ISSN: 2277-3754
ISO 9001:2008 Certified
International Journal of Engineering and Innovative Technology (IJEIT)
Volume 2, Issue 1, July 2012

A Movie Rating Approach and Application Based on Data Mining

S. Kabinsingha, S. Chindasorn, C. Chantrapornchai

Abstract— ***in this work, we are interested in the rating of movies. We apply the data mining to the movie classification. In the prototype model, the movies are rated into PG, PG-13 and R. The 240 prototype movies from IMdb (http://imdb.com) [1] are used. The data are divided into training and testing set with 4 fold cross validation. Among all various attributes of movies like actors, actress, directors, budget, genre, producers, etc., the total number of selected attributes is 8 which depends mainly on the genres of the movies and the words used in the movies. This corresponds to the decision used by most of the film rating organization. The prototype model is created based on the decision tree (J48) using Weka (http://www.cs.waikato.ac.nz/ml/weka/) [2]. The experiments have about 80%-88% precision for all the tested rating.***

Index Terms—**Data Mining; Movie Classification; Film Rating; Movie Rating; Decision Tree, IMDb.**

I. INTRODUCTION

Data mining has been popularly used in many areas such as marketing, customer relationships, banking, finance, statistics, inventory forecasting, bioinformatics, healthcare etc. [3], [4], [5], [6] the techniques have also been used for many recommender systems in many aspects like classification, clustering, predictions. It can be combined with many existing areas in the computer and information technology fields such as embedded system design, HCI, image/video processing. Currently, there are many interests in using data mining for video and image classifications, pattern mining, event extractions. The technique of data mining requires lots of data to be investigated. The data attributes must be studied in details to create an accurate model. The measurement of the accuracy of models can be F-measure, Precision, Recall, True positive/True negative/False positive/False negative etc. In this work, we are interested in the movie rating classification approaches. We attempt to apply data mining to create a model for a movie rating classification. A model prototype is created using the sample data from IMdb [1], which provides rich information about movies and the features selected are genres and words among many other attributes studied. From the experiments, about 80% accuracy is obtained. Several work has been done in movie classification. Some of them used image processing and some of use text mining to justify the rating. Some work used data mining to rate the movies. Amatriain, Jaimes, Oliver, and Pujol presented data mining methods for recommender systems [4]. The paper gives an overview of data mining processing, the measurement for feature selection, various classifiers. The work by Oh, Lee, Kote, and Bandi used data mining approach to extract information from the incoming video sequences [7]. They proposed an approach to detect interesting patterns from the video. The videos are segmented into pieces and the pieces are clustered to find abnormal events. Shijia, Liuzhang and Ming used data mining to predict popularity of the movies on the CDN system [8]. The predictions are used to recommends the number of copies kept in the CDN to aid accessing while the number of replications can be minimized for nonpopular movies. Bayesian Network and decision tree are compared for the prediction model. It is found that Bayesian Network is more appropriate. The correlation between popularity features are investigated such as actor, director, local, type, Showtime etc. Han presented a survey of how data mining can help extract from videos and images [9]. The paper particularly shows how pattern mining for images and videos, data mining to help create indexing useful for image and video retrieval, pattern-based classification, pattern-based clustering for images and videos. Saraee, White and Eccleston proposed the data mining method for the analysis of the movie rating [10]. Many features are interesting including keywords, names, titles, producers, companies, actors, composers, directors, genre, etc. The ratings are divided into levels as excellent, average, poor, and terrible. The features that affect the rating are actors, actress, directors, budgets and genre in the descending order. Fleischman, Decamp and Roy presented a method to extract temporal information from videos [11]. It is useful to extract events from videos. The tree kernel based on SVM is trained by a set of sample events. The results are compared with the Hidden Markov Model showing a high accuracy. The work by Changkaew and Kongkachandra also considered image properties to extract movie information [12]. They used colors of the images and the words. Image processing is used to extract the colors in the scenes to estimate the visual effect and the bad words are also used to help on the language issues. Also, the work by Chaovalit and Zhou presented the moving rating schemes using both supervised and unsupervised learning [13]. They are based on the movie reviews and used the opinion mining to help rate the movies. The paper is organized into the following sections. Section 2 presents some backgrounds on data mining. Section 3 discusses the methodology of the approach while Section 4 presents the experiment setup as well as the results. Section 5 concludes the work.

II. BACKGROUNDS

In this section, we present the backgrounds of the work as following.

A. Movie Rating

The purposed of movie rating is the guidance for parents about the films. They can suggest their children what they should see and what they should not see. However, it does not justify that the movie is good or bad. It provides the basic guidance about languages, sex, drug, and violence etc. (http://www.mpaa.org/ratings) [14].

There are several countries that have different ways to rate movies. For example, in Australia by the government, the movies can be rated as following (http://www.classification.gov.au/Pages/default.aspx) [15].

-E (Exempt from classification): This is for specific types of work for examples, educational purposes.

-G (General): The content is very mild.

-PG (Parental guidance recommended): The content is also very mild.

-M (Mature audiences): The content has a moderate impact.

-MA 15+ (Mature adult not suitable for people under 15): If under 15, the people must be accompanied with a parent or guardian. The content is strong.

-R18+ (Restricted to 18 and over): It is the high level content.

-Restricted to 18 and over: The classification contains the only sexually explicit content.

-RC (Refused classification): Banned from sales or hired in Australia.

Besides, in America, the Motion Picture Association of America (MPAA)'s film rating system is used since 1990 [14]. The ratings can be G (General audience), PG (Parental Guidance Suggested), PG-13 (Parents Strongly Cautioned), R (Restricted), NC-17 (No One 17 and under Admitted). Rating process is based on languages, drug content, sexual content, etc.

In British, there is a non-government organization called British Board of Films Classification (BBFC) funded by film industries in UK to classify the films (http://www.bbfc.co.uk/) [16] . The ratings are U (Universal), PG (Parental Guidance), 12A (Children under 12 must be accompanied with adults), 12 (Home media only, nobody younger than 12 can rent or buy), 15 (for those who is over 15 years only), 18 (only adults over 18 are admitted), R18 (can only be shown in license cinema), etc.

We can observe that the rating system is quite similar. So, in this work, we focus on the classification of G, PG, PG-13,R respectively on the MPAA system. It is quite interesting that rating of G, PG are very similar. As it is the prototype, we choose to classify into PG, PG-13, and R where the distinctions are clearer.

B. Data Mining

Data mining usually consists of three steps [4]. These are data preprocessing, data analysis, and result interpretation. The first phase, data preprocessing is the most important. The problem's data domain needs to be clearly studied. The attributes that should give affects to the results are investigated. Some statistical analysis such as correlation may be needed to analyze the relationships between them. This leads to the attribute selections. Also, the data need to be cleaned, filtered, or changed in some manner. For example, possible values of attributes are studied and may be grouped. Data records with improper attribute values are deleted and so on. Large data needs to be sampling to obtain the representative subset for the model.

After that, the data are analyzed. The techniques depend on the goal such as classification, clustering, association rule mining etc. In this paper, we focus on the clustering. The known clustering approaches decision tree, artificial neural net (ANN), supported vector machine (SVM), rule-based classifier, Baysian. For clustering, the popular one is k-nearest neighbor (k-NN). To measure the quality of the classifier, True Positive (TP), True Negative (TN), False Positive (FP), False Negative (FN), Accuracy, Precision, Recall and F1-measure are used.

Accuracy is defined as:

$$\frac{(TP+TN)}{(TP+TN+FP+FN)}$$

Precision is defined as

$$\frac{TP}{(TP+FP)}$$

Recall is defined as.

$$\frac{TP}{(TP+FN)}$$

F1-measure is defined as.

$$\frac{2TP}{2TP+FN+FP}$$

Data mining is a powerful for modeling for many applications areas. There are some limitations on it. Although, the created model confirms with the data profiles used, how in the real world behavior deviates from the model cannot be identified. Also, the model only shows the relationships between attributes but it does not imply the casual relationship [3] Thus, some more analysis is needed after the results are obtained.

III. METHODOLOGY

We perform according to the typical data mining process. The key of the presentation is the data preprocessing. In our work, the movie keywords are quite important since we find that the words that describe the movie tell about the context of the movies which imply the languages used. This has a major effect on the movie rating. In contrast with the work in [12] , we focus on the text processing while the work requires image processing to check the rating from the videos/images.

We also consider the genres and many groups of words to help classification. First, we collect the possible movie data from IMDb and other places. There are many records in the IMDb. We need to sampling the data related to our classifier. Since we focus on PG, PG-13, and R, we manually extract the records of movies for these rates and inspect them. Many attributes are created from the records such as actors, actresses, directors, producers, writers, budgets, genre, date, short story etc. Most of them are multivalue attributes. The encoding schemes are needed. For short stories the process of keywords extraction is required. Complexity occurs on these multivalue fields such as there are many major actors and actresses, producers, in a movie. A movie can be in more than one genres etc. The keyword extracting is also a little complex. A short description is first read as a text. The stop words or the words that are insignificant should be eliminated. This is such as how, to, and, etc. Next, we are remained with the nouns and verbs in many forms. This is also a difficult part. The verbs need to be transformed back to the original forms by eliminating –s, -es,- ing,-ed,-er,-ied, and so on and in some case, change the forms back. This part would be easy if we have a complete dictionary with the API that contains the original verbs in many forms. However, for us, we need to write a script to transform them. After transforming them back, we need to make sure that it is right by checking against the Lexitron. Next, we count the words' frequency and classify the group words. From observation, we classify the words in to the categories; that is; bad words, sexual, terror, religion, imprison, violence and drug. We need to levelize the number of frequencies in each group. Two approaches of the levelization are attempted: T-score and mid-point. For each, training data for each rating, we

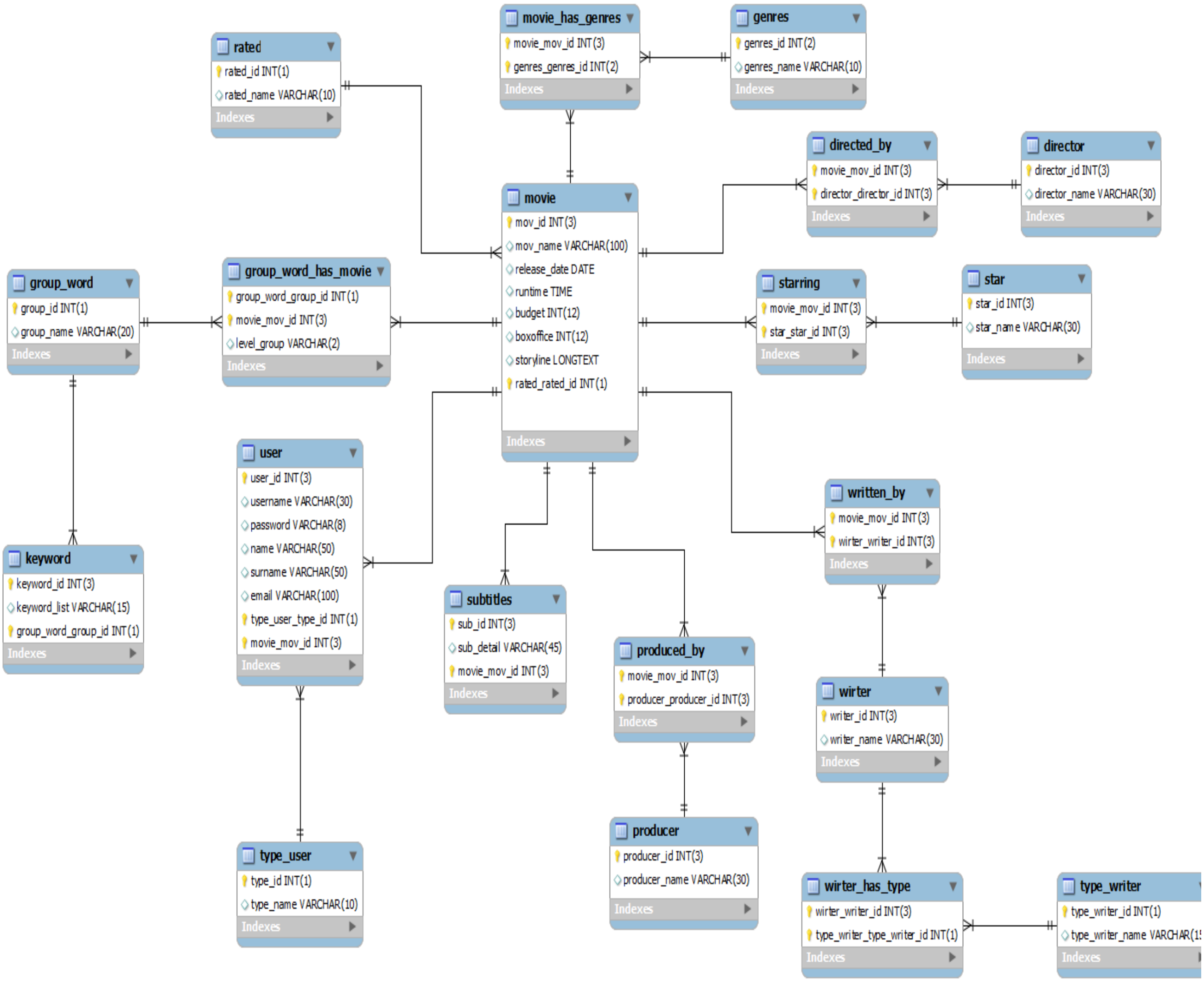


Fig 1. Database Contexts.

count the frequency of keywords found in each group. For the T-score approach, the frequency is the score itself. For all training movies, for each word group, we gather the score for each keyword, calculate the T-score and try to level them. For the mid-point approach, we calculate the midpoint for each word group for each rating. Thus, the levels obtained from both may be different. We have test the usability of each level style in the model creation as well.

After extracting records of the movies, we create a database to store these records according to the fields and the encoding of the field values. The database is shown in Figure 1. The core table is "movie". This stores the movie information like budget and it is linked to other tables which stores information about directors, producers, actors, actress, subtitles (in Tables "star", "producer", "writer", "subtitle") etc. Also, a movie can be of many genres (Table "genre") and related to many word groups (Table "group word").There are many tables in the database to completely store information about movies. We have the table of genres, directors, producers, writers, actors, actresses, etc. These are one to many relationships to each movie. The group word and keyword tables are used for keyword extraction from subtitles of the movies. The final goal here is to create the prototype application of moving rating system. The database is used to store the movies information as well as the new movie data to be classified when inserted in the web application. Next, the data from the database are exported to input to the data mining process. Note that the future work can integrate the features from the images/video processing to it.

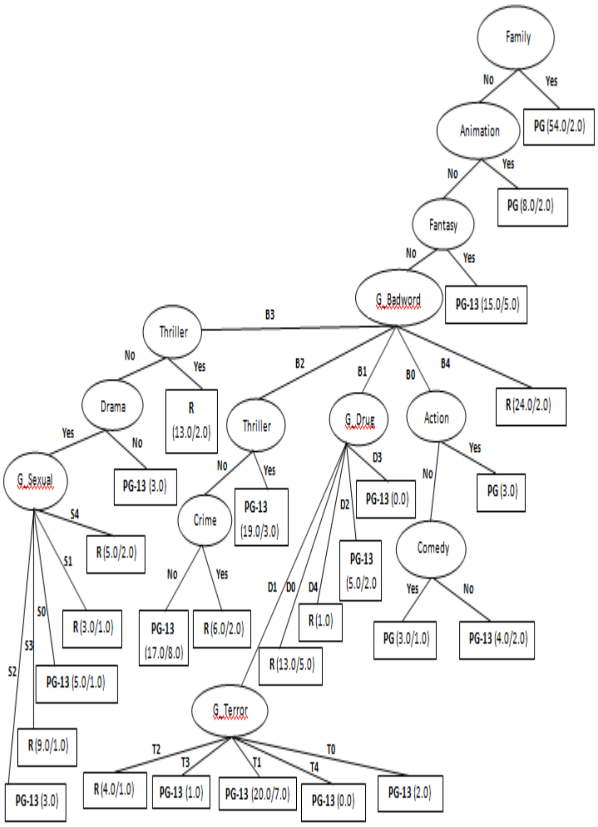
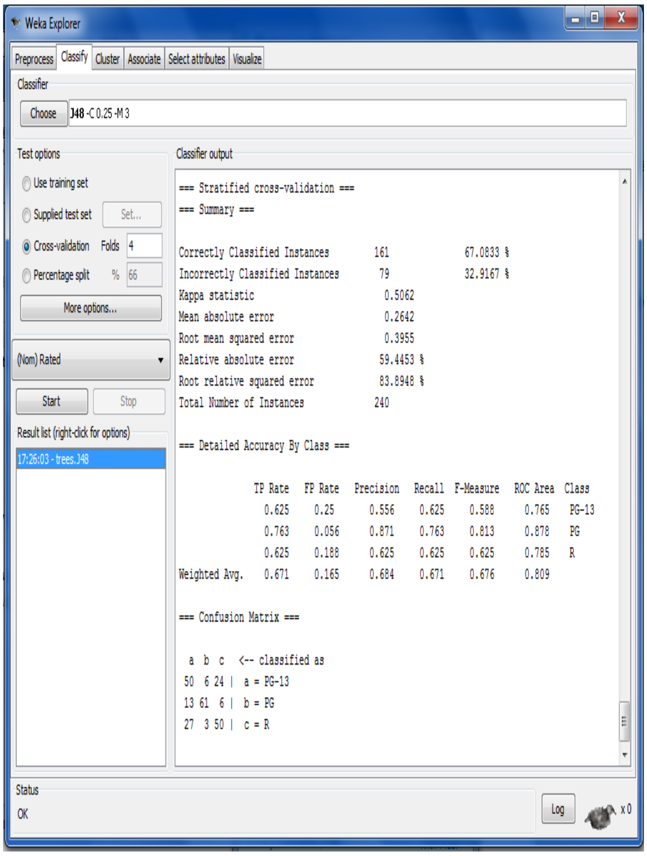


Fig 2. Decision Tree When Using T-Score to Levelize Word Groups.

IV. EXPERIMENTS

We create the database as well as keyword extraction. In the data mining process, we create a model using Weka [2]. There are many genres that we consider in the database: Action, Adventure, Animation, Biography, Comedy, Crime, Drama, Family, Fantasy, History, Horror, Music, Musical, Mystery, Romance, Sci-fi, Sport, Thriller, War and Western. Of all the attributes inserted, Weka selects only genre and group words as features in the decision tree (J48).Figure 2 shows the tree obtained when using T-score as a level method for word groups. In the tree, Weka uses the attribute "Genre" mainly to divide the rating. First, the genre "Family" is checked. Then, the genre "Fantasy" is checked. After that a group of "bad words" is used together with other genres. The difficult part is along the tree where genres and word groups are used alternatively. Totally, there are 9 genre nodes. Some of them are checked again and again while there are 4 word groups used in the tree uniquely.



J48 -C 0.25 -M 3

Folds 4

% 66

17:26:03 - trees.J48

```
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances         161               67.0833 %
Incorrectly Classified Instances        79               32.9167 %
Kappa statistic                          0.5062
Mean absolute error                      0.2642
Root mean squared error                  0.3955
Relative absolute error                 59.4453 %
Root relative squared error             83.8948 %
Total Number of Instances              240

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.625     0.25       0.556     0.625     0.588      0.765    PG-13
                 0.763     0.056      0.871     0.763     0.813      0.878    PG
                 0.625     0.188      0.625     0.625     0.625      0.785    R
Weighted Avg.    0.671     0.165      0.684     0.671     0.676      0.809

=== Confusion Matrix ===

  a  b  c   <-- classified as
 50  6 24 |  a = PG-13
 13 61  6 |  b = PG
 27  3 50 |  c = R
```

Status
OK

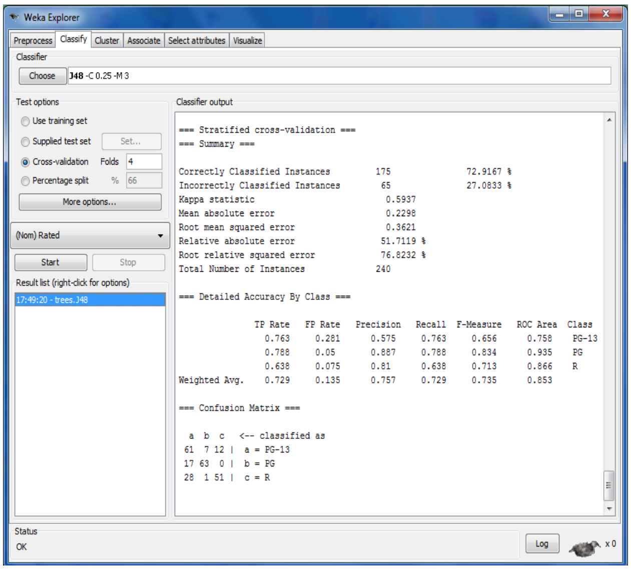
Fig 3. Weka Results When Using T-Score to Levelize Word Groups.

ISSN: 2277-3754
ISO 9001:2008 Certified
International Journal of Engineering and Innovative Technology (IJEIT)
Volume 2, Issue 1, July 2012

With the model in Figure 2, with the data set, Weka picks the attributes: Family, Animation, Fantasy, Thriller, Drama, Crime, Action, Comedy and the groups of words selected Bad word, Sexual, Terror, Drug with the J48 algorithm and cross validation for 240 sample movies. The prediction accuracy is from 161 movies which is 67.08%. This is shown in the results in Figure 3.

Figure 4 shows the tree obtained when using mid-point as a level method for word groups. When using mid-point, the tree is less complicated. Again, first the genre "Family" is checked. After that, the group of "bad words" is checked in the next level. Totally, there are 7 genres that are used to identify. The "comedy" node is the only one that is redundant. The tree uses only 2 word groups which is "bad words" and "Terror". With the model in Figure 4, with the data set using the mid-point level the prediction accuracy is from 175 movies which is 72.91%. This is shown in Figure 5.Table 1 compares both results. It compares the attributes used for each tree. For the training data, with the mid-point score, the genres and the word groups used are fewer while obtaining the better accuracy. This may be because the T-score divides into the smaller number of levels which affects the number of nodes in the tree.

In the experiments, we test the model against the movie samples from IMDb for 240 movies. We pick 80 each for PG, PG-13 and R ratings. Using the model with mid-point level, we can obtain about 80% accuracy. In particular, Table 2 shows the results of precisions, recalls, and F-measure.



Classifier: J48 -C 0.25 -M 3

Test options: ○ Use training set; ○ Supplied test set; ● Cross-validation Folds 4; ○ Percentage split % 66

(Nom) Rated

Result list (right-click for options): 17:49:20 - trees.J48

Status: OK

Classifier output

```
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances         175               72.9167 %
Incorrectly Classified Instances        65               27.0833 %
Kappa statistic                          0.5937
Mean absolute error                      0.2298
Root mean squared error                  0.3621
Relative absolute error                 51.7119 %
Root relative squared error             76.8232 %
Total Number of Instances              240

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.763     0.281      0.575     0.763     0.656      0.758    PG-13
                 0.788     0.05       0.887     0.788     0.834      0.935    PG
                 0.638     0.075      0.81      0.638     0.713      0.866    R
Weighted Avg.    0.729     0.135      0.757     0.729     0.735      0.853

=== Confusion Matrix ===

  a  b  c   <-- classified as
 61  7 12 |  a = PG-13
 17 63  0 |  b = PG
 28  1 51 |  c = R
```

Fig 5. Weka Results When Using Mid-Points to Levelize Word Groups.

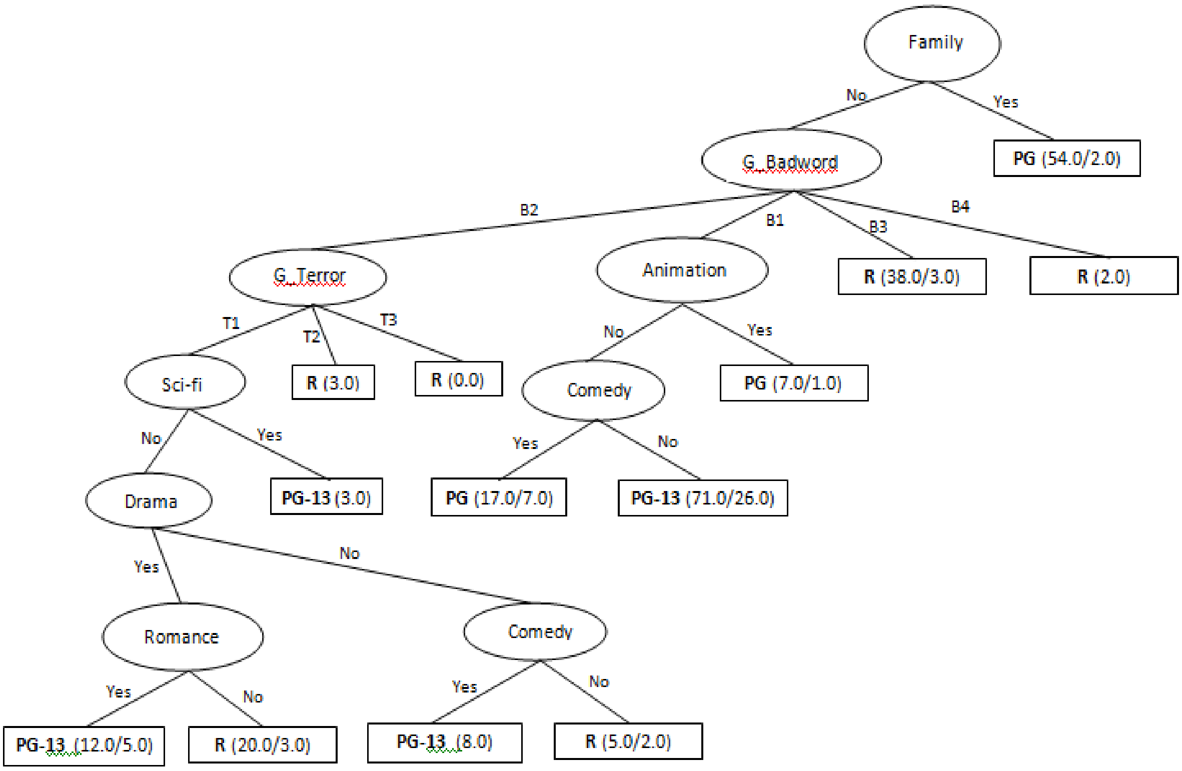


Fig 4. Decision Tree When Using Mid-Points to Levelize Word Groups.

Table 1. Comparison of Both Approaches in Training Data.

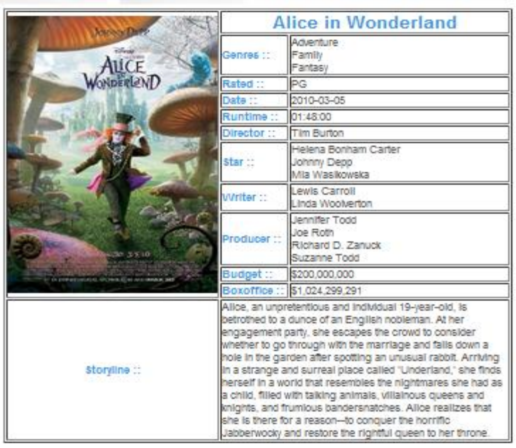
Level approaches	Weka Selected Attributes		Correctness from 240 movies			
	Selected Genres	Selected word groups	Correct		Incorrect	
Mid-point	Family Animation Drama Sci-fi Comedy Romance	Bad word Terror	175	73%	65	27%
T-Score	Family Animation Fantasy Thriller Drama Crime Action Comedy	Bad word Sexual Terror Drug	161	67%	79	33%

Table 2. Precision, Recall, F-Measure for Test Data.

Rating	Precision (%)	Recall (%)	F-measure (%)
PG	88.7%	78.8%	83.4%
PG-13	57.5%	76.3%	65.6%
R	81%	63.8%	71.3%
Average	**73%**	**76%**	**80%**

V. APPLICATIONS

To be complete, in Figure 6, we develop a web application that deploys the classification model. First, the application contains the movie information as stored in the database. Then, in Figure 7 the movies can be queried or searched from the database. Also, the new movie can be inserted with the information as well as the classified suggestion is performed with our Weka model in Figure 8.



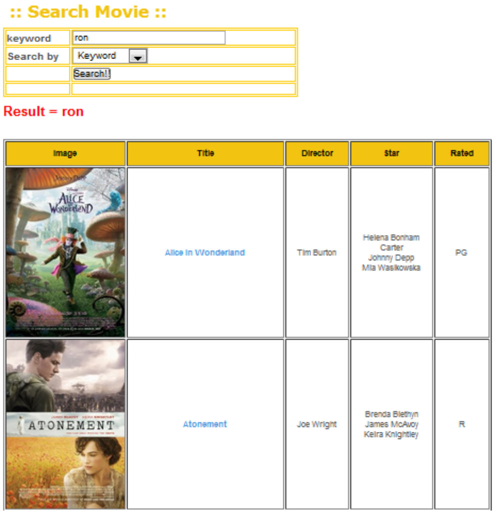
Alice in Wonderland

Genres ::	Adventure Family Fantasy
Rated ::	PG
Date ::	2010-03-05
Runtime ::	01:48:00
Director ::	Tim Burton
Star ::	Helena Bonham Carter Johnny Depp Mia Wasikowska
Writer ::	Lewis Carroll Linda Woolverton
Producer ::	Jennifer Todd Joe Roth Richard D. Zanuck Suzanne Todd
Budget ::	$200,000,000
Boxoffice ::	$1,024,299,291
Storyline ::	Alice, an unpretentious and individual 19-year-old, is betrothed to a dunce of an English nobleman. At her engagement party, she escapes the crowd to consider whether to go through with the marriage and falls down a hole in the garden after spotting an unusual rabbit. Arriving in a strange and surreal place called "Underland," she finds herself in a world that resembles the nightmares she had as a child, filled with talking animals, villainous queens and knights, and frumious bandersnatches. Alice realizes that she is there for a reason—to conquer the horrific Jabberwocky and restore the rightful queen to her throne.

Fig 6. Web Application User Interface Showing the Movie Information.

VI. CONCLUSION

In this work, we present a data mining application. We apply the data mining to perform the moving rating in the prototypes. Many movie attributes are studied such as the producers, writers, actors, actresses, genres, subtitles etc. Based on study, it is found that the most effective attributes are genre and words used in the movies. The studied methodology is described. The database of movie information are created. These information are extracted, cleaned and transformed to Weka to create a decision tree. The difficult part is the extraction of words in the subtitles/short stories, selection of keywords, and keyword classifications that affect the rating. The derived model can perform about 80% accuracy with the usage of genre and word group attributes. The model is embedded in the web application that stores movie information and suggested the moving rating. Also, the approach can be further extended to consider other attributes and image processing techniques to extract contents for classification.



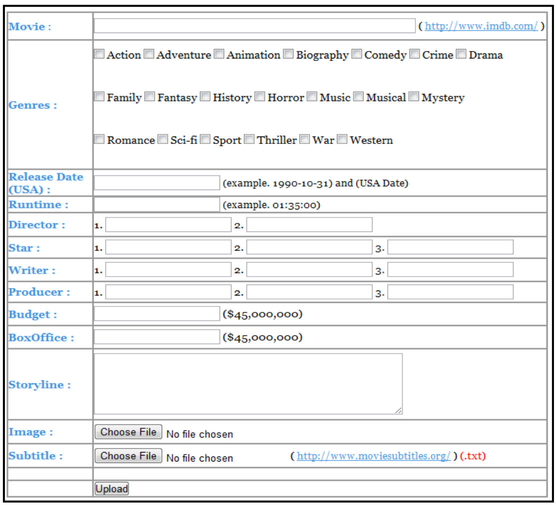
:: Search Movie ::

keyword	ron
Search by	Keyword
	Search!!

Result = ron

Image	Title	Director	Star	Rated
	Alice In Wonderland	Tim Burton	Helena Bonham Carter Johnny Depp Mia Wasikowska	PG
	Atonement	Joe Wright	Brenda Blethyn James McAvoy Keira Knightley	R

Fig 7. The User Interface of Searching Movie Information with the Rating.



Movie :	(http://www.imdb.com/)
Genres :	□ Action □ Adventure □ Animation □ Biography □ Comedy □ Crime □ Drama □ Family □ Fantasy □ History □ Horror □ Music □ Musical □ Mystery □ Romance □ Sci-fi □ Sport □ Thriller □ War □ Western
Release Date (USA) :	(example. 1990-10-31) and (USA Date)
Runtime :	(example. 01:35:00)
Director :	1. 2.
Star :	1. 2. 3.
Writer :	1. 2. 3.
Producer :	1. 2. 3.
Budget :	($45,000,000)
BoxOffice :	($45,000,000)
Storyline :	
Image :	Choose File No file chosen
Subtitle :	Choose File No file chosen (http://www.moviesubtitles.org/) (.txt)
	Upload

Fig 8. The User Interface For Entering Movie Information Used To Store in The Database and Classify.

APPENDIX

We list examples of word groups used in the model in the decision tree.

Table 3. Bad Words

ass	Asshole	bastard
bitch	Bullshit	crap
cunt	Damn	dick
fuck	Goddamn	idiot
motherfucker	Pussy	shit
suck		

Table 4. Sexual words

affair	Kiss	randy
rape	Sex	squeeze
strip	Tom	gay
assault		

Table 5. Terror words

dead	Death	die
afraid	Hell	kill
awesome	Soul	murder
awful	Blood	bloody
bury	Coffin	crime
dread	Evil	fear
funeral	Ghost	killer
murderer	Scare	scream
shout	Terrible	horror

Table 6. Drug words.

bar	Beer	bet
wine	Cheat	cigarette
drug	Illegal	smoke
heroin	Cocaine	alcoholism
addict	Dope	drink

REFERENCES

[1] "Internet movie database," IMDb.com, Inc. , [Online]. Available: http://imdb.com. [Accessed 25 June 2012].

[2] "Weka 3: Data Mining Software in Java," University of Waikato, [Online]. Available: http://www.cs.waikato.ac.nz/ml/weka/. [Accessed 25 June 2012].

[3] J. W. Seifert, "Data Mining: An Overview," Congressional Research Service, 2004.

[4] X. Amatriain, A. Jaimes, N. Oliver and J. Pujol, "Data Mining Methods for Recommender Systems," in Recommender Systems Handbook, F. Ricci, Ed., Springer Science Business Media, 2011, pp. 39-71.

[5] R. Groth, "Industry Applications of Data Mining," in Data Mining: Building Competitive Advantage, Prentice Hall, 2000, pp. 191-210.

[6] K. Raza, "Applications of Data Mining in Bioinformatics," Indian Journal of Computer Science and Engineering, vol. 1, pp. 114-118, 2012.

[7] J. Oh, J. Lee, K. Sanjay kumar and B. Bandi, "Multimedia Data Mining Framework for Raw Video Sequences," in ACM Third International Workshop on Multimedia Data Mining (MDM/KDD2002), 2002.

[8] L. Zhu, M. Zhu and S. Yao, "The popularity of movies predict system based on data mining technology for CDN," in IEEE International Conference on the 3rd Computer Science and Information Technology, , 2010.

[9] J. Han, "Data Mining for Image/Video Processing: A Promising Research Frontier," in CIVR, Ontario, Canada, 2008.

[10] M. Saraee, S. White and J. Eccleston, "A data mining approach to analysis and prediction of movie ratings," in Data Mining V, WIT Press, 2004, pp. 343-352.

[11] M. Fleischman, P. Decamp and D. Roy, "Mining Temporal Patterns of Movement for Video Content Classification," in MIR, Santa Barbara, CA, USA, 2006.

[12] P. Changkaew and R. Kongkachandra, "Automatic Movie Rating Using Visual and Linguistic Information," in NCCIT, Bangkok, Thailand, 2010.

[13] P. Chaovalit and L. Zhou, "Movie Review Mining: a Comparison between Supervised and Unsupervised," in the 38th Hawaii International Conference on System Sciences, Hawaii, USA, 2005.

[14] "Film Rating," Motion Picture Association of America, [Online]. Available: http://www.mpaa.org/ratings. [Accessed 25 June 2012].

[15] "Classification Website," Australian Government, [Online]. Available: http://www.classification.gov.au/Pages/default.aspx. [Accessed 24 June 2012].

[16] "British Board of Film Classification," [Online]. Available: http://www.bbfc.co.uk/. [Accessed 20 June 2012].