IR Game: How well do you know information retrieval papers?

Jan Rybak
Norwegian University
of Science and Technology
jan.rybak@idi.ntnu.no

Krisztian Balog
University of Stavanger
krisztian.balog@uis.no

Kjetil Nørvåg
Norwegian University
of Science and Technology
kjetil.norvag@idi.ntnu.no

Abstract

In this paper we demonstrate how a gamification approach increases the attractiveness of an assessment exercise in the context of expertise profiling. We present an online game, in two difficulty modes, where users have to guess the authors of publications. We analyze the collected data along different dimensions and identify four types of gaming personalities based on behavioral patterns. Further, we examine the relation between popularity and recognizability for both papers and authors. Finally, we provide insights into game mechanics that extend beyond our specific use case.

1 Introduction

Gamification is a method for keeping users involved in a task for longer periods of time or to encourage them to repeatedly undergo otherwise not so entertaining tasks. Another reason for gamification is the desire to generate useful data as a by-product of the playing activity [VAD08]. Our work takes place in the context of (temporal) *expertise profiling*, where we are concerned with identifying what topics people are knowledgeable about [RBN14b]. Specifically, we focus on the academic domain, where scientific publications constitute the best available evidence from which to draw conclusions regarding a person's expertise. Of course, what one holds as her most important publication(s) does not necessarily correspond with what others (i.e., the scientific community) consider as such. Arguably, the latter one is more important. In our experience, getting the first question answered (what a person considers his most important publications) is not easy; we developed an assessment interface for this purpose [RBN14a] and found that people were not especially willing to spend time with it. (It has to be mentioned, however, that selecting the most important publications was only part of the task, the assessment procedure was more involved than that.) The main motivation for us, therefore, is to attract users that can represent the relevant scientific community's general opinion and can generate data to be used in our efforts to evaluate temporal expert profiling approaches [RBN14b].

Gamification is not the only alternative we considered. In many scenarios, crowdsourcing is a valid option for the delivery of simple yet time-consuming or repetitive tasks such as data annotation or evaluation. Eickhoff [Eic14] examines the crowd-powered expert paradigm, where the majority of the workload is preprocessed by crowd workers and experts are only needed for specialized steps. None of these approaches, however, are applicable in our case; here, the entire task is strictly domain-specific and requires the involvement of domain experts.

We cast our assessment exercise as a simple question-answering quiz that tests the user's knowledge of information retrieval (IR) papers, i.e., the "IR game." The player is presented with the title of a publication, from selected top conferences, and her task is to attribute the paper to the corresponding author(s).[1] The game comes in two difficulty modes. In "beginner" mode, the user has to select the right set of authors, from three options, while in "advanced" mode authors have to be picked out individually. The goal in each mode is to answer as many questions, i.e., collect as many points, as possible. The game ends after three wrong answers. A "leader board" is provided to track the highest scoring players. The game is available at `http://bit.ly/ir-game`.

In the course of this work, we examine whether we are able to attract more interest (and collect more data) by presenting our assessment exercise indirectly, as a game, as

In: F. Hopfgartner, G. Kazai, U. Kruschwitz, and M. Meder (eds.): Proceedings of the GamifIR'15 Workshop, Vienna, Austria, 29-March-2015, published at http://ceur-ws.org

[1] This game should not be unfamiliar to academics; many of us perform a similar "authorship attribution" exercise, albeit not deliberately, when performing blind reviews. The main difference is that here we offer instant feedback while providing limited context.

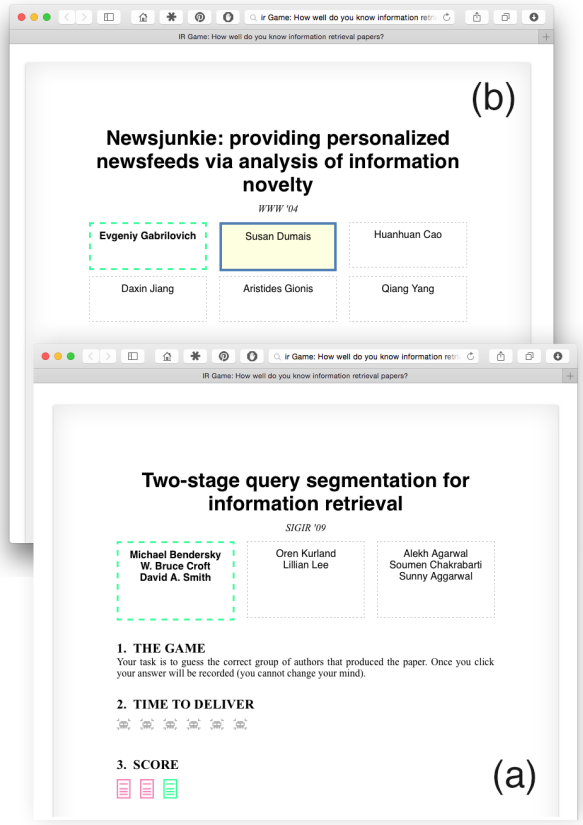


Figure 1: IR Game in (a) beginner and (b) advanced mode.

opposed to dealing with it explicitly using a purpose-built interface. In addition, we address a number of more specific questions:

- Which level of difficulty is preferred, the easy mode or the advanced one?
- Does a competitive element, such as a leader board, increase the level of engagement?
- When do users stop playing?
- Do users return to play again? After how long?
- What types of players can we identify?
- Are more cited papers also more easily recognized?
- Are more popular authors also more easily recognized?
- Do people prefer to play anonymously?

Our findings confirm the premise that interweaving game mechanics into a non-game environment is beneficial in terms of task attractiveness. We also demonstrate that the leader board is a powerful motivator for many people.

2 The Game

"IR game" is a simple knowledge quiz that tests users' knowledge of publications in the field of Information Retrieval.

2.1 Game rules

Given a publication title, the player's task is to select the correct authors within a given time limit. The game can be played in two modes, beginner and advanced.

Beginner mode The user has to select the (entire) group of authors from three options. Only one variant is correct and all authors are listed in the same order as on the paper. See Figure 1 (a).

Advanced mode In the more difficult game mode, individual author names are offered and the user has to decide which names belong to the paper. The number of authors does not necessarily correspond with the actual number of the paper's authors and the same applies to the order of names. See Figure 1 (b). The user is credited with the corresponding F1-score for each answer (i.e., correct vs. selected set of authors); answers below an F1-score of 0.5 count as wrong.

In both modes, the game ends after three wrong answers. A separate leader board is available for each game mode that lists the highest scoring players (with score, name, and timestamp). Players that made it to the leader board were offered the opportunity to "brag" about their achievement on Twitter.

2.2 Data

The collection of publications used in this game comprises of the 1111 top cited IR papers (according to the ACM DL[2]) from the period 2004-2014 that were presented in one of the following conference series: SIGIR, WWW, CIKM, KDD, and WSDM. For each publication, besides its own set of original authors, a set of "fictitious" authors is randomly selected from other documents within the same data set.

Usage data is collected while the game is being played. Specifically, for each user, we store the questions that have been asked in the game, correct and wrong answers, scores, date and time, time to answer, number of attempts to copy text from the webpage, and rough location. In order to recognize returning users, we use browser cookies with a unique identifier. We also track the site's traffic using Google Analytics.[3]

[2] http://dl.acm.org.
[3] http://www.google.com/analytics/.

2.3 Usage statistics

The game was promoted on Twitter[4] aiming at people from the IR community. In this paper, we analyze traffic from the first five days of the game's existence (i.e., from January 31 to February 4, 2015). During this period, 302 unique visitors from 33 countries visited the site and more than one third of them participated in the game. Figure 2 presents the geographic distribution of visitors; this roughly corresponds to the distribution of IR groups in the world (albeit Norway is admittedly over-represented on this figure). Figure 3 shows the time of the day when the game was played (normalized according to users' timezones).

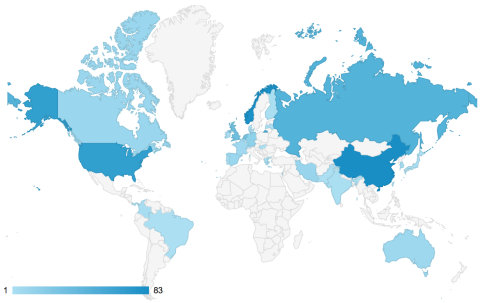


Figure 2: Geographic distribution of visitors.

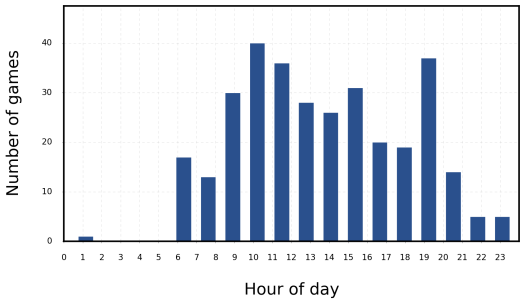


Figure 3: Time of day vs. number of games played.

Table 1 presents usage statistics, in terms of number of games played and number of unique players. We observe that the beginner game mode was almost 8 times more successful in terms of game counts and almost 9 times in terms of players counts than the advanced one. These statistics show that people who took part in the easier version of the game were more likely to play again.

Table 1: Usage statistics

	Game mode		Total
	Beginner	Advanced	
#unique players	111	16	116
#games played	347	39	387
avg. #games per player	3.14	2.44	3.34

[4] `#irgame`

3 Analysis of results

Next, we analyze the collected data in different ways: by answers (§3.1), by players (§3.2), by papers (§3.3), and by authors (§3.4).

3.1 Answers

Time to answer

In both game modes, users' response time is limited to 15 seconds. In case the time limit is exceeded, the answer is considered wrong. On average, it took about half of the specified time limit to provide an answer, more precisely, it was 7.58s. There is a notable difference, 1.8s, between average response times for correct (6.74s) and wrong (8.53s) answers. Looking at the distribution of answer times, Figure 4, we find that the higher response time for wrong answers is due to timeouts. It is also visible from this plot that when the user knows the correct answer, he is less likely to use up all the time available.

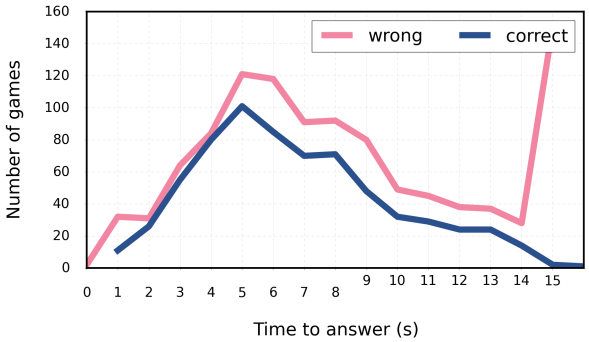


Figure 4: Time to answer in beginner mode.

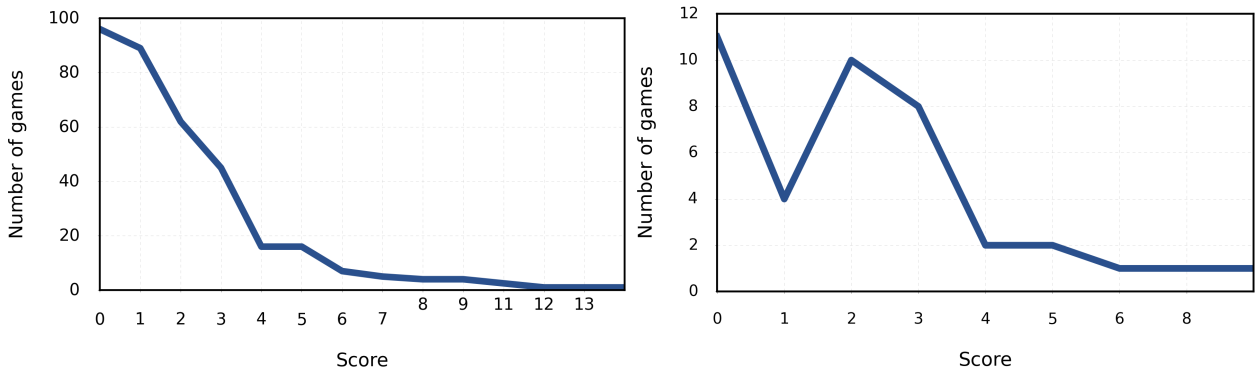
Total scores

Figures 5(a) and 5(b) depict the distribution of total game scores for beginner and advanced modes, respectively. These correspond to power-law shape distributions, although the existing data (esp. for the advanced mode) are too sparse to infer their parameters. In both cases there is a noticeable drop when going from score 3 to 4. This has to do with the fact that the game ends after 3 mistakes.

3.2 Players

Returning visitors

An interesting measure of success of a game is the number of returning players. We examined all games that were played more than once (56 games) searching for interesting patterns. It is most likely that a user plays again right after she finishes the game. In 42 cases, users played again within the same hour. The plot on Figure 6 presents time intervals between users' returns.



(a) beginner mode (b) advanced mode

Figure 5: Distribution of game scores.

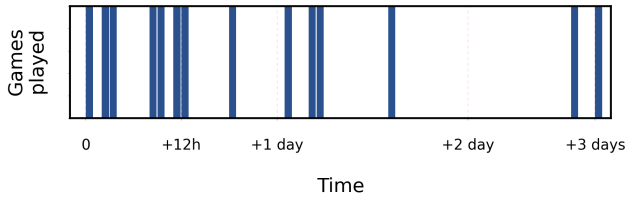


Figure 6: Time elapsed between games (from the first game) for returning visitors.

Player types

From the analysis of the game series, we can derive 4 types of players depending on when they decide to leave the game.

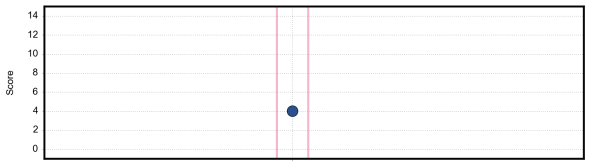
Jumpers are the type of visitors who come and play a single game; they leave after that no matter what the score is.

Give-upers are players who return repetitively but leave the game due to demotivation when in a series of games their score drops.

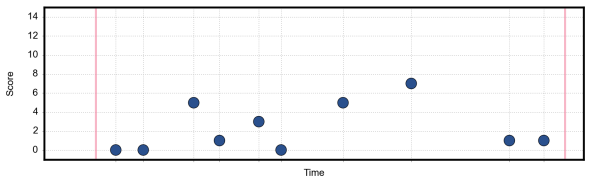
Fighters do the exact opposite. They leave the game at the top of their form, when in a series of games they reach their highest score.

Achievers care about winning. They keep returning and playing the game until they are back on the top of the leader board.

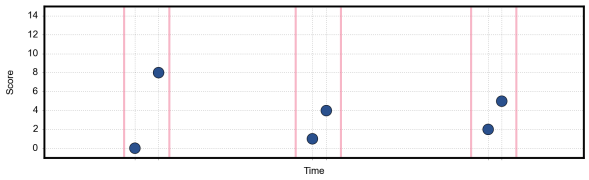
Figure 7 shows an example user for each of the player types.



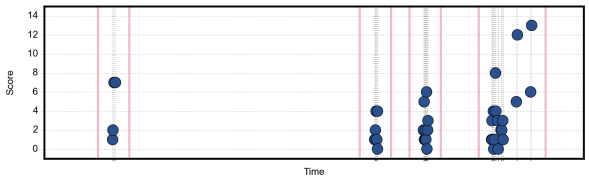
(a) Jumper



(b) Give-uper



(c) Fighter



(d) Achiever

Figure 7: Examples of player types. Session boundaries are marked with vertical red lines.

3.3 Papers

Are popular papers, i.e., papers with more citations, recognized more easily? In order to answer this question, we first introduce the concept of a paper's *recognition ratio*. It is defined to be the number of times the publication was successfully recognized by users (players) divided by total number of times it was shown to users. Next, we divide all publications that appeared in the game into three groups based on the number of citations they received (according to ACM DL). We report the average recognition ratio for

each group in Figure 8, where the leftmost bar represents papers with the highest number of citations. As expected, we find that more cited papers are in general better recognized (left and middle vs. right), albeit papers in the middle of the citation range seem to perform best in this regard. We note that these findings may not be conclusive due to data sparsity.

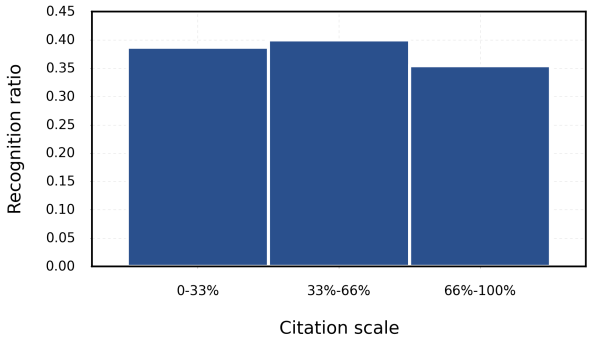


Figure 8: Citation counts vs. recognition ratio.

3.4 Authors

Are popular authors, i.e., people with more publications, recognized more easily? Similarly to papers, we define an author's *recognition ratio* to be the number of times the author's publications were successfully recognized by users (players) divided by total number of times her publications were shown to users. On Figure 9 we plot authors' popularity, measured in the number of publications (in our paper selection), against recognition ratio. We find that there is a significant difference between authors with a single publication and authors with multiple publications; not surprisingly, having multiple publications benefits recognition. On the other hand, it appears that having many more publications does not improve recognition any further.

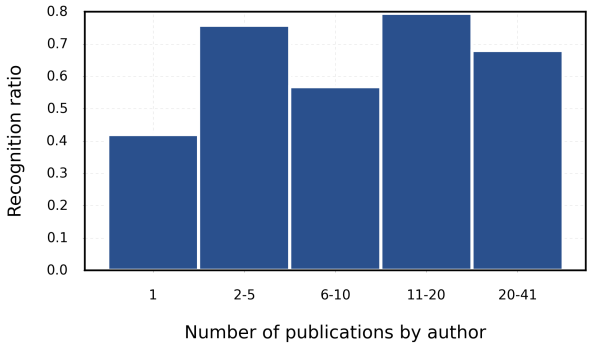


Figure 9: Author popularity (number of publications) vs. recognition ratio.

4 Observations

Based on the analysis of results as well as informal feedback from the users, we make a number of observations that may generalize beyond our specific use case.

Learning From user feedback we know that this game was also used to learn about relevant, previously unseen, publications. On the motto of Comenius' saying: *"Much can be learned in play that will afterwards be of use when the circumstances demand it."*, we believe that our game might also prove to be useful for exploring and discovering more about scientific literature.

Unfair behavior We have a suspicion that at least in one case a user acted dishonestly in order to get into the lead. This user's score was unreasonably high compared to the second best. More importantly, her average time to answer was 12.84s (compared to the average of 7.85s), which seems just about enough time to use a web search engine to look up the paper in question. This behavior was reported by another competitor (which supports the assumption that players do care about their position in the leader board).

Head-start At least in one case we observed that a user was restarting the game until he was able to answer the first question correctly. Beginning the game with a set of easier questions, then gradually increasing difficulty, might therefore be helpful in keeping users engaged.

Engaging users From the statistics (§3.2) we can see that users are not very likely to return to the game days after their first visit. However, chances that they repeatedly participate in the game within the first hour after their initial visit are much higher. It is therefore of vital importance to keep the user stay in the game as long as possible when she comes for the first time.

Identity Some people ($\sim 10\%$) opted to use their full civil name as opposed to a nickname. We hypothesize that it was a choice made deliberately in case they make it to the leader board.

5 Conclusions and Future work

This study presented in this paper has started with the following main question in mind: Could we make an assessment exercise, in the context of expertise profiling, more appealing for users? We have answered this question positively. Our experiment has shown that it is more desirable for users to participate in a game-like assessment task rather than having to evaluate results explicitly using a purpose-built interface. We have analyzed the data collected along different dimensions and have identified four

types of gaming personalities based on behavioral patterns. On top of the analysis of the game mechanics, this experiment has allowed us to gather valuable data about authors and publications. This has let us to perform an initial examination of the relation between popularity and recognizability for both papers and authors.

In future work we plan to enhance the game in several ways. The main purpose of the game is to indirectly measure how researchers recognize each others' publications. In this first version, fictitious publications were selected randomly; however, interesting experiments could be conducted if the selection of alternative authors was biased in a controlled way. This would allow us to adjust the difficulty of the questions as the game progresses. We also plan to add new game modes (e.g., time trial), expand the data set (i.e., add more publications), and possibly explore other research fields/communities.

References

[Eic14] Carsten Eickhoff. Crowd-powered experts: Helping surgeons interpret breast cancer images. In *Proceedings of the First International Workshop on Gamification for Information Retrieval*, pages 53–56, 2014.

[RBN14a] Jan Rybak, Krisztian Balog, and Kjetil Nørvåg. ExperTime: Tracking expertise over time. In *Proceedings of the 37th International ACM SIGIR Conference on Research & Development in Information Retrieval*, SIGIR '14, pages 1273–1274, 2014.

[RBN14b] Jan Rybak, Krisztian Balog, and Kjetil Nørvåg. Temporal expertise profiling. In *Proceedings of the 36th European conference on Advances in Information Retrieval*, ECIR '14, pages 540–546, 2014.

[VAD08] Luis Von Ahn and Laura Dabbish. Designing games with a purpose. *Communications of the ACM*, 51(8):58–67, 2008.