International Journal of Hybrid Information Technology
Vol.6, No.5 (2013), pp.209-218
http://dx.doi.org/10.14257/ijhit.2013.6.5.19

Rules Generation for Multimedia Data Classifying using Rough Sets Theory

M. Nordin A. Rahman, Yuzarimi M. Lazim, Farham Mohamed,
Syarilla Iryani A. Saany and M. Kamir M. Yusof

Faculty of Informatics and Computing,
Universiti Sultan Zainal Abidin, Gong Badak Campus
21300 Kuala Terengganu, Malaysia

mohdnabd@unisza.edu.my, ayulazim@yahoo.com.my, mahraf0212@yahoo.co.uk,
syarilla@unisza.edu.my, mohdkamir@unisza.edu.my

Abstract

An efficient multimedia data management process is very important in multimedia system application. Huge size of multimedia data that distributed in multi locations makes multimedia data management more complicated. Rapid development of multimedia applications created a vast volume of multimedia data and it is exponentially incremented from time to time. This situation requires for efficient data classification and organization technique for providing effective multimedia data manipulation process. Using rough sets theory and web services technology, this paper proposed a new rules generation for multimedia data classifying in collaborative environment. ROSETTA tool is applied to verify the reliability of the generated results. The experiments show that the rough sets theory based for multimedia data classifying is suitable to be executed in web services environment.

Keywords: *multimedia data management, rough sets theory, web services, ROSETTA*

1. Introduction

Multimedia data consist of texts, graphics, animations, video, sounds, music, *etc.* Multimedia is defined as combination of more than one media; they may be two types, static and dynamic media [1]. [8] describes text, graphics and images are categorized as static media, while objects like animation, music, audio, speech, and video are categorized as dynamic media. 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. In research on content based multimedia data management conducted by [5] stated that, audio data contains certain identifiable features such as words, sound, pitches, and silent periods as well as timing information.

The major research domains 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 process and provides a suitable environment for using and

ISSN: 1738-9968 IJHIT

managing multimedia data [1]. Technique of indexing and classification in multimedia data is developed in order to ease the query processing.

There are many issues and challenges faced by multimedia data providers to fulfill 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. Many studies have been done in multimedia data management. For example, [3, 4] proposed temporal elements such as valid time and transaction time into multimedia data management transactions. The important issue in multimedia data management is data quality and organization, which involved data uncertainty, data redundancy, data incomplete and data inconsistency [10]. 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*. [19]. Several weaknesses in current multimedia data management model have been found, and can be categorized as follow:

- Various models do not combine all type of multimedia data for classification purposes but focusing on one type of data in each research
- No specific technique for classifying multimedia data into different clusters of data type
- The existing multimedia data classification model based on media format (*e.g.*, .jpg, .txt, .mp3, .flv) and not by attributes of the objects.

Rough set theory that proposed by [16] is a data mining approach that can be used to discover the dependency of data, reduction and rule induction from databases. It can be utilized for representing and reasoning imprecision and uncertain information in data, requires no external parameters, and used only the information presented in the given data [7, 14, 15, 16]. Rough set theory also can be used to develop classification model for the particular datasets, where this classification is used to group the data into predefined group. The theory is proven to be very useful in practice of many real life applications, for example in medical, pharmacology, engineering, human resources, banking, financial, market analysis, *etc*. [15].

This research discussed how rough set theory could be used to predict accuracy of multimedia data type whether the data is audio, image or video. Then, web services technology is applied to execute the proposed model under collaborative environment. The paper is organized as follows: Section 2 describes the overview of the related works on classifying data using rough sets theory; Section 3 presents the proposed model and experimental results; Section 4 describes the architecture of the proposed model implementation under web services environment and; last section concludes the paper and addresses the future work of the proposed model.

2. Related Works

The goal of classification is to build a set of data models that can accurately predict the class of different objects. Nowadays, numerous successful implementation of data classification in various applications using rough set theory are available. Table 1 summarizes several previous work of data classification based on rough set theory.

Table 1. Previous works in rough sets theory based data classification

Author /Year	Description
McKee and Lensberg [11]	Classification rules of breast cancer data
Hassanien and Ali [6]	Bankruptcy Classification
Midelfart *et al.* [12]	Proposed a general rough set approach for classification of tumor samples analyzed
Wang *et al.* [18]	Presented a new approach to classify the four types of scene image
Shen and Chen [17]	Classifications of customer loyalty based on the data gather from the customer feedback.

3. Applying Rough Sets Theory in Multimedia Data Classification

The process of dividing a universe of objects into different categories is called clustering. If we have large data sets, acquired from measurements or from human experts, these data sets may represent vague knowledge, uncertain or incomplete knowledge. Pertaining to this research, rough set theory is used to discern and classify the multimedia objects in the multimedia dataset.

3.1 The Model Process Flow

There are five (5) steps are considered in the proposed rough sets theory based multimedia data classification. The steps can be denoted as:

> {*establishment of information system, development of indiscernibility relation, construction of set approximation, determination of reduct set and construction for decision rules*}.

The steps process flow is depicted in Figure 1 as a general algorithm. The detail descriptions of each step are discussed in the next paragraphs.

Input: A universe, *multimedia dataset*
Output: A set of *multimedia data clusters*

1. Construct information systems, IS = (U, A), where U is multimedia data set and A is a set of attributes of each U members.
2. Generate indiscernibility relation on U
3. Get the lower and approximations from U
4. Reduct ← {} // Set of reduct objects
5. ∀x ∈ U
   a. Eliminate redundancy of attribute
   b. T ← Reduct ∪ {x}
6. Reduct ← T
7. C_{image}← {}, C_{video}← {}, C_{audio}← {}
8. ∀x ∈ U // apply classification rules
   a. If (x is in image class)
      - C_{image}← C_{image}∪ {x}
   b. Else if (x is video clas)
      - C_{video}← C_{video}∪ {x}
   c. Else if (x is audio class)
      - C_{audio}← C_{audio}∪ {x}
9. Insert C_{image}, C_{video}, C_{audio} in <multimedia database>

Figure 1. The flow process of the proposed model

Establishment of information system. An information system is a data set that represented in a table as illustrated in Figure 2. Each row in the table represents an object, such as a case or an event. Each column in the table represents an attribute, *e.g.*, a variable, an observation or a property. For each object (row) is assigned with some attribute values. Formally, an information system can be signed as a system, $IS = (U, A)$ where U is non-empty finite set of objects called the universe, $U = \{O_1, O_2, O_3, \ldots, O_n\}$; and A is a non-empty finite set of attributes (features, variables), the attributes in A are further classified into disjoint condition attributes C and decision attributes D, such that $A=C \cup D$ and $C \cap D = \varnothing$.

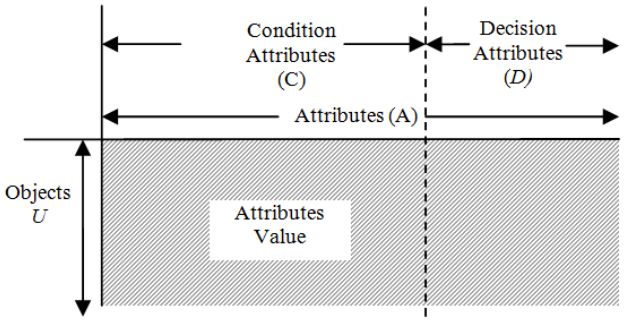

Figure 2. Information system

For the proposed model, the example of information system can be shown as in Table 2. The objects in this example represent the multimedia data. Different attributes (columns); *illustration*, *timeline* and *movement* are used to determine the class type of each object.

Table 2. Example of information systems

Object	Illustration	Timeline	Movement
O_1	Yes	Yes	Dynamic
O_2	Yes	No	Static
O_3	Yes	No	Dynamic
O_4	No	Yes	Static
O_5	Yes	Yes	Static
O_6	No	Yes	Dynamic
O_7	No	Yes	Dynamic

Development of indiscernibility relation. The relation between two objects or more is called indiscernibility, where if all the values of attributes are identical. Given a subset of attributes, $\boldsymbol{a} \in \boldsymbol{A}$ and $\boldsymbol{B} \subseteq \boldsymbol{A}$, each such subset defines an equivalence relation $IND_A(B)$ called an indiscernibility relation. This indiscernibility relation can be defined as:

$$IND_A(B) = \{(x, x') \in U^2 | \forall a \in B, a(x) = a(x')\} \quad (1)$$

Equation (1) states that the subset of attributes, B, will define a clustering process of the universe into sets such that each object in a set cannot be distinguished from other objects in the set using only the attributes in B. The sets which the object are divided into are called *equivalence classes*. From the information system shown in Table 2, the objects O_6 and O_7

can be classify under indiscernible relationship. The example of indiscernibility relation from the Table 2 can be signed as:

IND (Illustration) = {{O_1, O_2, O_3, O_5}{O_4, O_6, O_7}}

IND (Illustration, Movement, Timeline) = {{O_1}, {O_2}, {O_3}, {O_4}, {O_5}, {O_6, O7}}

If a new attribute is added to the information system, and this attribute represents some classification of the objects, then the system is called a decision system. It can be represented as:

$$IS = (U, A \cup \{d\}) \quad (2)$$

where, d is decision attribute.

The elements of A are known as condition attribute. The decision is not necessarily constants on the equivalence classes. Therefore, two objects are belonging to the same equivalence class, but the values of the decision attribute may be different. As an example, if a new data is added into information system in Table 2, then a decision system can be show as in Table 3. Those attributes could determine the objects classification whether an audio, an image or a video. From Table 3, object O_6 and O_7 are belonging to the same equivalence class, but they are classified differently. Meaning that, they have the same values for the conditional attributes, but different values for the decision attribute. Therefore, the information system can be categorized inconsistent. To remedy this, condition application of object is required.

Table 3. Example of decision system

Object	Illustration	Timeline	Movement	Decision
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

Construction of set approximation. In order to cluster an object based on the equivalence class in which it belongs, the concept of set approximation is needed. The approximation is based on the information contained in B. Therefore, given an information system, $\boldsymbol{IS} = (\boldsymbol{U}, \boldsymbol{A} \cup \{\boldsymbol{d}\})$, and a subset of attribute, $\boldsymbol{B} \subseteq \boldsymbol{A}$, to approximate a set of objects $\boldsymbol{X}$, the Equation (3) and (4) can be used to define the lower approximation and upper approximation of set *X*.

$$\underline{B}X = \{x|[x]_B \subseteq X\} \quad (3)$$

$$\overline{B}X = \{x|[x]_B \cap X \neq \emptyset\} \quad (4)$$

The lower approximation is the set containing all objects for which the equivalence class corresponding to the object which is a subset of the set. This set contains all objects which

with certainty belong to the set X. The upper approximation is the set containing the objects for which the intersection of the object equivalence class and the set we would like to approximate is not the empty set. This set contains all objects which possibly belong to the set X. For a given $B \subseteq A$ and $X \subseteq U$, the boundary of X in information system can be defined as:

$$BN_B = \overline{B}X - \underline{B}X \quad (5)$$

In this case, BN_Bconsists of objects that do not certainly belong to X on the basis of A. The concepts of lower approximation, upper approximation and B-boundary region are illustrated in Figure 3. Based on objects in Table 2 the lower and upper approximation of X are resulted as follows: $\underline{B}X = \{O_1, O_2, O_3, O_4, O_5\}$, $\overline{B}X = \{O_1, O_2, O_3, O_4, O_5, O_6, O_7\}$ and $BN_B = \{O_6, O_7\}$.

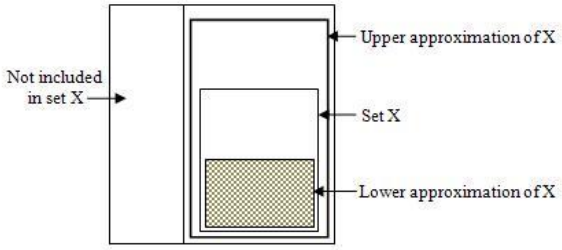

Figure 3. Set approximation

Determination of reduct set. A minimal set of attributes after the redundancy removal called reducts. Not all attributes in an information system is needed to classify the objects into particular classes. In these case, we can reduce the attributes by reducing the attribute which is it is not influenced the classification process. A reduct of A is defined as minimal set of attributes $\boldsymbol{B \subseteq A}$ such that $\boldsymbol{IND_A(B) = IND_A(A)}$. In the designed model, to discern between the different equivalence classes, only the attributes *illustration* and *timeline* are necessary. Thus, {*illustration, timeline*} is an example of reduct:

$$IND_A(\{Illustration, timeline\}) = IND_A(A)$$

Table 4 shows the example of reduct which attribute *movement* is dropped and indiscerbility objects are inserted in the same row. Decision rules in row 4th and 5th in Table 4 have the same conditions but different decisions. Such rules are called inconsistent; otherwise the rules are referred to as consistent.

Table 4. Reduct table

Object	Illustration	Timeline	Decision
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

Construction of decision rules. The approximation of the decision, d can be defined by constructing the decision rules. From the reduct computation, decision rules can be generated and objects clustering can be made. The decisions rules are created by combining rule reduct attributes. Each row of reduct table verifies a decision rule, which specify the decision that must be considered whenever conditions are indicated and the condition attributes are fulfilled. Table 4 shows the example of reduct set which attribute movement is dropped and the same object will be placed in the same row. The reduct set produced the following decision rule:

$$Illustration = yes \wedge timeline = no \Longrightarrow Decision = image$$

3.2 The Experimental Results

In the experiment, we have evaluated the dataset and accuracy of classification results using Rough Set Technical Analysis software (ROSETTA). ROSETTA is a toolkit application for knowledge discovery, data mining and analysis of tabular data that using the rough set methodology [20]. It has a friendly graphical user interface and easy to use. ROSETTA is applied for the entire rough set experimental processes are carried out, starting from data preprocessing stage to data classification stage.

The institution of higher education is an active agency that produced a vast volume of multimedia data and innovations [13]. Therefore, the dataset used in our experiment are collected from Multimedia Department, Information Technology Centre of Universiti Sultan Zainal Abidin. There are 350 objects are selected. A set of 175 objects are used for training, and a set of 175 objects are used for testing. The training data set includes 35 objects of audio, 115 objects of images and 25 objects of video. The value and result/meaning of condition and decision attributes is shown in Table 5. As an example of the classification result, from the first rule in Table 6, the condition shows that the object has a timeline but no illustration and movement, and the file type is Mp3. Therefore, the system will classify the object as an audio.

Table 5. Value and meaning of condition and decision attributes

Attribute	Name	Value	Result/meaning
C_1	Illustration	0,1	No, Yes
C_2	Timeline	0,1	No, Yes
C_3	Movement	0,1	Static, Dynamic
C_4	File Type	1~14	MP3, MP4, JPEG, FLV, wave, wma, gif, Movie clip, WMP, 3GP, video clip, bitmap, .RM, PNG
C_5	Category	1~21	Islamic, patriotic, people, technology, love, object, cartoon, traditional, abstract, sport, promo, funny, food, animal, celebration, nature, architecture, quote, entertainment, education, medical
D	Media Type	$d_{1\sim3}$	Video, image, audio

Table 6. Decision table for the object classification

Object	Condition Attribute, C_i (i=1~5)					Decision Attribute, d_j (j=1~3)
	C_1	C_2	C_3	C_4	C_5	
1	0	1	0	1	1	d_3
2	1	1	1	2	2	d_1
3	1	0	0	3	2	d_2
4	1	0	0	3	3	d_2
:	…	…	…	…	…	…
:	…	…	…	…	…	…
:	…	…	…	…	…	…
350	0	1	0	1	1	d_3

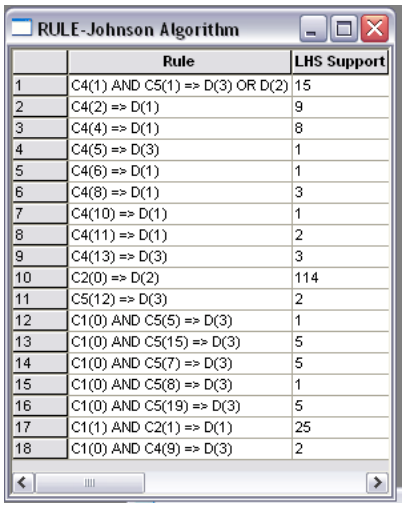

Figure 4. The generated rules

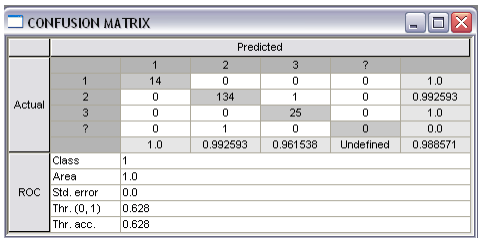

Figure 5. Accuracy of the classification

From screenshot of ROSETTA software as in Figure 4, there are 18 rules are generated. As an example, if the last rule in Figure 4 is considered: C1(No) $\wedge$ C4 (WMP)$\rightarrow$D3(Audio). Meaning that if the object does not have an illustration and file type of object is WMP, then we can classify that the media type is an audio. In Figure 5, value 1 denotes for video object, value 2 is image object and value 3 is an audio object. As shown in Figure 3, the average accuracy of the video object is the highest followed by image and audio and its recall rate is the lowest. Although the database images have a large number from the different object types, the overall accuracy of the system reached 98%, which verify the ability of rough set theory technique could precisely classify the uncertain knowledge of the multimedia data type.

4. Implementation the Model under Web Services Platform

The web services technology allows for interfacing, publishing and binding loosely coupled services through the web [2, 3, 9]. Nowadays, many organizations are sharing their multimedia data among other parties around the world. To manage those sharing application, web services facilities can be the best solution. The main function of web services platform is to allow a user from different platform to access multimedia data from different database and

multi-environment of operating systems. The development of the proposed application used XML-based web services with Java 2 Platform, Enterprise Edition (J2EE). This model supports three (3) main modules: *insertion*, *update* and *searching-clustering*. In order to evaluate the model, three (3) different database applications have been employed; which are MySQL, Oracle and Informix. The multimedia dataset used is replicated into these three (3) types of database systems. Furthermore, each database is installed in two different operating systems platforms; Microsoft Windows and Linux. The architecture of the proposed model is depicted in Figure 6. Figure 7 and Figure 8 shows the interface for *searching-clustering* module and its result respectively.

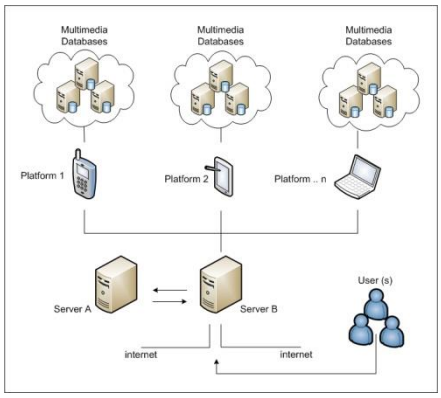

Figure 6. The proposed architecture model under web services based environment

Figure 7. Searching-clustering module

Figure 8. Searching-clustering result

5. Conclusions

The dramatic growth of multimedia information in diverse kind of data needs for efficient data management process. In this paper, we have proposed a new model for clustering multimedia data based on rough sets theory. The model attempts to classify the multimedia object into three standard multimedia data types. The experimental results have demonstrated that the rough set theory is effective to classify the multimedia data into their respective clusters. The accuracy level of classification result reached 98%. The classification process and results are validated with ROSETTA software. The proposed model has been implemented

under web services platform, where J2EE is employed to develop the application. The developed application promises independencies of communication between multi-platform environments. Currently, a comprehensive experiment is under way to combine the concept of temporal based data management to the proposed model in order to improve the classification efficiency.

References

[1] K. S. Candan and M. L. Sapino, "Data Management for Multimedia Retrieval", Cambridge University Press **(2010)**.
[2] S. Dai, H. Xu and H. Zhou, "Research on Multimedia Resource Management System Based on Web Services", Proc. of International Conference on Industrial and Information Systems, **(2009)**, pp. 259-262.
[3] M. Farham, M. A. R. Nordin and M. A. Lazim, "The Development of Temporal-Based Multimedia Data Management Application Using Web Services", Proc. Of International Conference on Intelligent Systems Design and Application, **(2011)**, pp. 487-492.
[4] m. Farham, M. A. R. Nordin, M. L. Yuzarimi and B. M. Saiful, "Managing Multimedia Data: A Temporal-Based Approach", International Journal of Multimedia and Ubiquitous Engineering, vol. 7, no. 4, **(2012)**, pp. 73-85.
[5] J. Griffioen, B. Seales, R. Yavatkar and K. S. Kiernan, "Content based multimedia data management and efficient remote access", Extrait de la Revue Informatique et Statisque dens les Sciences Humaines, vol. 1, no. 4, **(1997)**, pp. 213-233.
[6] A. E. Hassanien and J. M. H. Ali, "Rough Set Approach for Generation of Classification Rules of Breast Cancer Data", INFORMATICA, vol. 15, no. 1, **(2008)**, pp. 22-38.
[7] X. Hu, N. Shan, N. Cercone and W. Ziarko, "DBROUGH: A rough set based knowledge discovery system, methodologies for intelligent systems", Z. Ras and M. Zemankova, Eds. Springer-Verlag Press, **(1994)**.
[8] S. K. Jalal, "Multimedia Database: Content and Structure", Workshop on Multimedia and Internet Technologies, **(2001)**.
[9] M. Y. Kamir, M. A. R. Nordin and M. M. A. Atar, "Ontology and Semantic Web Approaches for Heterogeneous Database Access", International Journal of Database Theory and Application, vol. 4, no. 4, **(2011)**, pp. 13-23.
[10] M. Y. Kamir, M. A. R. Nordin and A. Atiqah, "Reducing of Inconsistent Data Using Fuzzy Multi Attribute Decision Making for Accessing Data from Database", International Journal of Database Theory and Application, vol. 6, no. 1, (2013), pp. 1-11.
[11] T. E. McKee and T. Lensberg, "Genetic Programming and Rough Sets: A hybrid approach to bankruptcy classification", European Journal of Operational Research, vol. 138, **(2002)**, pp. 436-451.
[12] H. Midelfart, J. Komorowski, K. Norsett, F. Yedetie, A. K. Sandwick and A. Largreid, "Learning Rough Set Classifier from Gene Expression and Clinical Data", Fundamental Informaticae, vol. 53, **(2002)**, pp. 155-183.
[13] M. A. R. Nordin, A. W. Fauziah, I. Rohana and U. Norlina, "A Comprehensive Innovation Management Model for Malaysians Public Higher Learning Educations", International Journal of Software Engineering and Its Applications, vol. 7, no. 1, **(2013)**, pp. 45-55.
[14] M. A. R. Nordin, M. L. Yuzarimi and M. Farham, "Applying Rough Set Theory in Multimedia Data Classification", International Journal on New Computer Architecture and Their Applications, vol. 1, no. 3, **(2011)**, pp. 706-716.
[15] M. A. R. Nordin, M. L. Yuzarimi, M. Farham, S. Suhailan, M. D. Sufian and M. Y. Kamir, "Rough Set Theory Approach for Classifying Multimedia Data", (Ed. Jasni, M.Z.) in Software Engineering and Computer Systems, Springer Verlag, **(2011)**, pp. 116-124.
[16] Z. Pawlak, "Rough set: Theoretical aspect of reasoning about data", Kluwer Academic Publisher, Dordrent, **(1991)**.
[17] L. Shen and S. Chen, "Research of Customer Classification Based on Rough Set using Rosetta Software", American Journal of Engineering and Technology Research, vol. 11, no. 9, **(2011)**, pp. 1279-1285.
[18] X. Wang, N. Liu and K. Xie, "Application of Rough Set Theory on Scene Image Classification", Chinese Control and Decision Conference, **(2008)**, pp. 2338-2342.
[19] M. L. Yuzarimi, M. A. R. Nordin and M. Farham, "Clustering Model of Multimedia Data By Using Rough Sets Theory", Proceedings of International Conference on Computers and Information Sciences, **(2012)**, pp. 336-340.
[20] The ROSETTA homepage http://www.idi.ntnu.no/~aleks/rosetta/. Norwegian University of Science and Technology, Department of Computer and Information Science. (Downloaded on January **2012**).

