On the effectiveness of a Mobile Puzzle Game UI to Crowdsource Linked Data Management tasks

Irene Celino
CEFRIEL - ICT Institute
Politecnico di Milano
Milano, Italy
irene.celino@cefriel.it

Emanuele Della Valle
DEIB
Politecnico di Milano
Milano, Italy
emanuele.dellavalle@polimi.it

Roberto Gualandris
DEIB
Politecnico di Milano
Milano, Italy
roberto.gualandris@mail.polimi.it

ABSTRACT

Linked Data publishing on the Web is a stably growing phenomenon, but its effective usage depends on the ability of consumers to assess the trustworthiness and the relevance of the published data. Pure automatic techniques are often inadequate to this end. Crowdsourcing is often advocated as a valuable solution. In this paper, we propose WikiFinder – a Games With A Purpose inspired by popular mobile puzzle games – and we report on its effectiveness in solving typical Linked Data Management tasks.

1. INTRODUCTION

The amount of structured data published on the Web is growing, but our ability to assess if a data item is true or false, or to understand if it is popular, remains limited. For instance, consider the following two statements: "The airport of Milano Malpensa is in Milan" and "Leornardo da Vinci was born in Vinci". Can a computer program tell if they are true or false? Assuming they are both true, which is the most popular one? These questions illustrate some of the Linked Data Management tasks [1]. These tasks are often difficult to automate. We know humans can solve them manually, but that they are normally not willing to. The central research question of crowdsourcing [2] is, indeed: which are the right incentives to motivate large groups of people to contribute their time in solving problems that machines are not good enough at?

In this paper, we focus on Games With A Purpose (GWAPs) [5]. In this crowdsourcing approach, designers embed in a game a task to be solved by the players (i.e., the purpose). In terms of incentives, GWAPs try to use game fun to engage people in solving the crowdsourced task. In particular, we study how effective is the user interface (UI) of a puzzle game in engaging casual mobile players in performing Linked Data Management tasks [4].

The remainder of the paper is organised as follows. Section 2 presents the UI of WikiFinder[1] – a GWAP inspired by popular puzzle mobile games such as Ruzzle. Section 3 reports on the effectiveness of WikiFinder in solving Linked Data Management tasks. Section 4 discusses the results while in Section 5 we draw some conclusions and we cast some light on future works.

2. THE WIKIFINDER GWAP

WikiFinder (see Figure 1) proposes its players two columns of people/teams/movies/actors/etc. (i.e., the blue column on the left and the red one the right) and several ways to link them (i.e., the grey column in the middle). The player has to find as many links as she can in 1 minute by sliding her fingers from any blue cell on the left to one of the three adjacent ones in the central grey column landing in one of the red cells of the right column. If the player finds a right link, she receives a positive feedback and her score is increased, otherwise she receives a negative feedback and her score is decreased. The decrement of the score is meant to deter random guessing. The increment is larger in the first seconds of the game and decreases over time so to reward faster players. When the time is over the player learns how many links were hidden in the puzzle and how many she found. A leaderboard challenges players to compete one against the others for the highest positions. WikiFinder, so far, has been played by 44 players for a total of 440 games (7.3 hours). The average life play of a user is 10 minutes.

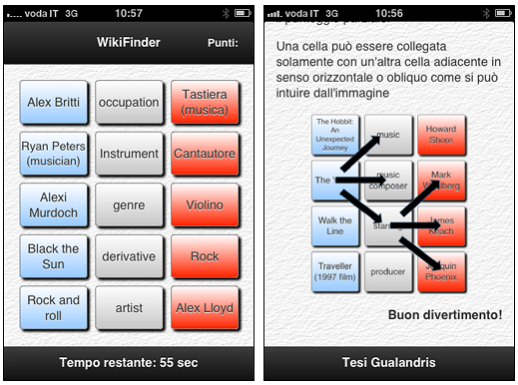

Figure 1: Screenshot of the WikiFinder game UI (left) and explanation of the gameplay in WikiFinder help (right)

The factual knowledge used to create the puzzles comes from DBpedia. An automatic algorithm builds the board by issuing SPARQL queries against DBpedia public endpoints.

[1]For Android check out `http://bit.ly/WikiFinder`, while for iOS check out `http://wikifinder.altervista.org/wikifinder/`

The algorithm works in two stages. The first one determines the paths in the puzzle that will contain the correct statements, while the second one looks in DBpedia for statements that can fit the puzzle. The paths can be direct (i.e., they link two and only two resources), or can contain branches (i.e., they can share a left blue, a middle grey or a right red cell). The algorithm uses different types of queries for filling in the different types of paths. For instance, if the path is straight it asks for a link between a seeding resource and another resource. For instance, if the seeding resource is the album "Black the Sun", the algorithm finds "Alex Lloyd" is the artist that produced it. If two paths share one left blue cell, one grey middle cell and two red right cell, a query using as seeding resource the artist "Alex Britti" may find that he plays two instruments: keyboards and violin. Once all the paths are filled up the remaining cells are filled up with queries that cannot introduce correct statements. The algorithm is non-deterministic, i.e., it may not be able to complete the puzzle. If that is the case, the puzzle is discarded and a new one is created.

GWAPs similar to WikiFinder were proposed by J. Waitelonis et al. in WhoKnows [6] and J. Hees et al. in BetterRelations [3]. WhoKnows is an online quiz that uses DBpedia to generate questions such "Spanish language is the language of ...?" proposing as answers Chile, Iraq, Brasil and Italy. As in WikiFinder, besides its entertainment side, WhoKnows hides the purpose of ranking statements to detect those that are inconsistent or doubtful. Also BetterRelations uses facts from DBpedia to build a game. The games casually pairs two players, who do not know each others, and presents two statement about the same subject (e.g., "Facebook is an online social network" and "Facebook's key people include Chris Hughesasking the players: "What would come first to your partner's mind?". The players scores if they agree. The purpose also in this case is ranking statements.

3. ANALYSIS OF WIKIFINDER RESULTS

Starting from the data collected through the WikiFinder game, we performed some evaluations aimed at understanding its effectiveness in processing linked data. In this section, we illustrate our analysis and results, grouped by the focus of the evaluation: the whole triple or one of its components (subject, predicate, object).

3.1 Results at Triple level

Regarding entire triples, we can evaluate WikiFinder's capability (1) to tell true and false facts apart and (2) to rank triples on the basis of their "popularity", i.e. the quality of being well known to an average (Italian) person.

3.1.1 Distinguishing between true and false triples

In total, during the gameplay, WikiFinder's players selected at least once a total of 1127 different triples, out of which 246 true facts and 881 false facts (in this context "false" means that the triple does not exist in DBpedia).

METHODOLOGY: we compute the relative *popularity* of a fact as the frequency of its selection by players (# of selections / # of times the triple was displayed in a grid to players), and the players' *precision* in selecting true triples (% of true triples among all selected triples); we then plot the true/false discernment precision as a function of the fact popularity in Figure 2.

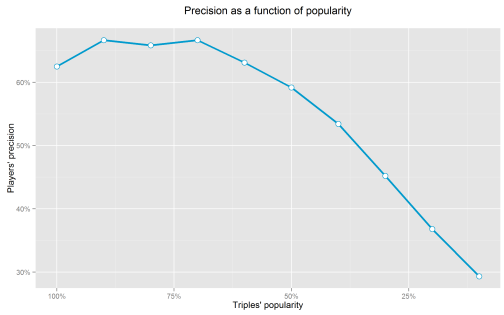

Figure 2: Distinguishing true/false triples

RESULTS: as illustrated in the graphics, WikiFinder's players were very precise in selecting true triples among the popular ones (those with popularity greater than 60%), while their ability decreases with the facts' general recognition.

3.1.2 Triples "popularity" – cf. with Web search

Focusing only on the 246 true facts (i.e. triples existing on DBpedia and selected at least once during gameplay), we can compare WikiFinder's ability to determine facts' popularity with some terms of comparison.

METHODOLOGY: we compute the *WikiFinder popularity* of a fact as the frequency of its selection by players (# of selections / # of times the triple was displayed in a grid to players), and the *Web search popularity* of a fact as the number of results for a Web search for the string "`<subject-label> <predicate-label> <object-label>`" by using Bing search API; we then compute the Spearman's rank correlation coefficient between the two popularity vectors and plot this correlation in Figure 3.

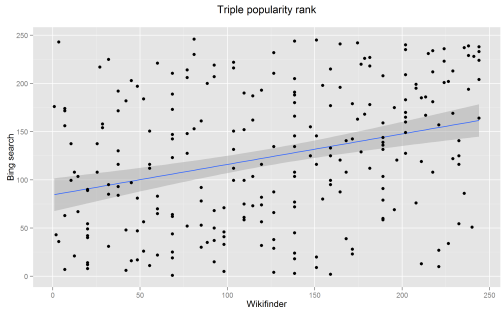

Figure 3: Triple popularity – cf. with Web search

RESULTS: there is a statistically significant Spearman's coefficient of 0.32 (p-value 4.05e-07) that supports the correlation hypothesis; also, a Web search for the string "`<subject-label> <object-label>`" (without the predicate) gives similar results (Spearman's coefficient: 0.33, p-value: 1.02e-07).

3.1.3 Triples "popularity" – cf. with NGD

Another term of comparison for WikiFinder's popularity is the Normalized Google Distance (NGD) between facts' subject and object. NGD is an indicator of the relationship strength between two strings *s1* and *s2*, computed as a function of the number of results for "*s1*", "*s2*" and "*s1 s2*".

METHODOLOGY: we compute the *WikiFinder popularity* as above and the NGD as the distance between "`<subject-label>`" and "`<object-label>`" as computed via Bing search API; we then compute their Spearman's rank correlation and plot it in Figure 4.

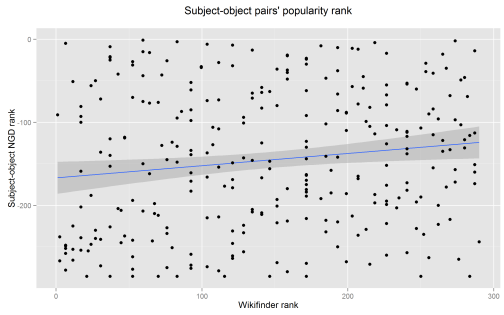

Figure 4: Triple popularity – cf. with NGD

RESULTS: while there is a statistically significant Spearman's coefficient of 0.15 (p-value: 0.01205), the correlation is too weak to conclude that WikiFinder is able to correctly predict subject-object NGD.

3.1.4 *Triples "popularity" – cf. with manual rank*

We also compared WikiFinder's popularity with another crowdsourced ranking mechanism, so to compare an indirect form of ranking (through the gameplay) to an explicit ordering activity.

METHODOLOGY: we selected a random sample of 50 true facts and we asked to 10 users (disjoint from the game players' group but with similar characteristics) to judge their popularity on a 5-level Likert scale (very high, high, medium, low, very low); then we computed the Spearman's rank correlation between the *WikiFinder popularity* of the sample and the *manual ranking* of those facts, obtained by aggregating the users' judgments.

RESULTS: from this analysis we obtain a statistically significant Spearman's coefficient of 0.39 (p-value 0.006) that supports the correlation hypothesis; also the distribution of judgments across the different popularity levels is similar between WikiFinder and the manual ranking (cf. Figure 5).

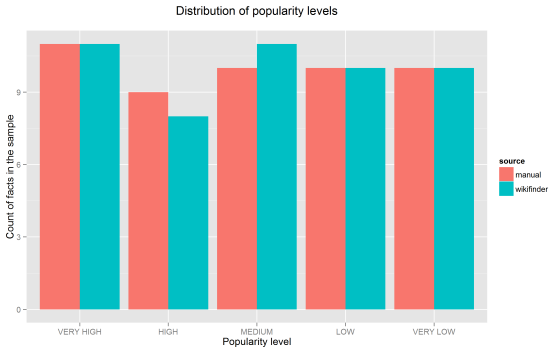

Figure 5: Distribution of triple popularity

3.2 Results at Subject/Object level

Similarly to triple level, we can evaluate the popularity of subjects and objects of each fact, by comparing with external terms of comparison: Wikipedia and Web search results.

3.2.1 *Cf. with Wikipedia*

METHODOLOGY: since all facts come from DBpedia, each subject and object has a respective page on Wikipedia. We obtained the number of visits of those subject/object pages in the same period of gameplay, i.e. 2013, for both the English and the Italian version of Wikipedia[2]. We then compared that *Wikipedia popularity* measure with the *WikiFinder popularity* (# of selections / # of times a triple was displayed in a grid to players) and computed the Spearman's rank correlation. For example, Figure 6 plots the correlation for subjects, using Italian Wikipedia visits.

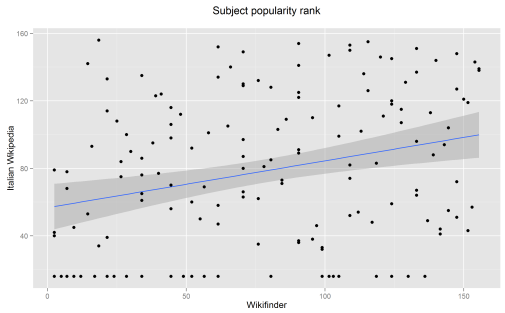

Figure 6: Subject popularity – cf. with Italian Wikipedia

RESULTS: for both subjects and objects, we obtain a statistically significant Spearman's coefficient:

- Subjects/English: correlation 0.26 (p-value: 0.0009)
- Subjects/Italian: correlation 0.28 (p-value: 0.0004)
- Objects/English: correlation 0.24 (p-value: 0.005)
- Objects/Italian: correlation 0.27 (p-value: 0.001)

This correlation value is slightly higher for Italian Wikipedia than for the English version, accordingly to the nationality of WikiFinder players.

3.2.2 *Cf. with Web search*

METHODOLOGY: we compute the *WikiFinder popularity* of a subject/object as above and the *Web search popularity* as the number of results for a Web search for the string `<subject-label>` or `<object-label>` by using Bing search API; we then compute the Spearman's rank correlation coefficient. For example, Figure 7 plots the correlation for subjects with Web Search.

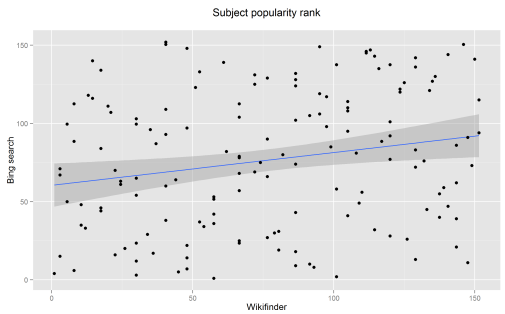

Figure 7: Subject popularity – cf. with Web search

RESULTS: we obtain a statistically significant correlation for both subjects (0.21, p-value 0.0096) and objects (0.29, p-value 0.0005).

3.3 Results at Predicate level

At predicate level, we can evaluate the capability of WikiFinder players to correctly identify the domain and range

[2] Cf. `http://stats.grok.se/`.

of properties. Every time they select a subject-predicate-object triple in the gameplay – whether the fact is true or false – we can assume that they are expressing the opinion that the subject belongs to the predicate's domain and the object belongs to the predicate's range.

METHODOLOGY: among all predicates appearing in WikiFinder facts, we selected the subset of 29 properties with an explicitly defined domain and range in DBpedia. We then computed the "*accuracy*" of WikiFinder players in correctly identifying domain (or range) as the # of selected triples with the subject belonging to the predicate's domain (or the object belonging to the predicate's range) divided by the # of selected triples with that specific predicate.

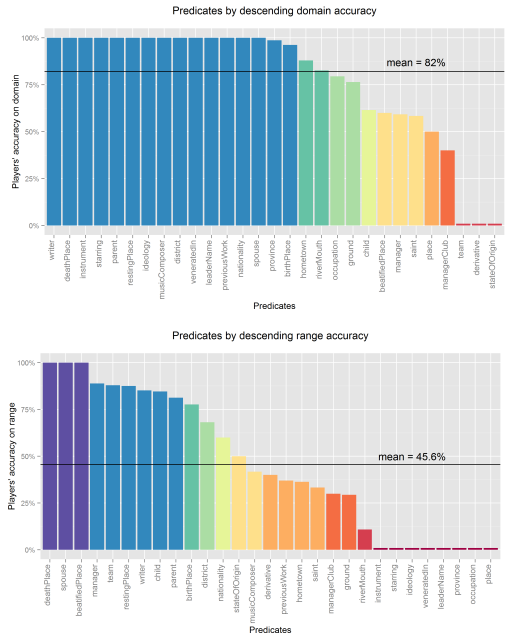

Figure 8: Identifying predicates' domain and range

RESULTS: as evident in Figure 8, WikiFinder players are very good at identifying predicates' domain (average accuracy of 82%, nearly half of predicates with 100% accuracy), while they show a much varied behaviour in identifying predicates' range (average accuracy of 45.6%, but with more than one third with accuracy greater than 75% and less than one third with 0% accuracy).

4. DISCUSSION OF RESULTS

In general, all our tests exhibited a positive correlation between WikiFinder outcomes (at triple, subject and object levels) and data that can be derived from other alternative sources; however, the correlation strength is never greater than 0.40 and explains only part of the variance. Our interpretation is that this weak correlation could be due to the limited number of players and also to the high diversity of facts in terms of topic (which was selected by design to challenge players with very disparate knowledge items).

Some results support very well the effectiveness of WikiFinder to execute Linked Data management tasks, like in the case of distinguishing true and false facts (cf. Section 3.1.1) and in the case of identifying predicates' domain (cf. Section 3.3). Some other evaluations, instead, highlight either room for improvement or the weakness to address specific tasks. This is the case, for example, of the identification of properties' range: WikiFinder players clearly show a much worse behaviour than in the case of properties' domain. Since it is not reasonable to expect different results for domain and range, this effect can be due to some specificity of WikiFinder; our interpretation is that, since the gameplay requires to "draw" the triple only from subject through predicate to object (i.e. from left to right of the mobile screen), probably the players start connecting the subject to the predicate (domain relation) even before looking at the possible objects (range relation) and, acting quickly to avoid running out of time, they tend to "close" the triple even when a suitable object is not available. To support this interpretation, WikiFinder gameplay could be modified to allow the players to "draw" the triple also in the opposite direction (from right to left).

5. CONCLUSIONS AND FUTURE WORK

A GWAP like WikiFinder can be exploited to get a full range of information required to address Linked Data management tasks from a mobile game UI. While not decisive, the results of WikiFinder evaluation are encouraging, because they show the game potential to crowdsource relevant tasks that could be otherwise expensive to perform.

Besides gathering additional information by involving a higher number of players, our future work will focus on improving the game to address the shortcomings highlighted in this paper; for example, we will build game grids with more homogeneous topics so to evaluate the effect of knowledge diversity on the tasks results.

Changes will also be made to the gameplay, in order to specifically assess the influence of different UI design choices on WikiFinder results; for example, we intend to check the influence of the fact selection "direction" (from right to left rather than from left to right) on domain and range identification. Another improvement could lie in the facts layout within the grid: it would be interesting to understand if WikiFinder experiences a top-to-bottom effect, i.e. if players tend to select facts displayed in the upper part of the screen more often than those appearing in the bottom area.

6. REFERENCES

[1] C. Bizer, T. Heath, and T. Berners-Lee. Linked Data - The Story So Far. *Int. J. Semantic Web Inf. Syst.*, 5(3):1–22, 2009.

[2] A. Doan, R. Ramakrishnan, and A. Halevy. Crowdsourcing systems on the World-Wide Web. *Commun. ACM*, 54(4):86–96, 2011.

[3] J. Hees, T. Roth-Berghofer, R. Biedert, B. Adrian, and A. Dengel. Betterrelations: using a game to rate linked data triples. In *KI 2011: Advances in Artificial Intelligence*, pages 134–138. Springer, 2011.

[4] E. Simperl, B. Norton, and D. Vrandecic. Crowdsourcing Tasks within Linked Data Management. In *Proceedings of COLD2011*, volume 782. CEUR-WS.org, 2011.

[5] L. von Ahn. Games with a Purpose. *IEEE Computer*, 39(6):92–94, 2006.

[6] J. Waitelonis, N. Ludwig, M. Knuth, and H. Sack. Whoknows? evaluating linked data heuristics with a quiz that cleans up dbpedia. *Interactive Technology and Smart Education*, 8(4):236–248, 2011.