Comparative Assessment of Data Quality Dimensions in Scientific Multimedia Indexing Process

Hamid Hassani*
PhD in Information Technology Management; Iranian Research Institute for Information Science and Technology (IranDoc); Tehran, Iran Email: hassani@alumni.irandoc.ac.ir

Azadeh Mohebi
Assistant Professor; Faculty of Information Technology; Iranian Research Institute for Information Science and Technology (IranDoc); Tehran, Iran Email: mohebi@irandoc.ac.ir

Mohammad Javad Ershadi
Information Technology Department; Iranian Research Institute for Information Science and Technology (IranDoc); Tehran, Iran; Email: ershadi@irandoc.ac.ir

Iranian Journal of Information Processing and Management

Received: 24, Jan. 2024 **Accepted: 25, Apr. 2024**

Iranian Research Institute for Information Science and Technology (IranDoc)
ISSN 2251-8223
eISSN 2251-8231
Indexed by SCOPUS, ISC, & LISTA
Special Issue | Winter 2025 | pp. 287-308
Exploring the Relationship Between Data Quality and Business Process Management
https://doi.org/10.22034/jipm.2024.2021178.1521

Abstract: Organizing a large volume of scientific multimedia data requires the use of appropriate indexing methods as one of the processes of information organization. Appropriate methods and algorithms are those that lead to the improvement of various aspects of quality in the process of organizing and retrieving information. For this reason, the purpose of this research is to identify the most important dimensions of data quality in the field of scientific multimedia indexing. In order to achieve this goal, a comparison of different dimensions of data quality has been made based on different criteria and the most important dimensions have been identified using Shannon entropy weighting approach and TOPSIS group ranking method. Also, using the correlation matrix, the intensity and direction of the relationship and correlation between the different dimensions of data quality have been evaluated. Based on the results of the first part of the research, the best ranks (priorities) were related to the data quality dimensions of recall, precision, completeness, appropriate amount of data, accuracy, relevancy, concise 1, consistency, concise 2, interpretability, value-added and accessibility, respectively. The results obtained from the second part of the research showed that

* Corresponding Author

the data quality dimensions of interpretability and relevancy had the highest correlation with the most important dimensions, i.e. recall and precision. As one of the implications of this research, it is possible to consider the measurement and evaluation of scientific multimedia data indexing methods based on different aspects of data quality and their importance.

Keywords: Data Quality, Scientific Multimedia Indexing, Prioritization and Ranking, Correlation and Relationship, Keyword Extraction

1. Introduction

Due to the advances in information and communication technologies, scientific organizations and institutions providing educational and scientific services are faced with a large amount of data (Guo et al., 2022; Martins, Gonçalves, & Branco, 2022). In order to provide access to the data, various retrieval services (Pandiaraja et al., 2022) using techniques such as artificial intelligence are needed. Each of these services consists of different activities and processes, which if done well lead to the improvement of the quality of the services. Indexing this data accurately is necessary for providing retrieval and access services (Goyal, Behera, & McGinnity, 2012). The scientific data come in different forms and formats. Some are textual, such as scientific articles and research reports while others are multimedia such as audio and video lectures used for educational purposes. Scientific multimedia data indexing includes several steps including extracting key frames from the video, extracting text from audio and images, preprocessing text, and eliminating potential noises (Hassani, Ershadi, & Mohebi, 2022). Since there are complex and distinct steps in scientific multimedia indexing, and the data from each step serves as the input for the subsequent steps, data quality issues arise throughout the entire process. Some of the previous studies have addressed these issues by considering limited dimensions of data quality at the end of the indexing process (Husain & Meena, 2019; Albahr, Che, & Albahar, 2021). Another study in this field has investigated various dimensions of data quality during the indexing process of scientific multimedia indexing (Hassani et al., 2023). In their study, nine different data quality dimensions are defined in the field of scientific multimedia indexing.

Since there are different data quality dimensions, and sometimes it is not possible to measure all of them, one approach is to consider the most important

ones. The main goal of this research is to compare the newly defined dimensions of data quality in the field of scientific multimedia indexing and to identify the most important dimensions. To achieve the ultimate goal of this research, three main objectives need to be considered: (1) comparing the data quality dimensions with each other and specifying the most important ones, (2) finding the relationships and correlation between different dimensions, (3) identifying dimensions whose value changes have a direct impact on the values of important dimensions. Thus, the following research questions need to be addressed:

- ◇ RQ1. How to compare the data quality dimensions and specify the most important ones?
- ◇ RQ2. What are the relationships and correlation between different dimensions?
- ◇ RQ3. What are the dimensions whose value changes have a direct impact on important dimensions?

In this article, the importance and correlation of data quality dimensions defined in Hassani et al. (2023) in the field of scientific multimedia data indexing are investigated. The dimensions are accuracy, value-added, relevancy, completeness, an appropriate amount of data, conciseness (with two definitions), consistency, interpretability, and accessibility. In addition to these dimensions, two well-known dimensions, precision and recall, are also considered to improve the quality of outputs and augment the final impact of the indexing process.

This paper is organized as follows: The theoretical background is reviewed in the next section. In the research method section, the main phases of the research and the steps of each are introduced. In the data analysis and findings section, the output of weighting steps, prioritization, and correlation and relationship between dimensions are presented. The last section of this article includes a discussion and conclusion and a presentation of scientific and practical suggestions.

2. Background

The issue of data quality has been investigated in various fields of study. In (Dakka et al., 2021), the issue of data quality in the field of health has been addressed. It has been covered in the field of machine learning (Priestley, O'Donnell, & Simperl,

2023), software engineering (Valverde et al., 2022), finance (Du & Zhou, 2012), and e-learning (Uppal, Ali, & Gulliver, 2018). The e-learning data developed in (Uppal, Ali, & Gulliver, 2018), is based on the SERVQUAL model (Parasuraman, Zeithaml, & Berry, 1988). The constituent dimensions of this model include (1) service dimension, consisting of five independent variables, “reliability”, “responsiveness”, “assurance”, “tangibility”, and “empathy”, (2) information dimension, including “learning content”, and (3) system dimension, including “course website”. This research suggests that in addition to “service”, it is critical to consider “information” and “system” quality to achieve an overall understanding of quality for e-learning systems.

Another field in which the issue of data quality has been investigated is organizing and retrieving scientific documents. Studies such as (Ershadi & Azizi, 2019) and (Ershadi et al., 2022) have investigated data quality and its various dimensions in this field. Data quality in organizing, indexing, and retrieving scientific multimedia data has received less attention compared to scientific textual data. However, some studies have been done in this field recently, for instance in (Hassani et al., 2023), new dimensions of data quality were defined and measured in the field of scientific lecture video indexing. Nine dimensions of data quality including accuracy, value-added, relevancy, completeness, appropriate amount of data, concise, consistency, interpretability and accessibility, were investigated. These dimensions were evaluated based on ten different criteria. The evaluation results showed that the well-known dimensions of data quality such as precision and recall are not superior in all criteria. For example, the dimension of completeness in the criterion of ease of implementation and the dimension of accuracy in the criterion of drill-down capability have had better results than other criteria. In (Rahman et al., 2024), the accessibility to different parts of a lecture video has been studied. This study uses artificial intelligence to generate visual and textual summaries of lecture video to improve navigation. The image summary is a subset of unique and important images obtained using image analysis. The textual summary is a collection of keywords selected by analyzing several factors such as font size, frequency, and time on screen. The framework developed in that research has been implemented in Videopoints lecture video portal which is available to educational institutions. In (Furini, Mirri, & Montangero,

2018), VLP, which stands for Video Lecture Playlist, was developed with the aim of improving the accessibility of disabled students to the content of lecture videos. In this research, three low-level audio/visual features, video segmentation and OCR analysis are used to "understand" the content of lecture videos. In this way, students search for a specific topic through keywords and the system finds all the pieces of lecture videos that cover the searched topic. These pieces are then provided through a playlist. In (Ghosh et al., 2022), the relevancy and usability dimensions were investigated so that an augmentation system was developed to identify off-topic concepts and link them to relevant video lecture segments, in order to provide a basic understanding of the concepts. Their system separated the video lectures by identifying the topical changes in the lectures using a technique based on word embedding. Video segments were indexed based on underlying concepts. The identification of off-topic concepts was done by modeling inter-concept relations in the semantic space. Then, the appropriate video segments for each off-topic concept were fetched. In addition to the system evaluation, feedbacks from some research scholars showed the usability of this system. Other previous studies in the field of summarizing, indexing and keyword extraction from scientific lecture videos have used well-known data quality dimensions such as precision and recall to evaluate the methods and algorithms (Sun & Tian, 2022; Davila et al., 2021; Abhilash et al., 2021). However, despite the appropriateness of the well-known dimensions of data quality such as precision and recall, for a more comprehensive evaluation of the indexing algorithms and methods, other aspects of data quality should also be considered in this field.

Although considering all data quality dimensions when developing indexing methods and algorithms may improve the entire indexing process, especially when dealing with multimedia, it may not be feasible to evaluate all dimensions due to certain limitations. Sometimes it is not possible to measure every dimension due to the lack of data or imprecise information. Thus, it is important to identify the most effective and informative dimensions and use them when there are limitations in measuring all dimensions. In this research, a ranking and prioritization approach is proposed for data quality dimensions in multimedia indexing. The dimensions are as proposed in (Hassani et al., 2023). In Table 1, the definitions of these dimensions are presented in the field of scientific multimedia indexing.

Table 1. Definitions of data quality dimensions in the field of scientific multimedia indexing (Hassani et al., 2023)

Data quality dimensions	Definitions
Accuracy	The correctness of the output words from ASR1 and OCR2 systems
Value-added	The extracted keywords have the advantage of speeding up and improving the video retrieval process. The usefulness of extracted key phrases for information retrieval.
Relevancy	The relevance of the extracted key phrases to the subject of the lecture video, according to their position in the final list of key phrases.
Completeness	Considering the four features including information content, time, Interruption of the video, and video presentation language.
An appropriate amount of data	Appropriateness of the number of key phrases according to the duration and volume (number of extracted words from OCR and ASR) of a video.
Conciseness	Definition 1: Not imposing useless information to the algorithms of extracting key phrases. Definition 2: The extent to which the algorithm generates distinct key phrases (in terms of concept and meaning).
Consistency	Minimal variations in the results of precision and recall values
Interpretability	The extent to which key phrases are understandable and meaningful to users (for phrases that are not in the original).
Accessibility	The ability of the algorithm to show the first occurrence of the extracted key phrase

3. Research method

This research is conducted in two main phases. In the first phase, the study investigates the significance of data quality dimensions as defined in the field of multimedia indexing, based on eight different criteria. These dimensions have been defined in Table 1. The first phase consists of two steps: evaluating the weight of criteria and prioritizing data quality dimensions based on these criteria. For this purpose, the opinions of five experts in the field of automatic indexing of multimedia and data quality are being collected. Among the selection

1. Automatic Speech Recognition
2. Optical Character Recognition

criteria for these experts was their knowledge and experience in the fields of information retrieval, multimedia indexing, and data quality. The first step in this phase, i.e., weighting, is based on one of the most well-known and widely used weighting methods, Shannon entropy. For this purpose, a matrix in the form of a questionnaire was sent to each of the experts. The rows of this matrix included data quality dimensions, and the columns included evaluation criteria. Based on this, and according to the evaluation criteria, the experts assigned values to each of the dimensions of data quality. Then, the prioritization step is done based on the TOPSIS (Hwang & Yoon, 1981) group ranking method using the weights obtained from the first step.

The evaluation criteria, which are weighted in the first step based on the Shannon entropy method, are shown in Table 2. This table also describes the criteria and their directions. The meaning of the criterion's direction is whether higher values indicate desirability or not. For criteria where higher values are more favorable, a positive sign (+) is used in the direction column.

Table 2. Evaluation criteria and their description

Code	Evaluation criteria	Description of criteria	The direction of the criterion
A1	Clarity of definition	To what extent the definition and objectives of the dimension are clearly stated	+
A2	Ease of implementation	To what extent the dimension can be easily implemented	+
A3	Drill-down capability	The extent to which the dimension is considered at the beginning of the indexing process	+
A4	Adaptivity	The extent to which the dimension can be defined in a context other than multimedia indexing	+
A5	Interpretability	How much definition of the dimension is interpretable for humans	+
A6	Acceptability	Existence of a threshold limit for improving dimension values	+
A7	Reportability	The extent of providing sufficient information for reporting	+
A8	Quantifiability	The extent to which dimension values can be expressed with numbers	+

4. Data analysis and findings

The findings obtained from the first phase are shown in Tables 3-5. Table 3 shows the results of weighting calculations based on Shannon entropy for each expert. Based on these findings, from the first four experts' point of view, the two criteria A1 and A8, i.e., quantifiability and clarity of definition, gain more weight for evaluating data quality dimensions for multimedia indexing. From the fifth expert's point of view, the two criteria A1 and A4, i.e., clarity of definition and adaptivity, gain more weight for evaluating data quality dimensions. The point that should be mentioned is that the clarity of definition was one of the criteria with the highest weight based on the opinion of all five experts.

Table 3. The calculated weights of the criteria based on Shannon Entropy

criteria code	The weight of criteria according to experts' opinions				
	1	2	3	4	5
A1	0.1644	0.1545	0.1606	0.2087	0.1788
A2	0.0881	0.0627	0.0900	0.1109	0.0689
A3	0.1528	0.1198	0.0752	0.1018	0.1283
A4	0.1617	0.1383	0.1505	0.1341	0.1689
A5	0.1430	0.1371	0.1450	0.1018	0.1046
A6	0.1189	0.0882	0.0960	0.0855	0.1077
A7	0.1459	0.1371	0.1492	0.1329	0.1046
A8	0.1685	0.1583	0.1722	0.1534	0.1641

The findings of the second step, which include ranking based on the TOPSIS method, are shown in Table 4. In this table, the geometric mean of Euclidean distances from the positive and negative ideal solutions and the similarity index are presented as the outputs of the TOPSIS group method. Based on the results of the similarity index, the higher the value of this index, the higher the priority of data quality dimensions. In this table, the "distance type" column, d+ represents the Euclidean distance from the positive ideal solution, and d- represents the Euclidean distance from the negative ideal solution. The following equation is used to calculate the similarity index:

$$\text{Similarity index} = \frac{d_i^-}{d_i^+ + d_i^-}$$

Where, d_i^+ represents the geometric mean of Euclidean distances between the values of each dimension and the positive ideal solution, and d_i^- represents the geometric mean of Euclidean distances between the values of each dimension and the negative ideal solution.

Table 5 shows the final prioritization of data quality dimensions based on similarity index. Based on this, dimensions recall, precision, completeness, appropriate amount of data, accuracy, relevancy, concise 1, consistency, concise 2, interpretability, value-added, and accessibility have received the highest priority respectively.

Table 4. Geometric mean of distances, similarity index and priority of data quality dimensions based on TOPSIS group method

Dimension	Distance type	Euclidean distances from the positive and negative ideal solutions according to experts					Geometric mean of distances	Similarity index	Priority
		1	2	3	4	5			
Precision	d+	0.0474	0.0238	0.0110	0.0252	0.0228	0.0235	0.7802	2
	d-	0.0833	0.0785	0.0829	0.0869	0.0865	0.0836		
Recall	d+	0.0474	0.0238	0.0110	0.0201	0.0228	0.0225	0.7909	1
	d-	0.0833	0.0785	0.0829	0.0955	0.0865	0.0851		
Accuracy	d+	0.0545	0.0513	0.0525	0.0539	0.0488	0.0522	0.5304	5
	d-	0.0666	0.0551	0.0507	0.0621	0.0616	0.0589		
Value-added	d+	0.0915	0.0812	0.0798	0.0756	0.0711	0.0796	0.2007	11
	d-	0.0144	0.0126	0.0140	0.0392	0.0316	0.0199		
Relevancy	d+	0.0560	0.0536	0.0482	0.0713	0.0522	0.0558	0.5192	6
	d-	0.0729	0.0613	0.0635	0.0474	0.0589	0.0602		
Completeness	d+	0.0485	0.0471	0.0392	0.0374	0.0423	0.0427	0.6076	3
	d-	0.0666	0.0559	0.0642	0.0780	0.0681	0.0662		
Appropriate Amount of Data	d+	0.0636	0.0525	0.0455	0.0465	0.0475	0.0507	0.5456	4
	d-	0.0579	0.0539	0.0571	0.0793	0.0593	0.0609		
Concise 1	d+	0.0572	0.0531	0.0530	0.0560	0.0488	0.0535	0.4837	7
	d-	0.0506	0.0422	0.0430	0.0602	0.0575	0.0502		

Dimension	Distance type	Euclidean distances from the positive and negative ideal solutions according to experts					Geometric mean of distances	Similarity index	Priority
		1	2	3	4	5			
Concise 2	d+	0.0809	0.0739	0.0745	0.0827	0.0529	0.0721	0.3657	9
	d-	0.0407	0.0357	0.0379	0.0374	0.0602	0.0416		
Consistency	d+	0.0652	0.0592	0.0497	0.0601	0.0622	0.0590	0.4434	8
	d-	0.0505	0.0404	0.0526	0.0478	0.0448	0.0470		
Interpretability	d+	0.0806	0.0718	0.0752	0.0821	0.0723	0.0763	0.2650	10
	d-	0.0297	0.0246	0.0235	0.0271	0.0336	0.0275		
Accessibility	d+	0.0970	0.0860	0.0858	0.1012	0.0908	0.0919	0.0000	12
	d-	0.0000	0.0000	0.0000	0.0000	0.0079	0.0000		

Table 5. Final prioritization of data quality dimensions

Priority	Data quality dimension	Similarity index
1	Recall	0.7909
2	Precision	0.7802
3	Completeness	0.6076
4	Appropriate Amount of Data	0.5456
5	Accuracy	0.5304
6	Relevancy	0.5192
7	Concise 1	0.4837
8	Consistency	0.4434
9	Concise 2	0.3657
10	Interpretability	0.2650
11	Value-added	0.2007
12	Accessibility	0.0000

In the second phase of this research, the correlation between data quality dimensions in various data sets is examined. The values of data quality dimensions considered in this section are obtained from the application of the LVTIA algorithm, one of the indexing algorithms for scientific lecture videos, on

four different datasets (Hassani, Ershadi, & Mohebi, 2022). The first dataset consists of 20 English scientific lecture videos from edX, an American Massive Open Online Course (MOOC) provider created by Harvard and MIT. The second dataset consists of 20 Persian scientific lecture videos from Faradars, one of the online education platforms in Iran. The third dataset consists of 20 English scientific lecture videos from tele-TASK, one of the online education platforms. The fourth dataset, E-learning, consists of 60 Persian scientific lecture videos from virtual education departments of various universities in Iran. In In fact, this study investigates how increasing or decreasing the values of one dimension of data quality affects the values of other dimensions. There are three types of correlation: positive (direct), negative (inverse), and no correlation. In a positive correlation, an increase in the values of one variable leads to an increase in the values of another variable. Negative correlation exists when an increase in one variable results in a decrease in the value of another variable. When there is no linear relationship between the values of two variables, they are considered uncorrelated or not linearly correlated (Aminpour, 2018). The intensity and direction of correlation is shown by the correlation coefficient and its values are in the range of [-1, +1]: (Lotfabadi, 1996):

- ◇ +0.85 to +0.99 positive and very strong correlation;
- ◇ +0.70 to +0.85 positive and strong correlation;
- ◇ +0.40 to +0.70 positive and relatively strong correlation;
- ◇ +0.20 to +0.40 positive and relatively weak correlation;
- ◇ +0.10 to +0.20 positive and very weak correlation;
- ◇ -0.10 to +0.10 random correlation;
- ◇ -0.10 to -0.20 negative and very weak correlation;
- ◇ -0.20 to -0.40 negative and relatively weak correlation;
- ◇ -0.40 to -0.70 negative and relatively strong correlation;
- ◇ -0.70 to -0.85 negative and strong correlation;
- ◇ -0.85 to -0.99 negative and very strong correlation.

In the following, the results of the correlation matrix for six dimensions of data quality, including accuracy, relevancy, Concise 1 (conciseness), interpretability, precision, and recall are presented in four data sets of video lectures.

The reason for choosing these dimensions was that in each data set for each video there was a value corresponding to these dimensions. In other words, for some dimensions of data quality such as completeness, consistency, and accessibility, there is no data for each video. For example, according to the definition of consistency, for each data set, only one value is calculated for this dimension, which cannot be used in correlation calculations. Figures 1 to 4 show the correlation matrix heat maps outputs for four datasets: edX, Faradars, tele-TASK, and E-learning.

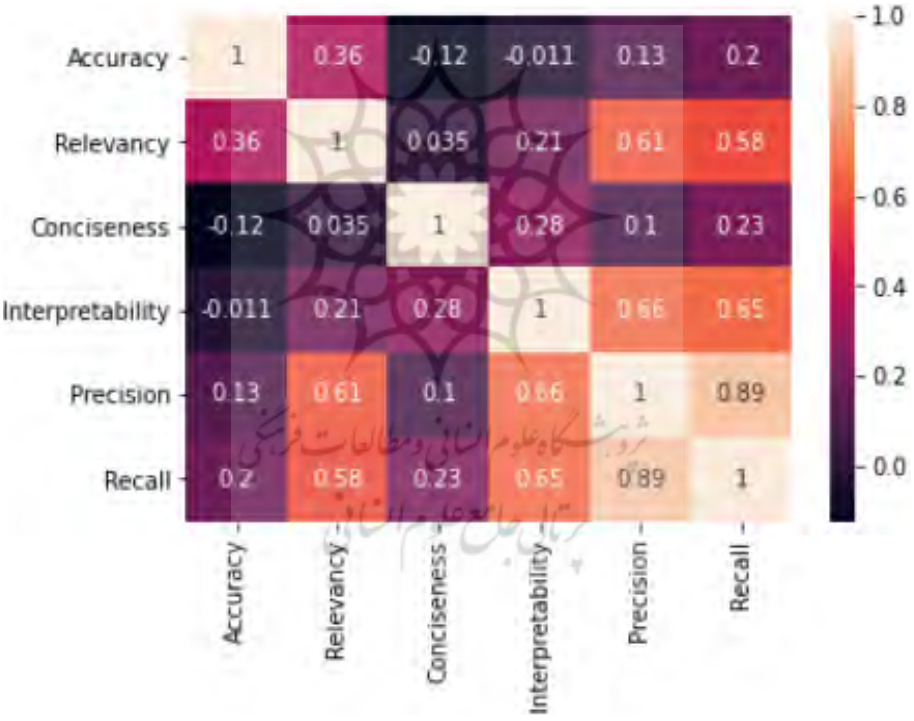

Fig. 1. Correlation matrix heatmap output for the edX dataset

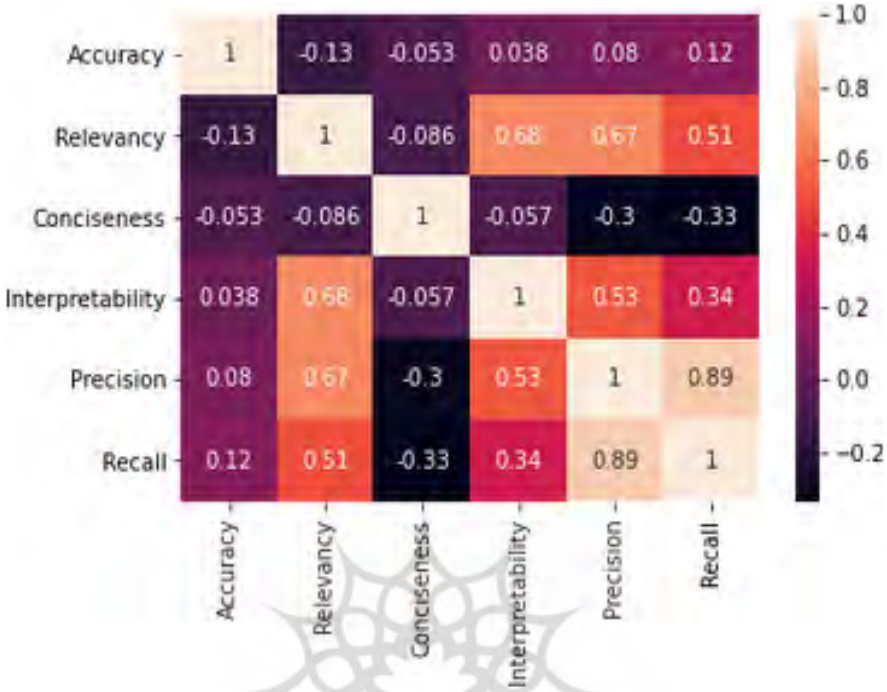

Fig. 2. Correlation matrix heatmap output for the Faradars dataset

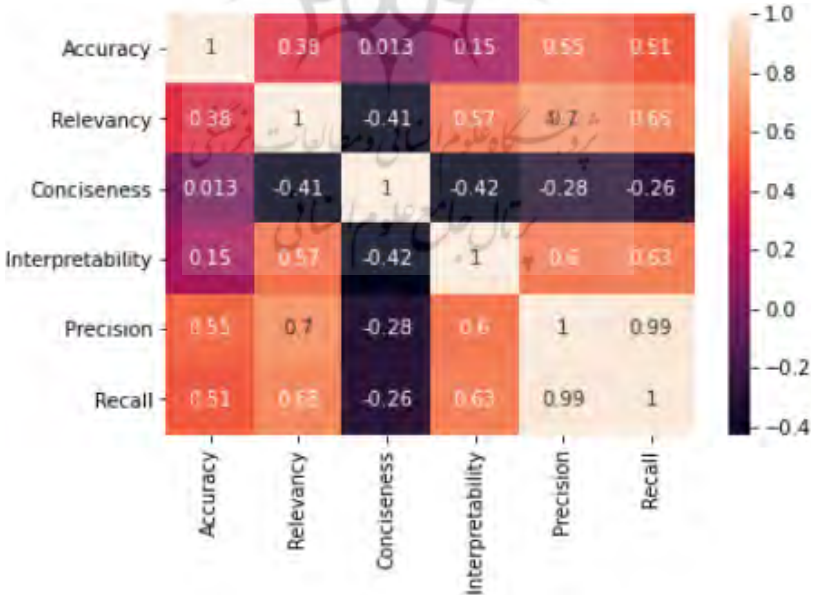

Fig. 3. Correlation matrix heatmap output for the tele-TASK dataset

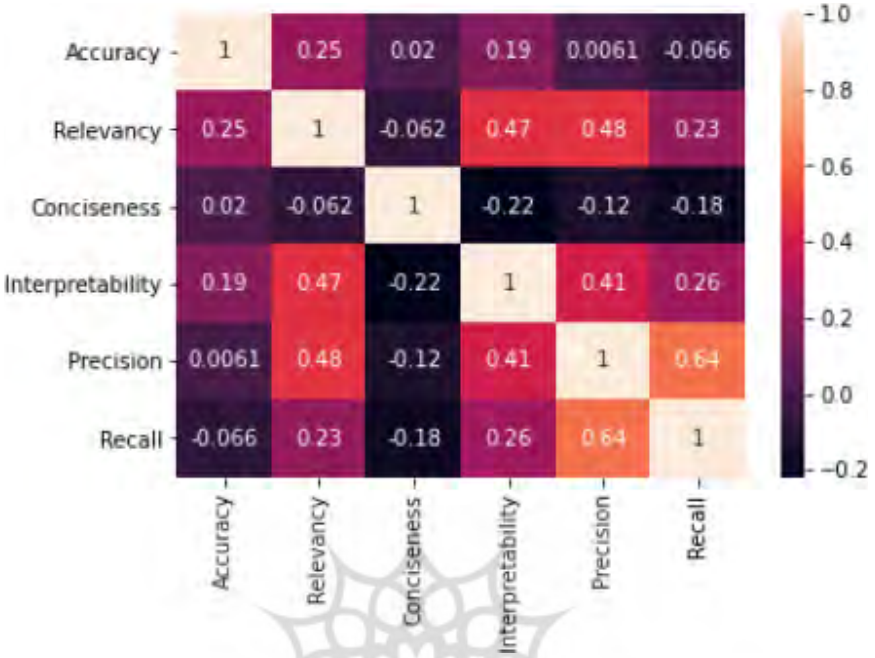

Fig. 4. Correlation matrix heatmap output for the E-learning dataset

As depicted in Figures 1, 2, and 3, after considering the random and relatively weak correlation values between the accuracy dimension and other dimensions in these three datasets, it is evident that the impact of this dimension on other dimensions is insignificant. Similarly, the effect of changes in the values of other dimensions on this dimension is also deemed insignificant. In other words, it can be said that the increase or decrease of this dimension is independent of other dimensions and is not affected by the increase or decrease of the values of other dimensions.

In all four datasets (except for recall in the E-learning dataset), the correlation between the relevancy dimension and precision and recall falls within the range of relatively strong and positive relationships. The reason for the emergence of such a relationship could be attributed to the nature of defining the relevance dimension. In the definition of this dimension, two parameters are considered: the degree of relevancy and the position of each key phrase in the final list of key phrases. It is possible to achieve a high value in a video for the precision and recall dimensions. This will enhance the relevancy parameter and ultimately lead

to an increase in the relevancy dimension. It is also possible that reducing the values of precision and recall dimensions has a direct effect on decreasing the value of the relevancy parameter, ultimately leading to a reduction in the relevancy dimension's value. In addition, from this perspective, we can examine the positive and relatively strong relationship between the relevance dimension and precision and recall. An increase in the relevance dimension will enhance precision and recall, as they are the most crucial dimensions derived from the weighting and prioritization process.

For the interpretability dimension, the intensity and direction of the correlation between this dimension and precision and recall in all datasets are relatively strong and positive, except for recall in Faradars and E-learning. The justification for this level of correlation could be that the higher the precision and recall dimensions of a video indexing method, the more closely the keyphrases align with the ground truth and the main keyphrases. In the same way, if the values of precision and recall decrease, the interpretability dimension also decreases.

Interpretability, in Figures 2, 3, and 4 has a positive and relatively strong correlation with relevancy. Two reasons can be put forward for this correlation between relevancy and interpretability. Considering that in the definition of the relevancy dimension, there is a parameter called the degree of relevancy, which is affected by the user, and also the interpretability dimension value is determined by the user. User may consider the meaning of the term "meaningful" (in the definition of interpretability) as being relevant. In other words, it is possible that the user may consider an extracted keyphrase as being meaningful and assign a high score to it, while he/she may also place a high value for the relevancy as well. However, a meaningful keyphrase may be completely unrelated to the topic of the video, and one should not make the mistake that every meaningful keyphrase is necessarily relevant. Therefore, the first reason for the existence of such a correlation between the relevancy and interpretability dimensions may be due to the misunderstanding of the difference between "meaningful" and "relevant" for the user.

Figure 2 shows the correlations of the conciseness dimension ranging from random to relatively weak. The important point is that all the correlations between this dimension and other dimensions are negative. In other words, the increase of conciseness is associated with the decrease of other dimensions, and vice versa

the decrease of conciseness is associated with the increase of other dimensions, having a relatively weak effect.

As shown in Figure 3, there are the highest correlations (in terms of intensity) between dimensions in this data set compared to other data sets. According to the heat map presented in this Figure, all dimensions, have a relatively strong correlation with at least two other dimensions. In this data set, precision and recall have relatively strong and positive correlations with all dimensions except conciseness, and their changes are associated with direct changes in other dimensions. The correlation and inverse relationship between the precision and recall dimensions with the conciseness dimension can be analysed in such a way that to improve the values of precision, recall and finally the F measure, the performance of the developed algorithm should be such that it leads to the lowest value of the conciseness dimension.

As shown in Figure 4, the accuracy dimension correlations range from random to relatively weak. In this Figure, the correlations of the relevancy dimension are also in the range of random to relatively strongly positive. Relatively strong correlations for this dimension are related to interpretability and precision dimensions and in a positive direction. There are random to relatively weak negative correlations for the conciseness dimension, so that the correlations of this dimension with relevancy, interpretability, precision and recall are in the opposite direction. The reason for the inverse direction of these correlations is that for the conciseness dimension, lower values are more favorable, while for the other mentioned dimensions, higher values are favorable.

In Figure 5, the mutual correlations between two dimensions are shown for different datasets. As shown in Figure 5, in all cases except for the correlation between conciseness and interpretability, the correlations are similar or closely similar. This issue shows the relatively similar behavior of dimensions in different data sets. Of course, there are cases such as the correlation between accuracy and relevancy, conciseness and recall, accuracy and precision, in which the correlation values in one data set is relatively different from the others. This can be due to various reasons such as the nature of the data and the values obtained from the data quality dimensions for each video in a dataset. The ultimate goal of studying the correlations between different dimensions is to discover the mutual

relationship between them, and consequently to determine that the measurement of which dimension can clarify the values of others. If the correlation between two dimensions cannot be revealed based on the above experiments then it is clear that we cannot predict the behavior of one based on the other. For instance, the correlation between conciseness and interpretability is different in four datasets. While, the correlation between relevancy and interpretability, and relevancy and precision, and relevancy and is relatively strong and the same in all datasets, meaning that knowing the value of one dimension may reveal the value of another.

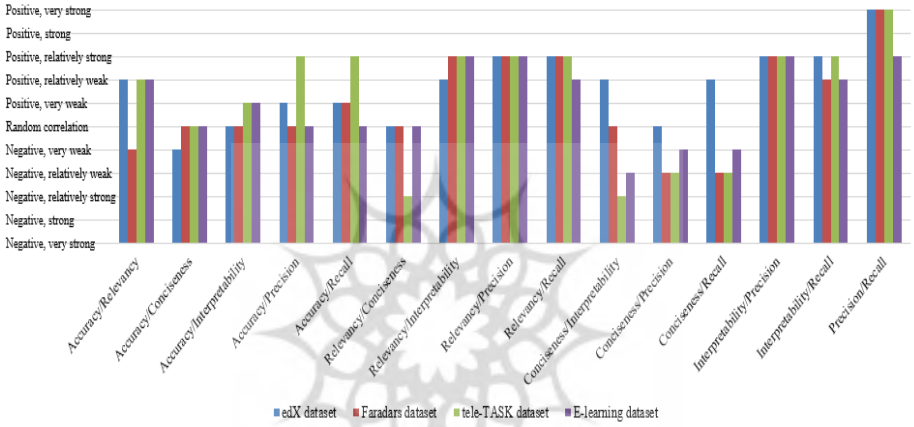

Fig. 5. Mutual correlations between each two dimensions in the data sets

5. Discussion and conclusion

Since the process of indexing scientific multimedia data consists of several steps (Yang & Meinel, 2014), and data from each step moves as input to the next step, poor quality of data can negatively affect the overall performance of the indexing method. Therefore, it is necessary to pay attention to the issue of data quality throughout the entire process and to consider indexing methods and algorithms from various perspectives. Various data quality dimensions have been defined in the field of scientific multimedia indexing (Hassani et al., 2023), but it may not be possible to measure all of them. Based on this issue and the main goal of this research, which is to compare the newly defined dimensions of data quality in the field of scientific multimedia indexing and to identify the most important dimensions, three research questions were raised in the introduction section. The first question pertains to comparing various dimensions and identifying the most crucial ones.

To answer the first question, the data quality dimensions were ranked using the Shannon entropy weighting method and the TOPSIS group ranking method to determine the most important dimensions. The second question pertains to the relationship and correlation between the dimensions, while the third question focuses on which dimension's increase or decrease directly influences the values of the most critical dimensions. To answer the second and third questions, the correlation matrix is used to investigate the relationships between different data quality dimensions. It has been shown that certain dimensions can either increase or decrease the values of the most critical dimensions. This approach can be beneficial when it is impractical to measure the most critical dimensions. By measuring and enhancing the less critical dimensions, the most important dimensions can be indirectly improved. For example, as shown in Figure 3, the correlation between the relevancy dimension and precision and recall is positive and relatively strong. Therefore, any increase or decrease in this dimension may result in a corresponding increase or decrease in the two metrics.

Compared to previous research in this field (Jiang, Miao, & Li, 2017; Albahr, Che, & Albahar, 2021), where only well-known dimensions of data quality such as precision and recall were investigated, this study explores various dimensions of data quality. Although these dimensions have been defined in (Hassani et al., 2023), their importance, correlation, and relationship have been examined in detail in this research. In other words, one aspect of contribution in this research has been to specify the most important dimensions of data quality in the field of scientific multimedia indexing.

Based on the ranking results in this research, the significance of various dimensions of data quality was determined. The best ranks were recall, precision, completeness, appropriate amount of data, accuracy, relevancy, conciseness 1, consistency, conciseness 2, interpretability, value-added, and accessibility, respectively. The following are some scientific and practical suggestions for future research:

- ◇ Measurement and evaluation of scientific multimedia data indexing methods based on different aspects of data quality and their importance;
- ◇ Customization of these dimensions in other fields of artificial intelligence;

◇ Implementation of more important data quality dimensions in databases, platforms and organizations that provide educational multimedia content.

References

Abhilash, R. K., Anurag, C., Avinash, V., & Uma, D. (2021). Lecture video summarization using subtitles. In *2nd EAI International Conference on Big Data Innovation for Sustainable Cognitive Computing: BDCC 2019* (pp. 83-92). Springer International Publishing. DOI: 10.1007/978-3-030-47560-4_7

Albahr, A., Che, D., & Albahar, M. (2021). A novel cluster-based approach for keyphrase extraction from MOOC video lectures. *Knowledge and Information Systems*, *63* (7), 1663-1686. DOI: 10.1007/s10115-021-01568-2

Aminpour, H. (2018). *Descriptive statistics in psychology and educational sciences*. Tehran: Payam Noor University.

Dakka, M. A., Nguyen, T. V., Hall, J. M. M., Diakiw, S. M., VerMilyea, M., Linke, R., ... & Perugini, D. (2021). Automated detection of poor-quality data: case studies in healthcare. *Scientific Reports*, *11* (1), 18005. DOI: 10.1038/s41598-021-97341-0

Davila, K., Xu, F., Setlur, S., & Govindaraju, V. (2021). Fcn-lecturenet: extractive summarization of whiteboard and chalkboard lecture videos. *IEEE Access*, *9*, 104469-104484. DOI: 10.1109/ACCESS.2021.3099427

Du, J., & Zhou, L. (2012). Improving financial data quality using ontologies. *Decision Support Systems*, *54* (1), 76-86. DOI: 10.1016/j.dss.2012.04.016

Ershadi, M. J., & Azizi, A. (2019). *Measurement and analysis of data quality metrics in the process of registering theses/dissertations of domestic graduates*. Tehran: Iranian research institute for information science and technology.

Ershadi, M. J., Ershadi, M. M., Fakhrazadeh, A., & Ghazizadeh. S. (2022). *Providing implementation solutions to improve the quality of theses/ dissertations information in the registration system based on data mining techniques*. Tehran: Iranian research institute for information science and technology.

Furini, M., Mirri, S., & Montangero, M. (2018, January). Topic-based playlist to improve video lecture accessibility. In *2018 15th IEEE Annual Consumer Communications & Networking Conference (CCNC)* (pp. 1-5). IEEE. DOI: 10.1109/CCNC.2018.8319246

Ghosh, K., Nangi, S. R., Kanchugantla, Y., Rayapati, P. G., Bhowmick, P. K., & Goyal, P. (2022). Augmenting video lectures: Identifying off-topic concepts and linking to relevant video lecture segments. *International Journal of Artificial Intelligence in Education*, *32* (2), 382-412. DOI: 10.1007/s40593-021-00257-z

Goyal, P., Behera, L., & McGinnity, T. M. (2012). A context-based word indexing model for document summarization. *IEEE Transactions on Knowledge and Data Engineering*, *25* (8), 1693-1705. DOI: 10.1109/TKDE.2012.114

Guo, T., Bai, X., Tian, X., Firmin, S., & Xia, F. (2022). Educational anomaly analytics: features, methods, and challenges. *Frontiers in big Data*, *4*, 811840. DOI: 10.3389/fdata.2021.811840

Hassani, H., Ershadi, M. J., & Mohebi, A. (2022). LVTIA: A new method for keyphrase extraction from scientific video lectures. *Information Processing & Management*, *59* (2), 102802. DOI: 10.1016/j.ipm.2021.102802

Hassani, H., Mohebi, A., Ershadi, M. J., & Jalalimanesh, A. (2023). A novel data quality framework for assessment of scientific lecture video indexing. *Library Hi Tech*. DOI: 10.1108/LHT-02-2023-0074

Husain, M., & Meena, S. M. (2019, February). Multimodal fusion of speech and text using semi-supervised LDA for indexing lecture videos. In *2019 National Conference on Communications (NCC)* (pp. 1-6). IEEE. DOI: 10.1109/NCC.2019.8732253

Hwang, C. L., & Yoon, K. (1981). Methods for multiple attribute decision making. *Multiple attribute decision making: methods and applications a state-of-the-art survey*, 58-191. DOI: 10.1007/978-3-642-48318-9_3

Jiang, Z., Miao, C., & Li, X. (2017). Application of keyword extraction on MOOC resources. *International Journal of Crowd Science*, *1* (1), 48-70. DOI: 10.1108/IJCS-12-2016-0003

Lotfabadi, H. (1996). *Assessment and measurement in educational sciences and psychology*. Mashhad: Samt.

Martins, J., Gonçalves, R., & Branco, F. (2022). A bibliometric analysis and visualization of e-learning adoption using VOSviewer. *Universal Access in the Information Society*, 1-15. DOI: 10.1007/s10209-022-00953-0

Pandiaraja, P., Boopesh, K. B., Deepthi, T., Laksmi Priya, M., & Noodhana, R. (2022, February). An analysis of document summarization for educational data classification using NLP with machine learning techniques. In *International Conference on Computing in Engineering & Technology* (pp. 127-143). Singapore: Springer Nature Singapore. DOI: 10.1007/978-981-19-2719-5_12

Parasuraman, A., Zeithaml, V. A., & Berry, L. L. (1988). Servqual: A multiple-item scale for measuring consumer perc. *Journal of retailing*, *64* (1), 12.

Priestley, M., O'donnell, F., & Simperl, E. (2023). A survey of data quality requirements that matter in ML development pipelines. *ACM Journal of Data and Information Quality*, *15* (2), 1-39. DOI: 10.1145/3592616

Rahman, M. R., Koka, R. S., Shah, S. K., Solorio, T., & Subhlok, J. (2024). Enhancing lecture video navigation with AI generated summaries. *Education and Information Technologies*, *29* (6), 7361-7384. DOI: 10.1007/s10639-023-11866-7

Sun, F., & Tian, X. (2022). Lecture video automatic summarization system based on DBNet and Kalman filtering. *Mathematical Problems in Engineering*, *2022* (1), 5303503. DOI: 10.1155/2022/5303503

Uppal, M. A., Ali, S., & Gulliver, S. R. (2018). Factors determining e-learning service quality. *British journal of educational technology*, *49* (3), 412-426. DOI: 10.1111/bjet.12552

Valverde, C., Marotta, A., Panach, J. I., & Vallespir, D. (2022). Towards a model and methodology for evaluating data quality in software engineering experiments. *Information and Software Technology*, *151*, 107029. DOI: 10.1016/j.infsof.2022.107029

Yang, H., & Meinel, C. (2014). Content based lecture video retrieval using speech and video text information. *IEEE transactions on learning technologies*, *7* (2), 142-154. DOI: 10.1109/TLT.2014.2307305

Hamid Hassani

Hamid Hassani has a PhD in Information Technology (IT) Management from the Iranian Research Institute for Information Science and Technology (IranDoc). Now he is a researcher at this research institute. His research interests include indexing, artificial intelligence, information systems, digital transformation, business continuity management (BCM), business process management, and data quality management.

Azadeh Mohebi

Azadeh Mohebi has a PhD in Systems Design Engineering, graduated in 2009 from University of Waterloo in Canada. She is currently an assistant professor at Iranian Research Institute for Information Science and Technology (IranDoc). Her main research themes are dedicated to developing intelligent decision support systems, human-computer interaction, and natural language processing.

Mohammad Javad Ershadi

Born in 1983, he holds a doctorate in Industrial Engineering from Iran University of Science and Technology, where he graduated in 2015. He received his MSc in Industrial Engineering from Sharif University of Technology in 2009 and his BSc in Industrial Engineering from Isfahan University of Technology in 2005. He began his academic career in 2015 as a faculty member at IranDoc and is currently an associate professor in the IT Management Research Group. His work at IranDoc focuses on applying quality engineering and management principles to information science and services. His research interests include statistical quality control (SQC), total quality management (TQM), business process reengineering (BPR), optimization, meta-heuristic algorithms, systems analysis, and data mining.