Rough Set Theory Approach for Classifying Multimedia Data

M. Nordin A. Rahman, Yuzarimi M. Lazim, Farham Mohamed, Suhailan Safei, Sufian Mat Deris, and M. Kamir Yusof

Faculty of Informatics, Universiti Sultan Zainal Abidin, Gong Badak Campus 21300 Kuala Terengganu, Malaysia
mohdnabd@unisza.edu.my, ayulazim@yahoo.com.my

Abstract. The huge size of multimedia data requires for efficient data classification and organization in providing effective multimedia data manipulation. Those valuable data must be captured and stored for potential purposes. One of the main problems in Multimedia Information System (MIS) is the management of multimedia data. As a consequence, multimedia data management has emerged as an important research area for querying, retrieving, inserting and updating of these vast multimedia data. This research considers the rough set theory technique to organize and categorize the multimedia data. Rough set theory method is useful for exploring multimedia data and simplicity to construct multimedia data classification. Classification will help to improve the performance of multimedia data retrieving and organizing process.

Keywords: Rough set theory, multimedia data management, approximation, classification, data clustering.

1 Introduction

Everyone deals with multimedia data at every walk of lives. Multimedia data consist of texts, graphics, animations, video, sounds, music etc. People are working with multimedia data and surrounded by them. Therefore, there are many issues and challenges faced by multimedia data providers to fulfill the user requirements. One of the issues is to organize and classify the huge multimedia data so that the information can be obtained easily at any point of time. An efficient multimedia data management is highly required because it will improve the process of multimedia information discovery especially for decision making application, business marketing, intelligent system, etc [1]. To do so, multimedia data management is a tool required to manage and maintain huge multimedia data.

Rough set theory is an effective tool for classification applications and, it has been introduced by Pawlak [2] [3]. Rough set theory is a mathematical tool that can be used for processing and analyzing of inexact, uncertain, and vague datasets. It is an extension of set theory for study of the intelligent system characterized by insufficient and incomplete information [3]. Various efforts have been made to improve the efficiency and effectiveness of classification with rough sets [4]. Practically, rough set theory has been applied to the number of application domains such as medical

J.M. Zain et al. (Eds.): ICSECS 2011, Part II, CCIS 180, pp. 116–124, 2011.

diagnosis, engineering reliability, expert systems, empirical study of materials data, machine diagnosis, business failure prediction, activity-based travel modeling, travel demand analysis, solving linear programming and data mining [5].

The intention of this research is to introduce a new representation for multimedia data management as an information system by integrating rough set theory elements such as decision table and approximation. This paper is organized as follows: Section 2 describes the related work on all the issue in multimedia data management while Section 3 gives details explanation on basic concepts of rough set theory. Section 4 demonstrates the proposed framework and Section 5 draws the conclusion.

2 Multimedia Data Management

Multimedia is defined as combination of more than one media; they may be of two types, static and dynamic media. Text, graphics and images are categorized as static media, while objects like animation, music, audio, speech, and video are categorized as dynamic media [6]. Multimedia data contains an enormous amount of information. This information is in the form of identifiable "features" in the multimedia data. For example, video data contains timing data that can be used to track the movement of an object from frame to frame or to identify transitions between scenes. Similarly, audio data contains certain identifiable features such as words, sound, pitches, and silent periods as well as timing information [7].

Multimedia data is attractive, informative and stored in compact format. It has been used by various categories of user; from technical up to management levels. The growing of consumer demand for multimedia information makes sophisticated technology is needed in multimedia data management. The applications are including representing, modeling, indexing, retrieving and manipulating. The explosion of multimedia content in various aspects such as in databases, broadcast, steaming media, etc. has generated new requirements for more effective access to these global information repositories. Multimedia data requires for a huge storage area and each media type requires different methods to store and retrieve.

The major issues that related to multimedia data management system are multimedia data modeling, huge capacity storage management, information retrieval capabilities, media integration and presentation. Multimedia database management system (MDMS) is developing in purpose to fulfill this requirement. MDMS supports facilities for the indexing, storage, retrieval and provides a suitable environment for using and managing multimedia data [8]. Technique of indexing and classification in multimedia data is created in order to ease the query processing.

Another study [9] designed a software tool known as web-based multimedia news management system (MNMS) which consist of a collection system, website management system, workflow system and publishing system. The MNMS tool enables delivery of TV and newspaper content trough ITV with internet connection. The integrated both publications are identify as multimedia interactive digital news. MNMS provides a service and application to allow collaboration and communication among journalists, reporters, editors and designers from multinational publication company

that distributed around the world. The concept of MNMS has been implemented as part of the data broadcasting in the project entitled News On Demand Kiosk Network (NODKN) [9]. This project is a collaborative between Multimedia University in Malaysia and Mitsushita in Japan.

The most essential activity involves in news content management process are authoring and publishing with additional of multimedia elements such as video, audio, image, text and animation. The MNMS tool provides an environment where the employees of a publication company can create, reviewing, publishing, versioning, archiving, changing and delete their news content or items. They can use this tool to support authoring and publishing operation of multimedia interactive news. The MNMS also provide a medium to communication with other user around the world to ease them shared the information. As a web application, the MNMS system allows end-users to search, view and read articles the interactive multimedia news based on their own personalization; for instance they can choose their own language.

Large volume patient related information like demographic data, historical data, data-rich multimedia diagnostic studies and longitudinal disease are one of the important parts in medicine information management. Some of general medicine clinics still used traditional record system such as handwritten notes to record information from patient examination and the physician just can review the record after accessing it in the office. To overcome this situation, Electronic Medical Record (EMR) application was suggested by [10] in order to manage patient data efficiently and accurately.

The EMR tool was develop in Spanish and English version language with the clinical practitioner at remote clinics in Ecuador. The tools support for patient information management in electronic form and electronic file sharing between the regional collaborating clinics in Ecuador. The implementation of EMR, migrates the patient information management from handwriting notes to distributed health care system. An authorize staff or clinicians easily recorded patient information and were able to achieve the information back for updating or for follow-up in patient care practices. Images or video data captured by radiographs, sonograms, microscopic and colposcopy examinations process were composed and stored within the workstation. The EMR user friendly interface allowed quick load and review of the digital images when the data is needed.

An unstructured data such as multimedia files are difficult to capture and store in the common database storage. In one of such study, [11] was interested to design a prototype tool to extract and classify unstructured data in any web pages. The prototype was focus on an efficient data classification and organization in providing effective multimedia data manipulation. Document Object Module (DOM) tree technique is applied to find the correct data in the HTML document in classification process. The prototype was implemented and tested on four main class data type in various formats to help end user get useful multimedia data stored for future retrieval usage. However, some of unnecessary information, such as script, style or other customized nodes in DOM tree cannot eliminate completely and may be minimized during extraction process.

Several weaknesses in current multimedia data management model have been found, and the weaknesses are as follow:

- Various models do not combine all type of multimedia data for classification but focusing on one type of data in each research
- Previous studies have emphasized that the issue of multimedia data storage and management, but little is so far known about the classification of multimedia data
- A lot of multimedia data classification model based on media format (e.g : .jpg, .txt, .mp3, .flv) but not by their attribute.

3 Rough Set Theory

Rough set theory, introduced by Pawlak [2][3] in the early 1980s. It is an extension of the set theory for the study of intelligent system characterized by inexact, uncertain or vague information and can serve as a new mathematical tool to soft computing [12]. An upper and a lower approximation of a set, the approximation space and models of sets are the fundamental concepts in rough set theory [13]. General elements engage in rough sets theory can be described as follows:

1) *Universe:* A non-empty finite of objects named training set, $U = \{x_1, x_2, \ldots, x_n\}$;
2) *Attributes:* A non-empty finite set of attributes, $A = \{a_1, a_2, \ldots, a_k\}$;
3) *Information system:* (also called decision table) is a pair of the universe and attributes, IS=<U,A> ;
4) *Indiscernibility relation:* (known as equivalence class) defines a partition in U. The indiscernibility relation is defined as, $R(B) = \{(x,y) \in U \times U : \text{for all } a \in B, a(x) = a(y)\}$ where, $a \in A$ and $B \subseteq A$;
5) *Approximation spaces:* define as a pair, AS = <U, R(C)> where, C be a set of condition attributes and R(C) be an indiscernibility relation on U.

- $[x]_B$ denotes the equivalence class of B containing x, for any element x of U;
- Based on singleton x, for a given $B \subseteq A$ and $X \subseteq U$, the *lower approximation* ($\underline{B}X$) of the set X in IS and the *upper approximation* of the set X in IS ($\overline{B}X$) are defined as follows:

$$\underline{B}X = \{x \in U : [x]_B \subseteq X\}. \quad (1)$$

$$\overline{B}X = \{x \in U : [x]_B \cap X \neq \varnothing\}. \quad (2)$$

- For a given $B \subseteq A$ and $X \subseteq U$, the boundary of X in IS can be defined as,

$$BND(X) = \overline{B}X - \underline{B}X. \quad (3)$$

BND(X) consists of objects that do not certainly belong to X on the basis of A.

Data used in rough set theory are often presented as a table which is initialized as decision table as illustrated in figure 1. In the table, columns correspond to attributes and rows of the decision table correspond to objects. Entries in the table are attribute values. The attributes of the decision table include condition attribute and decision attribute. The attributes in A can be further classified into two disjoint subsets, *condition attributes* (C) and *decision attributes* (D) such that A=C $\cup$ D and C $\cap$ D = $\varnothing$. The decision attributes can have some values though quite often it is binary [14]. Let B $\subseteq$ A, U/R (B) denotes a family of all equivalence classes of the relation R (B) called elementary sets. The elements of $\underline{B}$X are all and only those objects x $\in$ U which belong to the equivalence class generated by the indiscernibility relation contain in X. Meanwhile, the elements of $\overline{B}$X are all those objects x $\in$ U which belong to the equivalence classes generated by the indiscernibility relation containing at least on object x belong to X. The BND (X) indicates the objects in IS are inconsistent or vague. If upper and lower approximations of X are equal then X is an ordinary set. Clearly that, rough set theory mainly resolves to the problem how X $\subseteq$ U can be covered with a set of equivalence classes according to indiscernibility relation.

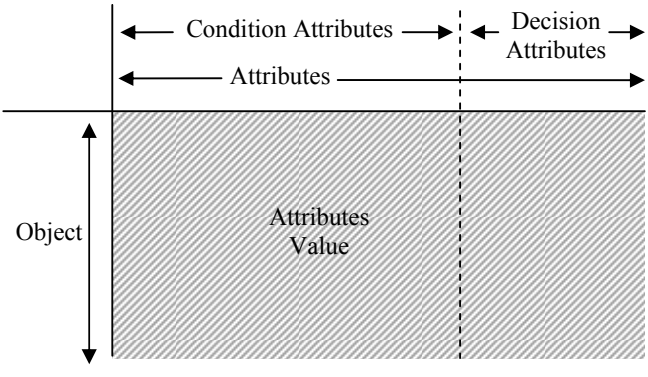

Fig. 1. Decision table

4 Classification Process for Multimedia Data Using Rough Set Theory

Classification of objects in the databases or information systems sources based on rough set theory has been done in many applications [5][15][16]. The goal of classification is to build a set of models that can accurately predict the class of different objects. Rough set mainly deals with data analysis in table format. The approach is generally to process the data in the table and then to analyze them. In this section, several multimedia data sets are shown with possible media type will be used as an example to illustrate the concept of rough set theory. Let, an information system is a set of objects represented in a data table, the rows are considered as objects for analysis and the columns represent a measureable attributes for each object, where

IS = <U,A>. Table 1 shows an information system which is a collection of multimedia data as object.

Given a set of universe, U= {O_1, O_2, O_3, O_4, O_5, O_6, O_7}, where U are all of the objects. The set condition attributes is represented by A= {*Illustration, Timeline, Movement*} and the set D represented the decision attribute, where D= {Media Types}. Table 1 can be shown in relation to the function of nominal values of the considered attributes, in the Table 2. Based on Table 1, classification of multimedia data produced based on condition attribute for each object. Theoretically, video and audio contain timing data [7] which can be used to track the movement of an object from frame to frame. Images are categorized as static media, while audio and video are categorized as dynamic media [6]. To build this information system, if C = {*Illustration*, yes} then, decision attribute can be defined as video or image media types. If C = {*Timeline*, yes} then decision attribute as video or audio types are admitted certainly. Justification based on attribute can be used to classify media types whether; it is a video, an audio or an image.

Table 1. Information system

	Condition Attributes			**Decision Attribute**
Object	*Illustration*	*Timeline*	*Movement*	Media Types
O_1	Yes	Yes	Dynamic	Video
O_2	Yes	No	Static	Image
O_3	Yes	No	Dynamic	Video
O_4	No	Yes	Static	Audio
O_5	Yes	Yes	Static	Video
O_6	No	Yes	Dynamic	Audio
O_7	No	Yes	Dynamic	Video

Table 2. Nominal Values of Attributes

	Attributes	**Nominal Values**
Condition Attributes	*Illustration*	Yes, No
	Timeline	Yes, No
	Movement	Static, Dynamic
Decision Attributes	Media Types	Video, Audio, Image

Indiscernibility relation is the relation between two objects or more, where all the values are identical in relation to a subset of considered attributes. In Table 1, it can be observed that the set is composed of attributes that are directly related to multimedia data, where A={*Illustration, Timeline, Movement*}, the indiscernibility relation is given to R(A). When Table 1 is broken down it can be seen that the set regarding {O_1, O_2, O_3, O_5} is indiscernible in terms of *Illustration* attribute. The set concerning

$\{O_1, O_4, O_5, O_6, O_7\}$ is indiscernible in term of *Timeline* attribute, and the *Movement* attribute generates two indiscernibility elementary sets are $\{O_1, O_3, O_6, O_7\}$ and $\{O_2, O_4, O_5\}$.

Approximations are fundamental concepts of rough set theory, it is can be defined as upper bounds and lower bounds. As define in section 3, based on objects in Table 1 the lower and upper approximation of X are categorize as follows,

$\underline{B}(X) = \{O_1,O_2,O_3,O_4,O_5\}$, $\overline{B}(X)=\{O_1,O_2,O_3,O_4,O_5O_6,O_7\}$. As a result, $BND(X) = \{O_6, O_7\}$. Figure 2 shows the relationship of the lower and upper approximation of an information system. The elements that enclosed by thick line belong to upper approximation. Meanwhile, the elements that enclosed by light line belong to the original set X. The elements that covered by grey color is belong to the lower approximation.

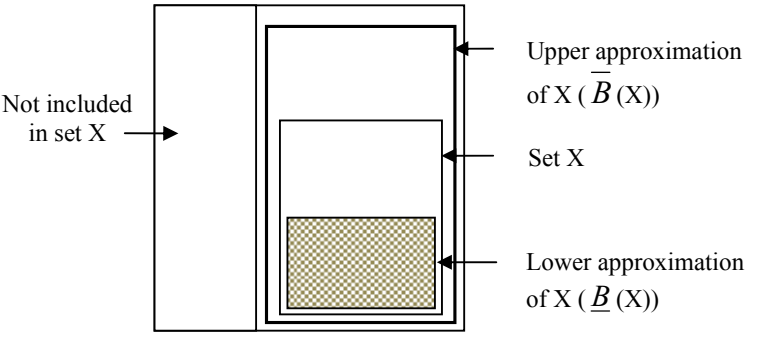

Fig. 2. Lower and upper approximation of set X

Table 3. Reducts table

	Condition Attributes		**Decision Attribute**
Object	*Illustration*	*Timeline*	Media Types
O_1, O_5	Yes	Yes	Video
O_2	Yes	No	Image
O_3	Yes	No	Video
O_4, O_6	No	Yes	Audio
O_7	No	Yes	Video

Based on lower and upper approximation, one way to facilitate data retrieving and manipulation is by reduction the set of data with reducing attributes. In indiscernibility relation, only attribute that do not contribute to the classification result can be omitted. Reduction means, the set of remaining attributes is the minimal set, and set which presents in all subsets call cores, in other words, removing repetitive or

overlapping data. The main purpose of reduction is to determine the attributes which can represent data in a database and dependencies between attributes. Table 3 shows the example of reduct which drops attribute *movement* and combined same objects in the same row. Decision rules 4) and 5) in table 3 have the same conditions but different decisions. Such rules are called inconsistent; otherwise the rules are referred to as consistent.

Decision rule created by combining rule reducts attributes. Each rows of reduct table verify a decision rule, which specifies the decision that must be taken when condition are indicated by condition attributes are fulfilled. Decision rules frequently presented as implication called "*if…then…*" rules. From the certainty factors of decision rules, the result as below:

- *if (illustration,yes) and (timeline,no) then (media types,image)*
- *if (illustration,no) and (timeline,yes) then (media types,audio)*
- *if (illustration,yes) and (timeline,yes) then (media types,video)*

Classifying data into several attributes is important because the attribute has to be matched with the corresponding data classes specified in the decision attribute. By applying rough set theory in classification of multimedia data, this model consists of six important elements; information system, indiscernibility relation, lower and upper approximation, reduction and decision rules. This model demonstrates that redefined indiscernibility can reduce the number of elementary sets. In addition sets of one object will enhance the approximation precision, but decrease the accuracy of decisions. Using the model provided, the theory of rough set proves to be an effective tool for multimedia data because:

- It reduces the data set without losing originality of characteristic set
- It is easy for clustering data
- It can manage multimedia data and effortless for application access.

5 Conclusion

The proposed model has been developed to help the end users to manage useful multimedia data (audio, video and image) in order to allow efficient process for storing, retrieving and updating data. The fundamental concept for classifying multimedia data in this research is based on attributes; *illustration, timeline* and *movement*. These attributes are mainly used to make a decision in classifying of media types. This research also has a new contribution in introducing rough set theory technique to organize and categorize multimedia data. The integration of classification data using rough set theory is believed to improve multimedia data management process. With more comprehensive study and investigation, we assume that some applications using classification of attribute in the rough set theory, using existing multimedia data in the real life multimedia organization through this view will be applicable. A future vision is to investigate the performance of this proposed model executed under the web services.

Acknowledgements

This research was supported by Fundamental Research Grants Scheme 2009 (FRGS-2009), Ministry of Higher Education, Malaysia and Universiti Sultan Zainal Abidin.

References

1. Nordin, M.A.R., Farham, M., Suhailan, S., Sufian, M.D., Kamir, M.Y.: Applying Time Granularity in Multimedia Data Management. In: Proc. International Conference on Computational Intelligence and Vehicular System (CIVS), pp. 60–64 (2010)
2. Pawlak, Z., Grzymala–Busse, J.W., Slowiriski, R., Ziarko, W.: Rough Sets. Comm. of the ACM. 38(11), 88–95 (1995)
3. Pawlak, Z.: Rough Set: Theoretical Aspects of Reasoning About Data. Kluwer Academic Publishers, Dordrent (1991)
4. Zhong, N., Dong, J.Z., Ohsuga, S.: Using Rough Sets with Heuristics for feature Selection. Journal of Intelligent Information Systems 16, 199–214 (2001)
5. Shyng, J.-Y., Wang, F.-K., Tzeng, G.-H., Wu, K.-S.: Rough Set Theory in Analyzing the Attributes of Combination Values for the Insurance Market. Journal of Expert Systems with Application 32, 56–64 (2007)
6. Jalal, S.K.: Mutimedia Database: Content and Structure. In: Workshop on Multimedia and Internet Technologies, Bangalore (2001)
7. Griffioen, J., Seales, B., Yavatkar, R., Kiernan, K.S.: Content Based Multimedia Data Management and Efficient Remote Access. Extrait de la Revue Informatique et Statistique dens les Sciences Humaines 1(4), 213–233 (1997)
8. Candan, K.S., Sapino, M.L.: Data Management for Multimedia Retrieval. Cambridge University Press, New York (2010)
9. Cheong, S.N., Azahar, K.M., Hanmandlu, M.: Development of Web-based Multimedia News Management System for News on Demand Kiosk Network. WSEAS Transaction on Computers 2(2), 360–365 (2003)
10. Azhar, R., Zhao, X., Cone, S., Merrell, R.: Electronic Multimedia Data Management for Remote Population in Ecuador. International Congress Series 1268, 301–306 (2004)
11. Abidin, S.Z.Z., Idris, N.M., Husain, A.H.: Extraction and Classification of Unstructured Data in WebPages for Structured Multimedia Database via XML. In: International Conference in Information Retrieval Knowledge Management (CAMP), pp. 44–49 (2010)
12. Xu, W.-H., Zhang, W.-X.: Knowledge Reduction in Consistent Information System Based on Dominance Relations. In: Liu, Y., Chen, G., Ying, M. (eds.) Optimization Techniques 1973. LNCS, vol. 3, pp. 1493–1496. Springer, Tsinghua University Press (2006)
13. Wang, Y., Ding, M., Zhou, C., Zhang, T.: A Hybrid Method for Relevance Feedback in Image Retrieval Using Rough Sets and Neural Networks. International Journal of Computational Cognition 3(1), 78–87 (2005)
14. Fomina, M., Kulikov, A., Vagin, V.: The Development of The Generalization Algorithm based on The Rough Set Theory. International Journal Information Theories & Application 13(13), 255–262 (2006)
15. Hu, X.: Using Rough Sets Theory and Database Operation to Construct a Good Ensemble of Classifiers for Data Mining Applications. In: Proc. of ICDM, pp. 233–240 (2001)
16. Nordin, M.A.R., Yazid, M.M.S., Aziz, A., Osman, A.M.T.: DNA Sequence Database Classification and Reduction: Rough Sets Theory Approach. In: Proc. of 2nd International Conference on Informatics, pp. 41–47 (2007)