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 modulewise 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

Copyright © 2021, IGI Global. Copying or distributing in print or electronic forms without written permission of IGI Global is prohibited.

TERMINOLOGY USED

The terminology used is presented in the table below -

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 pre- scribed 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>i</i> in question <i>j</i> . It is calculated using the formula: $tf_{ij} = freq_{ij}/max freq_{ij}$
Inverse Document Frequency $(id f_i)$	<i>id</i> f_i refers to the discriminating power of term i and is calculated as: <i>id</i> $f_i = log_2(Q/t_i)$
tf-idf weighting (Wij)	It is a weighting scheme to determine weight of a term in a question. It is calculated using the formula: $W_{ij} = tf_{ij} \times idf_i$
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, \delta	User input threshold value to find the similarity

Table 1. Terminology used for question clustering

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

12 more pages are available in the full version of this document, which may be purchased using the "Add to Cart" button on the publisher's webpage: <u>www.igi-</u> <u>global.com/chapter/grouping-of-questions-from-a-question-</u> bank-using-partition-based-clustering/268461

Related Content

Methodologies and Technologies to Retrieve Information From Text Sources Anu Singhaand Phub Namgay (2018). *Modern Technologies for Big Data Classification and Clustering (pp. 99-123).* www.irma-international.org/chapter/methodologies-and-technologies-to-retrieve-informationfrom-text-sources/185980

Optimization of Mean and Standard Deviation of Multiple Responses Using Patient Rule Induction Method

Jin-Kyung Yangand Dong-Hee Lee (2018). *International Journal of Data Warehousing and Mining (pp. 60-74).*

www.irma-international.org/article/optimization-of-mean-and-standard-deviation-of-multipleresponses-using-patient-rule-induction-method/198974

Organizational Impact of Spatiotemporal Graph Convolution Networks for Mobile Communication Traffic Forecasting

Pan Ruifeng, Mengsheng Wang, Jindan Zhang, Brij Guptaand Nadia Nedjah (2025). *International Journal of Data Warehousing and Mining (pp. 1-19).*

www.irma-international.org/article/organizational-impact-of-spatiotemporal-graph-convolutionnetworks-for-mobile-communication-traffic-forecasting/368563

Modified Single Pass Clustering Algorithm Based on Median as a Threshold Similarity Value

Mamta Mittal, R. K. Sharma, V.P. Singhand Lalit Mohan Goyal (2017). *Collaborative Filtering Using Data Mining and Analysis (pp. 24-48).*

www.irma-international.org/chapter/modified-single-pass-clustering-algorithm-based-on-medianas-a-threshold-similarity-value/159493

Knowledge Management Portals for Empowering Citizens and Societies

Hakikur Rahman (2009). Social and Political Implications of Data Mining: Knowledge Management in E-Government (pp. 42-63).

www.irma-international.org/chapter/knowledge-management-portals-empowering-citizens/29064