Chapter 2
Grouping of Questions From a Question Bank Using Partition-Based Clustering

ABSTRACT

During automatic test paper generation, it is necessary to detect percentage of similarity among questions and thereby avoid repetition of questions. In order to detect repeated questions, the authors have designed and implemented a similarity matrix-based grouping algorithm. Grouping algorithms are widely used in multidisciplinary fields such as data mining, image analysis, and bioinformatics. This chapter proposes the use of grouping strategy-based partition algorithm for clustering the questions in a question bank. It includes a new approach for computing the question similarity matrix and use of the matrix in clustering the questions. The grouping algorithm extracts n module-wise questions, compute $n \times n$ similarity matrix by performing $n \times (n-1)/2$ pair-wise question vector comparisons, and uses the matrix in formulating question clusters. Grouping algorithm has been found efficient in reducing the best-case time complexity, $O(n\times (n-1)/2 \log n)$ of hierarchical approach to $O(n \times (n-1)/2)$.

DOI: 10.4018/978-1-7998-3772-5.ch002

TERMINOLOGY USED

The terminology used is presented in the table below -

Table 1. Terminology used for question clustering

Term	Meaning
Subject (S)	S is a subject/paper offered in different semesters of a course.
Modules/Units	For each subject, there is a university prescribed syllabus which consists of different modules/units.
Question Bank (QB)	QB is a database which stores module wise questions with its details such as question- no, question-content, question-type, question- marks and question-answer-time
Q	Q is the total number of questions stored under a module
t_i	t_i refers to the total number of questions in which term i appears
$f\,req_{ij}$	$f\,req_{ij}$ is the frequency of term i in question j
maximum frequency ($max\,freq_{ij}$)	$max\,freq_{ij}$ is the maximum frequency of a term in question j
term frequency (tf_{ij})	tf_{ij} refers to the importance of a term *i* in question *j*. It is calculated using the formula: $tf_{ij} = freq_{ij}/\max freq_{ij}$
Inverse Document Frequency ($id\,f_i$)	$id\,f_i$ refers to the discriminating power of term i and is calculated as: $id\,f_i = log_2\ (Q/t_i)$
tf-idf weighting (Wij)	It is a weighting scheme to determine weight of a term in a question. It is calculated using the formula: $W_{ij} = tf_{ij} \times id\,f_i$
Question-Term-Set, T_i (question q_i)	A set of terms extracted from each question by performing its tokenization, stop word removal, taxonomy verb removal and stemming
Theshold, δ	User input threshold value to find the similarity

2. 2 PARTITION-BASED GROUPING ALGORITHMS FOR QUESTION CLUSTER FORMULATION

The similarity matrix computation has been carried out by using matrix representation of vectors which is a natural extension of existing Vector Space Model (VSM) (Jing, L., Ng, M. K.& Huang, J. Z. 2010; Turney, P. D., Pantel, P. 2010; Wong, S. K. M. and Raghavan, V.V. 1984). VSM is a popular information retrieval system implementation which facilitates representation of a set of documents as vectors in term space. Similarity matrix generates its