Understanding Latent Factors Using a GWAP

Johannes Kunkel, Benedikt Loepp, Jürgen Ziegler
University of Duisburg-Essen, Duisburg, Germany
{firstname.lastname}@uni-due.de

ABSTRACT

Recommender systems relying on latent factor models often appear as black boxes to their users. Semantic descriptions for the factors might help to mitigate this problem. Achieving this automatically is, however, a non-straightforward task due to the models' statistical nature. We present an output-agreement game that represents factors by means of sample items and motivates players to create such descriptions. A user study shows that the collected output actually reflects real-world characteristics of the factors.

KEYWORDS

Recommender Systems; Matrix Factorization; Game with a Purpose

ACM Reference Format:
Johannes Kunkel, Benedikt Loepp, Jürgen Ziegler. 2018. Understanding Latent Factors Using a GWAP. In *Proceedings of the Late-Breaking Results track part of the Twelfth ACM Conference on Recommender Systems (RecSys '18), Vancouver, BC, Canada, October 2–7, 2018*, 2 pages.

1 INTRODUCTION AND RELATED WORK

Recommender Systems (RS) make it often difficult for users to understand the results, in particular when using model-based techniques such as *Matrix Factorization* (MF) [3]: While latent factor models are known for accuracy and efficiency, they are typically considered non-transparent [9]. Some works indicate that learned factors are related to actual real-world characteristics [3, 6], but only few steps have been taken to automatically explain the abstract dimensions, e.g. by associating them with mined topics [9] or tags [6]. Others visualized relations between predefined tags and factors [8] or reduced the dimensionality to display a map [4]. Still, making factor meanings explicit can be considered difficult, especially without predefined data or complex visualizations. Thus, it seems promising to count on voluntary user contribution. *Games with a Purpose* (GWAP), with their prominent method of *Output-Agreement* (OA) [11], are well known for motivating users to solve such a human computation problem. In OA, randomly matched pairs of players are presented with a common input and have to come up with the same output without any means of communication [11]. The winning strategy is thus to type in terms that describe the shared content as best as possible. Such games are often used to annotate images [11], but also for e.g. eliciting preferences in RS [2].

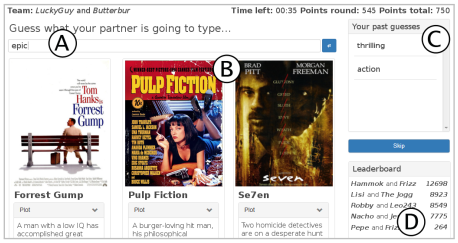
In this paper, we present a GWAP that follows the OA method for collecting semantic descriptions of latent factors.

2 THE GAME

In our game[1], representative items are shown as shared input for one factor after the other. For producing the same output, players have to type in commonalities of these representatives. As descriptions, we collect all terms typed in for each factor. Such term-factor relations may then enable to e.g. explain MF results by showing terms related to factors relevant for the current recommendation.

Factor Model and Representatives: First, we use a *Mahout ParallelSGDFactorizer* for offline model training. With *MovieLens 20M* dataset, 10 factors, $\lambda = 0.001$ and 16 iterations, results were up to standard ($RMSE = 0.86$, $NDCG@10 = 0.82$). While our approach is in principle independent of MF algorithm and parametrization, we deliberately choose a comparatively small number of factors. This is in line with earlier suggestions [7] and allows us to a) collect more terms per factor with less effort, and b) decrease likelihood of factors being redundant, thus diversifying game experience.

Next, for selecting sets of representatives both distinguishable and reflecting the semantics of model dimensions, we basically follow [7]: We calculate for each factor f and each movie i that has a value in the upper 25 % of values for f a score $s_{if} = 0.4 \cdot pop(i) + 0.3 \cdot rel(i, f) + 0.3 \cdot spec(i, f)$, taking *popularity* (high number of ratings), *relevance* (high value for f), and *specificity* (high value for f but neutral for others) into account. For each factor, we then select the 25 movies with highest s_{if} as representatives. Weights and numbers are the result of pretesting.



Team: *LuckyGuy* and *Butterbur* **Time left:** 00:35 **Points round:** 545 **Points total:** 750

Guess what your partner is going to type...

epic

Your past guesses
thrilling
action

Forrest Gump
Plot
A man with a low IQ has accomplished great

Pulp Fiction
Plot
A burger-loving hit man, his philosophical

Se7en
Plot
Two homicide detectives are on a desperate hunt

Leaderboard

Hammok and *Frizz*	12698
Lisi and *The Jogg*	8923
Robby and *Leo243*	8549
Nacho and *Je*	7775
Pepe and *Friz*	264

Figure 1: *LuckyGuy* guesses (A) how his partner would describe the factor representatives (B). He made two guesses already (C). A leaderboard shows overall performances (D).

Game Mechanics: We implemented the game as a web application (Fig. 1), as recommended for OA [11]. A *game* lasts 3 min., during which two randomly matched players seek to play as many *rounds* as possible with the goal of gaining *points* and scoring high in a *leaderboard*. Rounds are randomly related to factors of the underlying MF model. In each round, 3 movies are randomly chosen from the 25 factor representatives, and displayed by means of poster, plot description, cast and director. A round ends a) as soon as a *match* is found in the terms entered by both players, i.e. they *guess* the same, or b) if both decide to skip e.g. because the movies are too hard to describe (leading to penalty points). Either way, they then proceed to the next round, i.e. an item set for another factor is shown.

LBRS@RecSys '18, October 2-7, 2018, Vancouver, BC, Canada

[1] Introduced in [5] in German language, see also: http://interactivesystems.info/lfg/.

Table 1: Latent factors with sample representatives, number of guesses and matches, and terms that led to at least two matches.

Factor	1	2	3	4	5	6	7	8	9	10
Sample representatives	*The Lion King, Bad Boys, Home Alone*	*Forrest Gump, Fight Club, Se7en*	*Jurassic Park, Rocky V, Star Wars*	*Cast Away, Meet the Parents, Waterworld*	*The Doors, The Beach, Casino*	*Indiana Jones, Speed, Aliens*	*Nude Girls, American Pie, Harold & Kumar*	*The Net, Dave, Groundhog Day*	*Batman Forever, Deep Impact, Twister*	*Pretty Woman, Big, E.T.*
Guess./match. (ratio)	620 / 54 (11.48)	612 / 57 (10.74)	589 / 51 (11.55)	565 / 49 (11.53)	562 / 55 (10.22)	580 / 65 (8.92)	423 / 42 (10.07)	438 / 50 (8.76)	754 / 66 (11.42)	598 / 56 (10.68)
Matches (#)	comedy (10) funny (8) disney (4) action, love (3) fight, sex (2)	action (13) fight, man, serious (4) thrilling (3) comedy, dog, drama, thriller, war (2)	action (12) war (5) fight (4) comedy (3) drama (2)	action (13) comedy (8) drama, funny, spooky, thriller (2)	action (7) comedy, horror, love (5) spooky (3) erotic, mystery, old (2)	action (24) comedy (3) adventure, alien, fight, love, old, weapons (2)	comedy (10) sex (7) action, college, drama (2)	love (12) comedy (5) family (3) action, america, boring, romance, sex, woman (2)	action (18) horror (6) sci-fi, spooky (3) alien, aliens, batman, comedy, future (2)	love (11) comedy (6) family (4) action, animals, romance, romantic (3) dramatic (2)

3 EVALUATION

We conducted a user study to evaluate subjective game experience and collect a first baseline of factor descriptions. We recruited 84 (42 female) participants (age: $M = 21.23$, $SD = 4.53$), which were asked to play the game and fill in a questionnaire. The study took place at our lab, at participants’ home and in university classes. A supervisor was present and controlled that no communication occurred. Yet, few participants played again later without being supervised. Due to the different settings, players sometimes knew each other very well (21 %), while in many other cases, they did not (42 %). We used self-generated items to assess game-specific aspects, elicited demographics and domain knowledge, and measured user experience (SUS [1]) and enjoyment (IMI [10]). If not stated otherwise, we used 5-point Likert scales. We logged all interaction data, especially guesses and matches, to analyze the produced output.

Results: 173 games were played. Participants were required to play only once, but each player played on average 4.12 games. Together with the mean score of 4.79 ($SD = 1.26$) on the 7-point enjoyment subscale of IMI, this indicates that they liked playing. Participants reported that they were only sometimes in doubt when entering guesses ($M=2.75$, $SD=1.05$) and found the game overall rather easy to play ($M=3.21$, $SD=1.24$). They tended to love movies ($M = 3.31$, $SD = 1.03$) and had average knowledge about recent ones ($M = 2.60$, $SD = 0.98$). Accordingly, representatives were to some extent known ($M = 2.77$, $SD = 0.99$), but it was pointed out that a few old movies should have been omitted. Still, participants reported to have somehow understood why movies were displayed together ($M=2.74$, $SD=0.96$), and that they, considering all rounds, seemed diverse ($M=3.58$, $SD=1.01$). Provided information appeared sufficient ($M = 3.06$, $SD = 0.95$), with posters ($M = 4.32$, $SD = 0.91$) being most informative. Usability was *good* (SUS-score of 79).

In total, 5 741 guesses were made, on average 574.10 per factor ($SD=93.29$) and 33.18 per game ($SD=20.60$). This resulted in a total of 545 matches, on average 54.50 per factor ($SD=7.23$) and 3.15 per game ($SD=5.05$). Thus, each player had an *expected contribution* [11] of 68.35 guesses and 6.49 matches. For further analysis, we cleaned the dataset and set $X=2$ as *good label threshold* [11], i.e. minimum number of matches for a term to be considered meaningful. This left us with 325 matches comprising 35 distinct terms. Tab. 1 shows the collected data. Based on these terms, we created a dictionary and calculated content vectors by means of *TF-IDF*, representing how often terms led to a match for a factor in relation to how often this was overall the case. This allowed us to compare the sets of terms, i.e. descriptions created for factors, by means of vector cosine similarity. Overall, similarities were very low ($M=0.09$, $SD=0.13$).

Discussion: Questionnaire results and cosine similarities indicate high diversity between factors. Apparently, the method for selecting representatives ensures that shown items reflect different factor semantics. Accordingly, differences can be found in the sets of collected terms: Matches and their frequency seem consistent within factors, but vary between (see Tab. 1). Only in few cases, factors seem less unique, e.g. 8 and 10 ($sim=0.54$). This could be due to insufficient data, making items less distinguishable. On the other hand, factors might actually express similar aspects.

Some factors appear to have more obvious semantics: For instance, guess-match ratios in Tab. 1 show that players arrived at a match more often for factors 6 and 8, i.e. they are easier to describe.

Overall, current game mechanics seem to favor rather general terms, e.g. genres such as “action” or “comedy”. This effect has also been shown earlier [11], and can be prevented e.g. by taboo lists. However, implementing such mechanics would have required output data which we had not had prior to our study. On the other hand, very specific terms led to matches as well, e.g. “dog” for factor 2. This does not appear to result from participants describing a factor’s general meaning, but from certain movie posters being displayed. Yet, as the descriptions are only affected to a small degree, this issue will most likely vanish with more output data and $X > 2$.

4 CONCLUSIONS AND OUTLOOK

Study results show that our GWAP is fun to play and thus motivates users to produce output useful to better understand the hidden semantics of common RS models: Terms entered by players allow deriving meaningful and distinguishable latent factor descriptions, which may be used e.g. to explain recommendations by presenting keywords related to factors relevant for the active user and the respective items. Yet, investigating application areas is subject of future work, as is implementing a single player version as well as advanced game mechanics such as taboo lists for collecting further output data and more specific terms via gameplay.

REFERENCES

[1] J. Brooke. 1996. SUS – A quick and dirty usability scale. In *Usability Evaluation in Industry*. Taylor & Francis, 189–194.

[2] S. Hacker and L. von Ahn. 2009. Matchin: Eliciting user preferences with an online game. In *CHI ’09*. ACM, 1207–1216.

[3] Y. Koren and R. M. Bell. 2015. *Recommender Systems Handbook*. Springer US, Chapter Advances in collaborative filtering, 77–118.

[4] J. Kunkel, B. Loepp, and J. Ziegler. 2017. A 3D item space visualization for presenting and manipulating user preferences in collaborative filtering. In *IUI ’17*. ACM, 3–15.

[5] J. Kunkel, B. Loepp, and J. Ziegler. 2018. Ein Online-Spiel zur Benennung latenter Faktoren in Empfehlungssystemen. In *M&C ’18*. Gesellschaft für Informatik.

[6] B. Loepp, T. Donkers, T. Kleemann, and J. Ziegler. 2018. Interactive Recommending with Tag-Enhanced Matrix Factorization (TagMF). *IJHCS* (2018).

[7] B. Loepp, T. Hussein, and J. Ziegler. 2014. Choice-based preference elicitation for collaborative filtering recommender systems. In *CHI ’14*. ACM, 3085–3094.

[8] B. Németh, G. Takács, I. Pilászy, and D. Tikk. 2013. Visualization of movie features in collaborative filtering. In *SoMeT ’13*. 229–233.

[9] M. Rossetti, F. Stella, and M. Zanker. 2013. Towards explaining latent factors with topic models in collaborative recommender systems. In *DEXA ’13*. 162–167.

[10] R. M. Ryan. 1982. Control and information in the intrapersonal sphere: An extension of cognitive evaluation theory. *J. Pers. Soc. Psy.* 43, 3 (1982), 450–461.

[11] L. von Ahn and L. Dabbish. 2008. Designing games with a purpose. *Commun. ACM* 51, 8 (2008), 58–67.