_1 \)} & 1/17 & 0 & 0 & 0 & 3/17 & 0 & 0 \\
\text{\( c_2 \)} & 0 & 0 & 1/17 & 3/17 & 0 & 1/17 & 0 \\
\text{\( c_3 \)} & 0 & 1/17 & 3/17 & 3/17 & 0 & 0 & 1/17 \\
\end{array}
\]

\[
P(d_k | t_i) = \frac{P(t_i | d_k) \times P(d_k)}{\sum P(t_i | d_j) \times P(d_j)}
\]

where \( d_k \) represents the \( k \)th document bed and \( t \) the \( i \)th term for which the predicted document bed is to be calculated among ‘\( m \)’ number of documents i.e : holding maximum value of \( P(d_k | t) \).

The seven conditional probability values for predicting the content coverage of term \( u_i \) among seven distinct documents were computed as shown in the table 7. It was observed that the conditional probabilities \( P(d_j | u_i) \) and \( P(d_j | u_i) \) attain non-zero values indicating some what coverage, while \( P(d_j | u_i) \), \( P(d_j | u_i) \), \( P(d_j | u_i) \), \( P(d_j | u_i) \) and \( P(d_j | u_i) \) and \( P(d_j | u_i) \) take over zero values depicting nil coverage.

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

Similarly, the non-zero content coverage for the term \( u_i \) were computed as \( P(d_1 | t_i) \), \( P(d_1 | t_i) \), \( P(d_1 | t_i) \) and \( P(d_1 | t_i) \) of which the conditional probability measures in context with documents \( d_1 \) and \( d_4 \) obtain the highest
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 \( n \times n \) 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>
<thead>
<tr>
<th>Table 7</th>
<th>Computed Posterior Probability Measures for Terms ( u_1 ) and ( u_2 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( i )</td>
<td>( j )</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>1</td>
<td>3</td>
</tr>
<tr>
<td>1</td>
<td>4</td>
</tr>
<tr>
<td>1</td>
<td>5</td>
</tr>
<tr>
<td>1</td>
<td>6</td>
</tr>
<tr>
<td>1</td>
<td>7</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>2</td>
<td>4</td>
</tr>
<tr>
<td>2</td>
<td>5</td>
</tr>
<tr>
<td>2</td>
<td>6</td>
</tr>
<tr>
<td>2</td>
<td>7</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Table 8</th>
<th>The Predicted Documents of Relevance for the Unique Tuple-ids</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tuple-id</td>
<td>Syllabus terms</td>
</tr>
<tr>
<td>( u_1 )</td>
<td>( t_1 )</td>
</tr>
<tr>
<td>( u_2 )</td>
<td>( t_1 )</td>
</tr>
<tr>
<td>( u_3 )</td>
<td>( t_3 )</td>
</tr>
<tr>
<td>( u_4 )</td>
<td>( t_3 )</td>
</tr>
<tr>
<td>( u_5 )</td>
<td>( t_8 )</td>
</tr>
<tr>
<td>( u_6 )</td>
<td>( t_9 )</td>
</tr>
<tr>
<td>( u_7 )</td>
<td>( t_{10} )</td>
</tr>
<tr>
<td>( u_8 )</td>
<td>( t_{11} )</td>
</tr>
<tr>
<td>( u_9 )</td>
<td>( t_{12} )</td>
</tr>
<tr>
<td>( u_{10} )</td>
<td>( t_{13} )</td>
</tr>
<tr>
<td>( u_{11} )</td>
<td>( t_{13} )</td>
</tr>
<tr>
<td>( u_{12} )</td>
<td>( t_{14}, t_{15} )</td>
</tr>
<tr>
<td>( u_{13} )</td>
<td>( t_{14}, t_{15} )</td>
</tr>
<tr>
<td>( u_{14} )</td>
<td>( t_{16}, t_{17} )</td>
</tr>
<tr>
<td>( u_{15} )</td>
<td>( t_{16}, t_{17} )</td>
</tr>
<tr>
<td>( u_{16} )</td>
<td>( t_{16}, t_{17} )</td>
</tr>
<tr>
<td>( u_{17} )</td>
<td>( t_{16}, t_{17} )</td>
</tr>
<tr>
<td>( u_{18} )</td>
<td>( t_{16}, t_{17} )</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Table 9</th>
<th>Confusion Matrix for Assigned and Predicted Relevant Documents</th>
</tr>
</thead>
<tbody>
<tr>
<td>Predicted document bed</td>
<td>( d_1 )</td>
</tr>
<tr>
<td>Observed document bed</td>
<td>( cf_{11} )</td>
</tr>
<tr>
<td>( d_2 )</td>
<td>0</td>
</tr>
<tr>
<td>( d_3 )</td>
<td>0</td>
</tr>
<tr>
<td>( d_4 )</td>
<td>0</td>
</tr>
<tr>
<td>( d_5 )</td>
<td>0</td>
</tr>
<tr>
<td>( d_6 )</td>
<td>0</td>
</tr>
<tr>
<td>( d_7 )</td>
<td>0</td>
</tr>
</tbody>
</table>

[where \( cf_{ij} = fd_i(d_j) \), \( cf_{ij} = fd_i(d_j) \) etc.]

7. RANKING CRITERIA

Decision making for document relevancy could have been done on the basis of crisp ordinal ranking, if relevancy 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}{ccccccc}
\text{\( d_1 \)} & \text{\( d_2 \)} & \text{\( d_3 \)} & \text{\( d_4 \)} & \text{\( d_5 \)} \\
\text{\( d_1 \)} & f_{d_1}(d_1) & f_{d_1}(d_2) & f_{d_1}(d_3) & f_{d_1}(d_4) & f_{d_1}(d_5) \\
\text{\( d_2 \)} & f_{d_2}(d_1) & f_{d_2}(d_2) & f_{d_2}(d_3) & f_{d_2}(d_4) & f_{d_2}(d_5) \\
\text{\( d_3 \)} & f_{d_3}(d_1) & f_{d_3}(d_2) & f_{d_3}(d_3) & f_{d_3}(d_4) & f_{d_3}(d_5) \\
\text{\( d_5 \)} & f_{d_5}(d_1) & f_{d_5}(d_2) & f_{d_5}(d_3) & f_{d_5}(d_4) & f_{d_5}(d_5) \\
\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 \( f_{d_i}(d_j) \) [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.

\[
\begin{array}{ccccccc}
\text{\( d_1 \)} & \text{\( d_2 \)} & \text{\( d_3 \)} & \text{\( d_4 \)} & \text{\( d_6 \)} & \text{\( d_7 \)} \\
\text{\( d_1 \)} & 0 & 0 & 0 & 0 & 0 & 0 \\
\text{\( d_2 \)} & 0 & 0 & 1 & 0 & 0 & 0 \\
\text{\( d_3 \)} & 0 & 0 & 3 & 0 & 0 & 0 \\
\text{\( d_5 \)} & 0 & 0 & 3 & 0 & 0 & 0 \\
\text{\( d_6 \)} & 0 & 0 & 0 & 1 & 0 & 0 \\
\text{\( d_7 \)} & 0 & 0 & 1 & 0 & 0 & 0 \\
\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 \( d_j \) in the form of relativity function as denoted in the expression below [12].

\[
f(d_j | d_i) = \frac{f_{d_i}(d_j)}{\max [f_{d_i}(d_j), f_{d_i}(d_j)]}
\]

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

\[
C = \begin{pmatrix}
1 & f(d_1 | d_2) & \ldots & \ldots & f(d_1 | d_5) \\
f(d_2 | d_1) & 1 & \ldots & \ldots & \ldots \\
f(d_3 | d_1) & f(d_3 | d_2) & 1 & \ldots & \ldots \\
f(d_4 | d_1) & f(d_4 | d_2) & f(d_4 | d_3) & 1 & \ldots \\
f(d_5 | d_1) & f(d_5 | d_2) & f(d_5 | d_3) & f(d_5 | d_4) & 1
\end{pmatrix}
\]

Hence, for the put up observations, the values subsequently computed as relativity function measures can be shown below as in matrix ‘\( C \)’:

\[
\begin{array}{ccccccc}
\text{\( d_1 \)} & \text{\( d_2 \)} & \text{\( d_3 \)} & \text{\( d_4 \)} & \text{\( d_5 \)} & \text{\( d_6 \)} & \text{\( d_7 \)} \\
\text{\( d_1 \)} & 0 & 1 & 1 & 1 & 0 & 1 & 1 \\
\text{\( d_2 \)} & 1 & 1 & 0 & 0 & 1 & 1 & 1 \\
\text{\( d_3 \)} & 1 & 1 & 1 & 0.75 & 1 & 1 & 1 \\
\text{\( d_4 \)} & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\
\text{\( d_5 \)} & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\
\text{\( d_6 \)} & 1 & 1 & 1 & 0 & 1 & 1 & 1 \\
\text{\( d_7 \)} & 1 & 1 & 0 & 0 & 1 & 1 & 1 \\
\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 ‘\( d_j \)’.
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 \(d_1\) and \(d_2\), 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

<table>
<thead>
<tr>
<th>(di / Ij)</th>
<th>(P(A, d_1, Ij))</th>
<th>(P(A, d_2, Ij))</th>
<th>(P(A, d_3, Ij))</th>
<th>(P(A, d_4, Ij))</th>
<th>(P(A, d_5, Ij))</th>
<th>(P(A, d_6, Ij))</th>
<th>(P(A, d_7, Ij))</th>
<th>Most topic-relevant document</th>
<th>Relevance Number (Maximum)</th>
</tr>
</thead>
<tbody>
<tr>
<td>(I_1)</td>
<td>5· 4/18</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>(d_1)</td>
<td>56/18</td>
</tr>
<tr>
<td>(I_2)</td>
<td>0· 4/18</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>(d_1)</td>
<td>49/18</td>
</tr>
<tr>
<td>(I_3)</td>
<td>5· 4/18</td>
<td>1· 1/18</td>
<td>1· 4/18</td>
<td>8· 7/18</td>
<td>2· 7/18</td>
<td>0</td>
<td>0</td>
<td>(d_4)</td>
<td>413/18</td>
</tr>
<tr>
<td>(I_4)</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>6· 7/18</td>
<td>7· 7/18</td>
<td>0</td>
<td>0</td>
<td>(d_5)</td>
<td>49/18</td>
</tr>
<tr>
<td>(I_5)</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>(d_5)</td>
<td>413/18</td>
</tr>
<tr>
<td>(I_6)</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>(d_5)</td>
<td>413/18</td>
</tr>
<tr>
<td>(I_7)</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>(d_5)</td>
<td>413/18</td>
</tr>
<tr>
<td>(I_8)</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>(d_5)</td>
<td>413/18</td>
</tr>
<tr>
<td>(I_9)</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>(d_5)</td>
<td>413/18</td>
</tr>
<tr>
<td>(I_{10})</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>(d_5)</td>
<td>413/18</td>
</tr>
<tr>
<td>(I_{11})</td>
<td>0</td>
<td>0</td>
<td>1· 4/18</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>(d_5)</td>
<td>413/18</td>
</tr>
<tr>
<td>(I_{12})</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>(d_5)</td>
<td>413/18</td>
</tr>
<tr>
<td>(I_{13})</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>(d_5)</td>
<td>413/18</td>
</tr>
<tr>
<td>(I_{14})</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>(d_5)</td>
<td>413/18</td>
</tr>
<tr>
<td>(I_{15})</td>
<td>2· 4/18</td>
<td>0</td>
<td>5· 4/18</td>
<td>6· 7/18</td>
<td>7· 7/18</td>
<td>1· 3/18</td>
<td>0</td>
<td>(d_5)</td>
<td>49/18</td>
</tr>
<tr>
<td>(I_{16})</td>
<td>2· 4/18</td>
<td>0</td>
<td>5· 4/18</td>
<td>6· 7/18</td>
<td>7· 7/18</td>
<td>1· 3/18</td>
<td>0</td>
<td>(d_5)</td>
<td>49/18</td>
</tr>
<tr>
<td>(I_{17})</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>3· 7/18</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>(d_5)</td>
<td>231/18</td>
</tr>
<tr>
<td>(I_{18})</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>3· 4/18</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>(d_5)</td>
<td>231/18</td>
</tr>
</tbody>
</table>

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


