Evaluation of a TV Programs Recommendation using the EPG and Viewer's Log Data

Kazuki Ikawa
Graduate School of Engineering,
The University of Tokyo
7-3-1 Hongo, Bunkyo-ku
Tokyo, Japan

Tomohiro Fukuhara
National Institute of Advanced Industrial Science and Technology
2-3-26 Aomi, Koto-ku
Tokyo, Japan
tomohiro.fukuhara@aist.go.jp

Hideki Fujii
Graduate School of Engineering,
The University of Tokyo
7-3-1 Hongo, Bunkyo-ku
Tokyo, Japan
fujii@save.sys.t.u-tokyo.ac.jp

Hideaki Takeda
National Institute of Informatics
2-1-2 Hitotsubashi, Chiyoda-ku
Tokyo, Japan
takeda@nii.ac.jp

ABSTRACT

Evaluation results of a TV program recommendation using EPG (Electronic Program Guide) and viewer's log data are described. Two experiments using the log data of a Japanese video service provider are conducted: (1) an experiment of prediction of TV programs that each viewer usually watched, and (2) a survey of TV program recommendation using the log data and a questionnaire. High precision values are obtained in the first experiment using *keyword-based* and *celebrity-based* recommendation methods for terrestrial and satellite broadcasting programs. Meanwhile, the *channel & time-based* recommendation method for programs in paid channels of the video service provider obtained the high precision in the second experiment. An overview of experiments and results are described.

Categories and Subject Descriptors

H.4.2 [**Information Systems Applications**]: Types of Systems

General Terms

EXPERIMENTATION

Keywords

TV program recommendation, evaluation

.

1. INTRODUCTION

Today, many TV programs are broadcasted on various channels. Terrestrial digital broadcasting services and satellite broadcasting services brought us many channels and programs. As the increase of channels and programs, recommendation services for TV programs and channels are needed that allow us to choose suitable programs to watch for each use[4]. For designing a good TV program recommendation service, fundamental data that compares performances among various recommendation methods is needed[6].

We had two experiments of a TV program recommendation using EPG (electronic program guide) and the viewer's log data: (1) an experiment of prediction of TV programs that each viewer watches usually, and (2) a survey of TV program recommendation using the log data and a questionnaire. We created a prototype system of a TV program recommender system that recommends programs for each user using EPG and the viewer's log data. An overview of experiments and results are described.

In the following, section 2 describes the related work. Section 3 describes an overview of our approach for TV program recommendation. Section 4 describes an overview of our experiment and its results. Section 5 discusses the results of the experiments. We summarize arguments in section 6.

2. RELATED WORK

There are many work on TV program recommendation. TiVo, which is a TV program recommendation system based on the collaborative filtering (CF) method, is a popular product in U.S[1]. Although the CF method works well where a large number of judgement data is available, we consider that the CF method have difficulties when new programs that few people watched before.

Smyth and Cotter introduces an overview and performance of a personalized TV guide system[8]. They also adopted the CF method. In their experiment, 310 households attended to evaluate the precision of recommendation, and found that

Table 1: An example of a viewer's log (dummy data)

Household ID	Date	Time	Channel ID
1000	01/Nov/2009	07:18	235
1000	01/Nov/2009	07:19	235
1000	01/Nov/2009	07:20	234

Table 2: An example of EPG data.

Channel ID	Start Time	End Time	Title	Description
235	01/Nov/2009 07:00:00	01/Nov/2009 07:59:00	***	***

61% of households answered positively. Their result showed that the CF method showed the best performance, compared to the content-based (CB) recommendation and a naive random recommendation.

Although there exist recommender systems based on the CF method, few data is reported with respect to the evaluaion of the CB recommendation. In this paper, we will evaluate the performance of a CB recommendation method using the real data of a video service provider.

3. APPROACH FOR THE TV PROGRAM RECOMMENDATION

3.1 Log data

We used the viewer's log data of a video service provider in Japan. We call this provider *TV-999* in this paper. Each household has a set top box (STB) that receives TV programs of *TV-999*. *TV-999* broadcasts terrestrial programs, satellite broadcasting (BS[1]) programs, and *paid channels* such as movie, drama, sports, animation and so on[2]. The view history of a household is recorded at the server machine of the provider[3].

Table 1 shows an example of log data. The log data consits of following parts: (1) *household ID*, (2) *date and time* when the household watched a channel, and (3) *Channel ID* at which the household watched. The data is recorded in each minute. In this case, household ID 1000 watched channel ID 235 and 234 during 7:18 through 7:20 a.m. on November 1st, 2009. For zapping action, we removed the view history of which duration is under five minutes.

3.2 EPG data

EPG data contains *content ID*, *date and time* when the content is broadcasted, *duration* of the content (minute), *title* and *summary* of the content. Table 2 shows an example of EPG data. We extract keywords and names of celebrities appeared in title and description. We can identify which household watched which programs by combining EPG and the log data[5].

3.3 Recommendation method

Based on programs that each household watched, we first calculate feature values of keywords and celebrities using following formulas[2][3][7].

$$P_{key}(k) = \frac{|Watched(k)|}{|Programs(k)|} \quad (1)$$

$$P_{person}(p) = \frac{|Watched(p)|}{|Programs(p)|} \quad (2)$$

$P_{key}(k)$ is a function that calculates a feature value for a keyword k. $Watched(k)$ is a set of programs that a household watched. $Programs(k)$ is a set of programs that contains the keyword k in their EPG data. $P_{person}(p)$ is a function that calcuates feature value for a celebrity person p. $P_{key}(k)$ and $P_{person}(p)$ are calculated for each household.

For calculating recommendation values of a program *title*, we use following formulas[9].

$$R_{key}(title) = \sum_{k \in Keywords(title)} P_{key}(k) \quad (3)$$

$$R_{person}(title) = \sum_{k \in Persons(title)} P_{person}(p) \quad (4)$$

$$R_{mix}(title) = \frac{R_{key}(title)}{\max_{t \in TITLES} R_{key}(t)} + \frac{R_{person}(title)}{\max_{t \in TITLES} R_{person}(t)} \quad (5)$$

$R_{key}(title)$ is a function that calculates a recommendation value for a program *title* containing a keyword k in its EPG data. $Keywords(title)$ is a function that returns a set of keywords contained in the EPG data of a program *title*. Formula (3) is used as a *keyword-based* recommendation in the next section.

$R_{person}(title)$ is a function that calculates a recommendation value for a celebrity *person* appeared in the EPG of program *title*. $PERSONS$ is a set of all of persons appeared in all of EPG data. Formula (4) is used as a *celebrity-based* recommendation.

$R_{mix}(title)$ is a function that calculates recommendation value for a program *title*. $TITLES$ is a set of all titles appeared in all of EPG data. Formula (5) is used as a *keyword & celebrity-based* recommendation.

4. EXPERIMENTS

4.1 Prediction of daily viewing programs

Approach

We evaluated the accuracy of prediction of programs that each household daily watches. We used the log data in February 2008. The data is divided into twofold: (1) the first part (February 1st to 15th 2008) is the *training data* which is used for calculating feature values of keywords and celebrities. (2) the second part (February 16th to 29th 2008) is the *test data* that is used for evaluating prediction of programs.

The log data contains the view history of 923 households. 111, 343 programs are contained in this data. Among these programs, 29, 887(26.8%) are terrestrial/BS programs, and 81, 456(73.2%) are programs of paid channels.

[1] BS is a common name of one of the direct broadcast satellites in Japan.

[2] There are 30+ paid channels in *TV-999*.

[3] The log data is recorded under the agreement of each household. If a household does not want, the data is not recorded.

With respect to programs watched by households, 669 programs are watched by each household in the test stage. Among these programs, 70.1% (469) are terrestrial / BS programs, and 29.9% (200) are programs of paid channels.

We prepared two types of recommendations: (1) *keyword-based* recommendation, and (2) *celebrity-based* recommendation. For the first type of recommendation, we use the formula (3) to calculate a recommendation value of a program for a household. For the second type of recommendation, we use the the formula (4) to calculate recommendation a value of a program for a household. We randomly picked up 10 households, and evaluated 2*100 programs, which is created by two types of recommendations, for each household. We used terrestrial / BS programs because the large amount of log data was available for these programs.

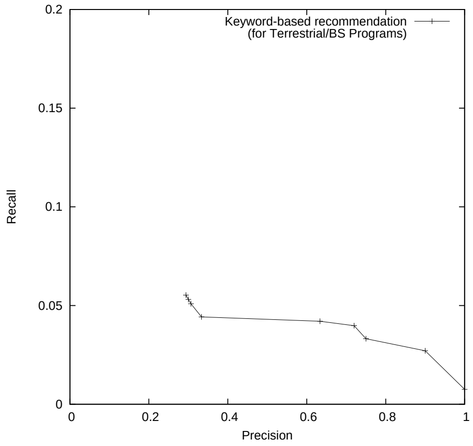
As evaluation measures, we used precision and recall measures described in the following.

$$\begin{aligned} Precision &= R/N & (6) \\ Recall &= R/C & (7) \end{aligned}$$

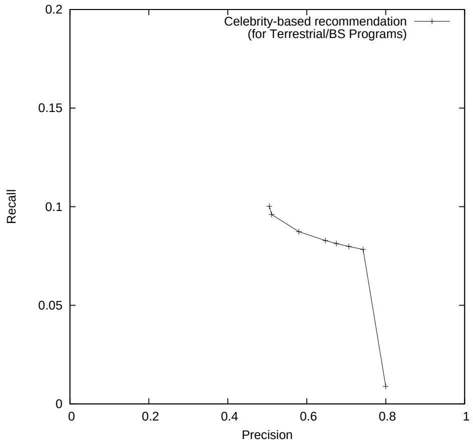
R is the number of programs that the system suggested and a household watched, and N is the number of programs that the system suggested. C is the number of programs that a household watched.

Results

Figure 1(a) and Figure 1(b) show the results for *keyword-based* and *celebrity-based* recommendations respectively. In these figures, the recall values are quite low because the number of C in the formula (7) is much larger than the number of programs that we recommended in the experiment[4]. So, the recall values are low. With respect to the precision, both of recommendation types showed high precision values. We considered that predicting programs that a household usually watches based on *keywords-based* and *celebrities-based* recommendation was possible.



(a) Recall and precision of the keyword-based recommendation for terrestrial/BS programs.



(b) Recall and precision of the celebrity-based recommendation for terrestrial/BS programs.

4.2 Evaluation of recommendations using a questionnaire

Approach

Based on the results of the first experiment, we had a survey of TV program recommendation using keywords and celebrities. We prepared a questionnaire containing recommended programs for households using the log data. In this experiment, we chose programs in paid channels.

Table 3 shows the periods of this experiment. As a training period, we used the log data during November 9th through 22nd (14 days), 2009. As a test period in which TV programs are recommended in the questionnaire, we chose the period during November 28th to December 4th (7 days). For avoiding the change of preferences of each household[5], we sent the questionnaire after the test period, i.e., each household received the questionnaire after December 5th.

The questionnaire contains programs recommeded by *keyword-based* and *celebrity-based* recommendations for each household. We asked each household up to five programs for each day, i.e., up to 35 programs broadcasted in the test period are asked. We asked each household to judge programs according following four categories: (1) *Watched & Good*, (2) *Watched & ¬Good*, (3) ¬ *Watched & Expected*, and (4) ¬ *Watched & ¬Expected.* Each household checked one of these categories for each recommendation.

For comapring performances of recommendation types, we classified households into four types. Table 4 shows the list of recommendation types. Type A&B is the mixture of type A and B, which is calculated using the formula (5). Type C is the baseline type where programs are chosen randomly according to the joint probability of channels and time-slots that a household usually watches TV. This probability is calculated from the log data of a household in the training period. We used the log data of 907 households of the service. We sent a questionnaire containing recommendation of programs for each household. Questionnaires are sent to 875 households via mail on December 4th, 2009[6]. By December 25th, 2009, 605 answers (69.3%) are returned.

[4] We only evaluated 100 programs for each household

[5] This was a request from *TV-999.*

[6] Households received the questionnaire after December 5th, 2009.

Table 3: Periods of experiment.

Dataset	Term	# of Days
Training	Nov.9th - Nov.22nd, 2009	14 days
Test	Nov.28th - Dec.4th, 2009	7 days

Table 4: Types of recommendation.

Rec.		# of Received/
Type	Description	Recommended (Ratio)
A	*Keyword-based*	112/160 (70.0%)
B	*Celebrity-based*	141/200 (70.5%)
A&B	*Keyword&Celebrity*	132/200 (66.0%)
C	*Channel&Time*	220/315 (69.8%)

Results

Table 5 shows the summary of the result. Most of recommendations were judged as ¬*Watched & ¬Expected*. Except for ¬*Watched & ¬Expected* category, type *C* obtained the best score in *Watched & Good* category. This was an unexpected result.

Table 6 shows another summary of the result which is organized according to viewpoints of *prediction* and *recommendation*. The *prediction* viewpoint is the union of *Watched & Good* and *Watched & ¬Good* evaluations, and *recommendation* viewpoint is the union of *Watched & Good* and ¬*Watched & Expected* evaluations. For both of viewpoints, type *C* (*channel & time*) showed the best score. Whereas type *B* showed the lowest score in both of *prediction* and *recommendation* viewpoints. Althgouh type *A* at the *recommendation* viewpoint showed the second score, it did not exceed the type *C*.

5. DISCUSSION

One of our unexpected things is that the *keyword-based* and *celebrity-based* recommendations did not work in the second experiment. Although we expected that both of these recommendations would show good scores, the results was a contrary. Especially, type *celebrity-based* recommendation was the worst among four recommendations in Table 6.

For reasons of this failure, we consider that there are preferences for paid channels. Because paid channels in *TV-999* are highly specialized, we consider that each household already has its own preference for specific channels. For understanding this phenomenon, further investigation using the qualitative data such as the age of a respondent, number of people in his/her household, and channels and celebrities that s/he likes is needed.

6. CONCLUSION

We described evaluation results of experiments on TV program recommendation using EPG and viewer's log data. We had two experiments: (1) a TV program prediction experiment using the log data, and (2) a survey of TV program recommendations using the log data and a questionnaire. In the latter experiment, we found that the proposed methods did not work well, and the baseline method recommendation showed the best score. We will continue to investigate this phenomenon by analyzing the obtained questionnaire data.

Table 5: Results of the experiment (%).

Rec.	Watched&	Watched&	¬Watched&	¬Watched&
Type	Good	¬Good	Expected	¬Expected
A	9.9	0.5	17.5	72.1
B	3.1	0.4	12.1	84.4
A&B	6.3	0.4	16.7	76.6
C	13.9	1.1	16.3	68.7

Table 6: Performances with respect to prediction and recommendation.

	Prediction	*Recommendation*
Rec.	(*Watched&Good* ∪	(*Watched&Good* ∪
Type	*Watched&¬Good*)	¬*Watched&Expected*)
A	10.4 (%)	27.4 (%)
B	3.5 (%)	15.2 (%)
A&B	6.7 (%)	22.9 (%)
C	**15.0 (%)**	**30.2 (%)**

7. REFERENCES

[1] K. Ali and W. van Stam. Tivo: making show recommendations using a distributed collaborative filtering architecture. In *KDD '04: Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 394–401. ACM, 2004.

[2] L. Ardissono, C. Gena, P. Torasso, F. Bellifemine, A. Chiarotto, A. Difino, and B. Negro. Personalized recommendation of TV programs. In *AI*IA 2003: Advances in Artificial Intelligence, Lecture Notes in Computer Science*, volume 2829, pages 474–486, 2003.

[3] A. Bär, A. Berger, S. Egger, and R. Schatz. A lightweight mobile TV recommender. In *EUROITV '08: Proceedings of the 6th European conference on Changing Television Environments*, pages 143–147. Springer-Verlag, 2008.

[4] A. Basso, M. Milanesio, and G. Ruffo. Events discovery for personal video recorders. In *EuroITV '09: Proceedings of the seventh european conference on European interactive television conference*, pages 171–174. ACM, 2009.

[5] G. Graefe. Query evaluation techniques for large databases. *ACM Comput. Surv.*, 25(2):73–169, 1993.

[6] M. Gude, S. M. Grünvogel, and A. Pütz. Predicting future user behaviour in interactive live TV. In *EUROITV '08: Proceedings of the 6th European conference on Changing Television Environments*, pages 117–121. Springer-Verlag, 2008.

[7] R. J. Mooney and L. Roy. Content-based book recommending using learning for text categorization. In *DL '00: Proceedings of the fifth ACM conference on Digital libraries*, pages 195–204. ACM, 2000.

[8] B. Smyth and P. Cotter. A personalized television listings service. *Commun. ACM*, 43(8):107–111, 2000.

[9] W. E. Spangler, M. Gal-Or, and J. H. May. Using data mining to profile TV viewers. *Commun. ACM*, 46(12):66–72, 2003.