2014

ISSN : 0974 - 7435

Volume 10 Issue 19

BioTechnology

An Indian Journal

FULL PAPER

BTAIJ, 10(19), 2014 [11536-11542]

Effective application of information model in the data sorting process of music recommendation

Rongli Ju
North China University of Water Resources and Electric Power, Zhengzhou, 450046, (CHINA)

ABSTRACT

In the process of effective collection and sorting, music recommendation data needs to make appropriate model analysis to give full guarantee to its effectiveness. Combining the data of Knowledge Discovery Cup Challenge in 2011 Track2, the study makes the corresponding research process, in which it collects and sorts the total number of users and the total entry of data effectively to make the data can be effectively established. And then, it defines the training data and the users' item data to make sure the positive example and counter example maintain a high degree of clear. In the process, it transmits these item data and positive example, counter example to the two information models (regression model and scheduling model) to estimate the error probability effectively and determine the final information model. This is the main purpose and related research process of this study. Through the application of this process, it guarantees the accuracy and validity of the information model, making the music recommendation data to achieve the appropriate transparency and the data analysis process have a more strong rationality. From the aspect of the effect of information model, the analysis process of this study not only can make the specific results of the experiment more convincing, but also can ensure the analysis process and the analysis step of the data achieve more comprehensive, so as to improve the validity of recommendation data continuously.

KEYWORDS

Music recommendation; Data sorting; Information model; Application research.

INTRODUCTION

From the data sorting process of music recommendation, the accuracy of the data analysis process is the essential prerequisite for the formation of the ideal standard of recommendation data. However, the importance of effective application of the information model can be fully reflected. In the process of research and discuss, first, this study introduces the data sets in detail. At the same time, it determines the evaluation criteria effectively. Then, it enables to play its role and maintain a high degree of clear for the establishment direction of the best model. This makes the research and discussion of this study more reasonable and scientific, so as to lay a solid foundation of theory and practice for the thorough development of the future research work.

INTRODUCTION TO DATA SETS

In the discussion process of this section, the study mainly makes the elaboration for the data sets combining with the corresponding experimental process. However, before the introduction to the function of the data sets, the readers should know purpose of data sets. Now, this point will be introduced in the following part. During the course of the experiment, the data sets are the scoring data sets opened by Yahoo Institute for the Knowledge Discovery Challenge Cup in 2011 Track2. The number of the users of training data sets has reached 249,012 and the scoring data items have reached 296,111. In the effective scoring process of these data, the total items of scoring data have reached 61,944,406From the data in TABLE 1, among the different data sets; it is easy to see the number and the proportion of items[1]. In TABLE 2, the total number and proportion of different items and scoring data in the training data sets can also be fully reflected. Making contrast between TABLE 1 and TABLE 2, this can effectively calculate the scoring intensity of the different types of items, so as to make a result with a high degree of accuracy. However, from the translation of the data in TABLE 3, we can see that in the training data, the items of the singer's style, the album’s type and the singer's characteristics are more than the average number of scoring significantly. Under such condition, the song itself reflects less information. When it can not meet the basic requirements, then you can make effective supplement about the information of the singer's style and the album’s type. The fundamental prerequisite of consciousness can get corresponding satisfaction.

It can sum up the specific relationship existed in the scoring data and scoring items from the above characteristics of the data set. However, this data set can also reflect another feature, which is the structural information among data items has certain regularity. This study will make an effective discussion for the feature in the following part and show the affect of structural information to the Music data recommendation to prove the important position of the data sets in the construction and application of information model[2].

TABLE 1 : The number of different types of items

	Track	Album	Artist	Genre	Total
#	224041	52829	18674	567	296111
%	75.66	17.84	6.31	0.19	100

TABLE 2 : The number of scoring data of different items in the training data

	Track	Album	Artist	Genre	Total
#	27167857	11928316	19289882	3558351	61944406
%	43.86	19.26	31.14	5.74	100

TABLE 3 : The number of average score of different items in the training data

	Track	Album	Artist	Genre	Total
#(Average)	121.26	225.79	1032.98	6275.75	209.19

From Figure 1, the readers can see the scores’ specific distribution of the given training data. And in the figure, the scoring data of the users themselves demonstrate an obvious trend. This trend is mainly concentrated on polarization and the extreme phenomenon is serious. The proportion of zero and 90 is more than others in data points, while the proportion of other scores is small relatively.

In the research and discussion of this paper, the test data also comes from the data sets disclosed by Yahoo Institute. There are 101,172 data users in it. For each user, the organizer gives them six corresponding items. And the types of the six items all are songs. In the scoring process of the six items, half of them adopt the method of the users scoring themselves, so the scores are higher than others[3].

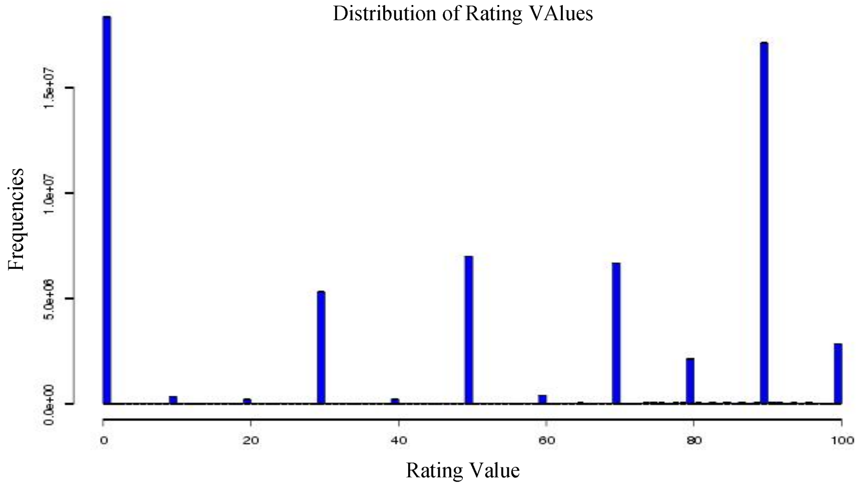

Figure 1 : The distribution of scores in training data

EVALUATION CRITERION

However, in this process, there are three users are not involved in the scoring process. But from the foregoing discussion, it can be obtained that the users involved in scoring give the high scores. So here regards the three users' items involved in scoring as the positive examples in the test data, and the other three did not participate in scoring are called counter examples. From this discussion, the probability to be selected as examples between scoring data items and the items involved in data scoring is equal. At the same time, they have positive relationship. The calculation formula is:

$$p_r(i) = \frac{\sum_{u,j} I(j = i \,\&\&\, r_{u,j} \geq 80)}{\sum_{u,y} I(r_{u,j} \geq 80)} \tag{1}$$

In the formula mentioned above, its main purpose is to build the corresponding training model by the training data effectively, making the test data model in the process of training can be embodied in each user. The six items contained in this data are divided into three groups[4]. Thus, the data which can make effective evaluation are determined by the error rate, also by the proportion of the total samples, which is calculated as follows:

$$ErrorRate = \frac{\sum_{u \in \Gamma_{(u)}} \sum_{i=1}^{6} I(\bar{\hat{l}}_{u,i} \neq l_{u,i})}{6 * |\Gamma(u)|} \times 100 \tag{2}$$

In the formula mentioned above, $\Gamma_{(u)}$ represents the number collection of users in the data sets. $\hat{l}_{u,j}, l_{u,j} \in \{+1, -1\}$ Represent the predicted and actual results. However, in the course of research and discussion, this paper usually regards the series of questions as recommendation questions of flow analysis. From the simple sense, each user corresponds to the 6 items. By the model, the data is analyzed effectively and the prediction order is ranked correspondingly, while the top three items as positive examples, the prediction result is $\hat{l} = 1$, the last three items were called counter examples, the test result is expressed by $\hat{l} = -1$. At last, it makes the final effective evaluation process by using the data evaluation methods in TABLE 1.

THE FUNCTION OF INFORMATION MODEL AND THE ESTABLISHMENT OF OPTIMAL MODEL

The function of information model

In the research and discussion process of this part, the paper mainly makes an effective discussion for the specific function of information model by the specific results of the experiment. During the course of the subsequent discussion, it will involve some relevant abbreviations of the professional terminologies and TABLE 4 will give a representation. TABLE 4 mainly involves in the abbreviations of loss function and sorting model pairing method. The specific content is shown in TABLE 4.

TABLE 4 : The corresponding abbreviations of different loss functions and parameters

ReSVD	Loss function(4-1)
RaSVD+GAP+PAIR	Loss function(4-12), $\delta_{uij} Equation(4-13)$
RaSVD+BOUND_PAIR	Loss function(4-12), $\delta_{uij} Equation(4-14)$

TABLE 5 : The abbreviations of different prediction models

BSVD	Basic SVD, Equation(5-2)
Tax-SVD	Taxonmv-Aware SVD, Equation(5-6)
+IMFB	Implicit feedback, Equation(5-9)
+Item-10NN	Top 10 Item neighborhood, Equation(5-11)
+Tax-CLF	Taxonomy based classifier, Equation(5-14)

From TABLE 5, the readers can see that the different information models have corresponding abbreviated form, in which the "+" represents the continuous accumulation. For example, the last row in TABLE 5, the basic model it represents is an information model with appropriate hierarchical matrix factorization, implicit feedback.

As can be seen from TABLE 6, in the selection process of counter-examples, this paper makes a random selection, making the number of counter-examples become 3 times than before, and the error probability of different information models can be reflected more clearly. From this table, a conclusion can be obtained. It is that in the information models, for the new information integration process, its error probability will reduce significantly and the accuracy of its information can be guaranteed effectively[5]. For example, the ranked SVD, during the build process of the matrix decomposition model, increasing the item structure information effectively, the error probability can be reduced by 6.5%. Based on this, joining the implicit feedback information, the error probability is reduced by 9.2%. And, adding the proximity information between each item in this form continually, the error rate can be reduced 32.4%. In the end, adding the hierarchy resolver into the model, it also can reduce the error probability by 22.6%, and can get the final error probability.

TABLE 6 : The corresponding error rate of different information models

	ReSVD f=100	RaSVD+GAP_PAIR T=20,f=100	RaSVD+BOUND_PAIR $t_{lb} = 20, t_{ub} = 0, f = 100$
BSVD	6.52	6.42	6.76
Tax-SVD	6.15	6.03	6.03
+IMFB	5.71	5.52	5.58
+Item-10NN	3.98	4.17	4.18
+Tax-CLF	3.76	3.40	3.33

TABLE 7 : The corresponding error rate of different information models

	ReSVD f=100	RaSVD+GAP_PAIR T=20,f=100	RaSVD+BOUND_PAIR $t_{lb} = 20, t_{ub} = 0, f = 100$
BSVD	23.99	25.20	29.40
Tax-SVD	20.79	22.74	26.92
+IMFB	17.09	17.03	16.51
+Item-10NN	14.23	15.70	15.63
+Tax-CLF	9.63	12.19	13.78

From the data in TABLE 7, it can be seen that when each information model extracts the counter examples effectively, the error probability produced by them are also different. In this table, both the regression model and ordering model, when there is a new added item, the error probability of the model itself can always be constantly reduced[6]. The results reflected in the table are same with the results discussed in the previous form, so as to further demonstrate the effectiveness and accuracy can be guaranteed when the information models carry out the analysis process of the recommendation data. And in the specific discussion of the part, the minimum error probability of each model also was introduced.

Optimal model

In the research and discussion process of this part, firstly, the paper identifies the best results of the regression model and the sort model. Then, it analyses the data obtained from the experiments through all the information models. And the corresponding data obtained from the experimental results is divided into two parts to be reflected, respectively, comparing with the best single model of each result and comparing with the hybrid model of every result.

TABLE 8 : The lowest error rate of regression model and sort model

	Best Results	Parameters
ReSVD	3.78	$\theta = 60, k = 3, f = 100$
RaSVD+BOUND_PAIR	3.16	$\theta = 60, k = 5, t_{lb} = 20, t_{ub} = 0, f = 100$
RaSVD+GAP_PAIR	3.10	$\theta = 60, k = 5, t = 40, f = 300$

In the above TABLE 8, the error rata of regression models and sort models is reflected precisely and the setting of parameters is shown clearly in this table. Among this, f is the specific dimensions of the hidden space, which are the specific dimensions of the users' vectors and the items' vectors. And θ represents the specific range of the items of counter examples, in which only the data that exceeds the range of scores should be recorded so that the counter examples can make a random selection effectively. K represents a specific multiple of the counter example. The meaning of t_{lb} and t_{ub} are the scoring on-lines of direct proportion and the scoring on-lines of inverse proportion in BOUND-PAIR. T represents the minimum point's difference between two opposed items in GAP-PAIR. However, comparing the data in TABLE 6 and TABLE 8, the readers can draw a conclusion: in the condition of not guarantying the adequate items of the inverse proportion, during the process of the recommendation music data, the regression models have more advantages than sorting models. In the continuous increasing process of the items of inverse proportion, during the process of the recommendation music data, sorting models are superior to the regression model. The error probability can be controlled effectively and its error probability can be reduced continually. However, through the study on the data in TABLE 8, the conclusion that the best results of sorting model were superior to regression model can be drawn.

As can be seen in TABLE 9, under the circumstance of giving the information model, making a corresponding scientific comparison for the top ten' results of this competition, the data in this table can reflect such a conclusion: information model is good at analyzing the best results for a single model, so that in the analysis process of the data of model results, it has been beyond the hybrid model.

Influence of counter examples extraction

In TABLE 10, the paper shows the influence of extracting counter examples multiples to the recommendation accuracy. As can be seen from this table, increasing the counter examples multiples is always can reduce the error probability of the test data sets, but the decreasing range becomes smaller and smaller. For example, in RaSVD + BOUND_PAIR, when the multiples of counter examples increase from one time to twice, the error probability decreases by 17.8%. Based on this, when the counter examples increase one time, the error rate decreases by 7.4%. When the counter examples multiples increase to four times, the error rate decreased by 5.6%. At last, when extracting five times counter examples, the error rate is decreased by 1.9% than four times counter examples extraction and becomes 3.16%. Comparing TABLE 6 and TABLE 7, under the circumstance of extracting three times counter examples, the effect of the easiest BSVD model is better than the effect of Tax-CLF model, which also explains the important role of counter examples extraction for the improvement of the accuracy.

TABLE 9 : Comparison between the best results of information model and the results which have been published on this issue

Team name	Best single predictor	Ensemble modle
InfoSVD	3.10	
Nation Taiwan University	4.04	2.47
The Art of Lemon	3.49	2.48

commendo	4.28	2.49
The Though Gang	3.71	2.93
The Core Team		3.87
False Positives	5.70	3.89
Opera Solutions	4.45	4.38
MyMediaLite	6.04	4.49
KKT’s Learning Machine	5.62	4.63
coaco		5.20

TABLE 10 : The influence of counter examples extraction multiples to the experimental results

	ReSVD **f=100**	**RaSVD+GAP_PAIR** **T=20,f=100**	**RaSVD+BOUND_PAIR** $t_{lb}=20, t_{ub}=0, f=100$
k=1	4.48	4.09	4.30
k=2	4.14	3.66	3.65
k=3	4.05	3.54	3.40
k=4	3.96	3.45	3.22
k=5	3.92	3.40	3.16

Parameters adjustment

In the research and discussion of this part, the paper mainly explores the influence of parameters on data in the sorting model as well as the formation process error probability. As can be seen in Figure 2, when giving a maximum limit and minimum limit of the examples and counter examples, it will has an impact on the accuracy of music recommendation data. However, it can be found in the figure that after determining the upper limit of counter examples, with the upper limit of positive examples increasing, the error probability of models is also increasing constantly[7].

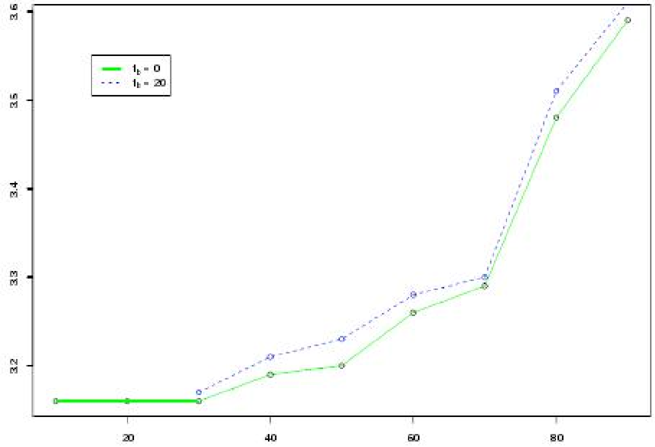

Figure 2 : The influence of different tub and tlb to the accuracy of recommendation in BOUND_PAIR

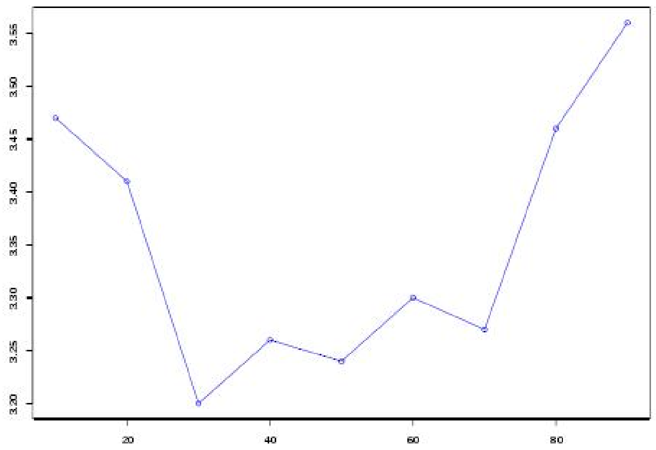

Figure 3 : The influence of different t to the accuracy of recommendation in GAP_PAIR

As can be seen in Figure 3, in the GAP-PAIR model, different t causes different effects on the error probability of models. From the figure above, if stopping the growing trend, the error probability of models will continue to decrease, and subsequently showing an increasing trend. The main reason is that when the t is too small, the relationship between the items of positive examples and counter examples can not be effectively guaranteed, leading to the error data of training data is introduced. However, when the t increases constantly, even to excessively large, although the positive examples and counter examples reliability can be ensured between the entries, the condition of the items of training data reducing gradually would results in the insufficient of training data. However, when ensuring the error probability can be reduced to a minimum, t should be in the range of about 30.

CONCLUSION

The above is the study process of the effective application of information model in the data sorting process of music recommendation. The paper mainly studies the data sets and the function of the information models. Also, it analyzes the main factors which influence the error probability of models specifically, so as to guarantee the efficiency and accuracy of the application of information model in the data sorting process of music recommendation.

ACKNOWLEDGEMENT

Fund project: the philosophy and society planning project of Henan Province in 2013, number: 2013BTY011

REFERENCE

[1] Ou Xiaoping, Wang chaokun, Peng zhuo, Qiu ping, Bai yiyuan; A graph-based music data model and query language, Journal of Computer Research and Development, **48(10)**, 1879-1889 **(2011)**.

[2] Zhou lijuan, Lin hongfei, Yan jun; Music information retrieval model based on TLDA and SVSM, Computer Science, **41(2)**,174-178 **(2014)**.

[3] Lv lanlan; Music emotion fuzzy computing model based on evolving kernel clustering, Pattern Recognition and Artificial Intelligence, **25(1)**, 63-70 **(2012)**.

[4] Hao da; Music emotion theory in aesthetics perspective, Social Science Front bimonthly, **(12)**, 244-245 **(2012)**.

[5] Jiang shengyi, Li xia; Survey on music emotion automatic analysis, Computer Engineering and Design, **(18)**, 4112-4115 **(2010)**.

[6] Liuyi; From construction to achieve: Talk about music emotion system and significance, Sichuan Drama, **(2)**, 69-71 **(2014)**.

[7] Huang zongquan; "Music rhetoric" and the historical perspective of baroque music emotion expression paradigm. People's Music: Comment, **(8)**, 88-91 **(2012)**.