Collecting Links between Entities Ranked by Human Association Strengths

Jörn Hees[1,2], Mohamed Khamis[3], Ralf Biedert[2,4],
Slim Abdennadher[3], and Andreas Dengel[1,2]

[1] Computer Science Department, University of Kaiserslautern, Germany
[2] Knowledge Management Department, DFKI GmbH, Kaiserslautern, Germany
{joern.hees,andreas.dengel}@dfki.de
[3] Computer Science & Engineering Department, German University in Cairo, Egypt
{mohammed.khamis,slim.abdennadher}@guc.edu.eg
[4] Tobii Technology AB, Stockholm, Sweden
rb@xr.io

Abstract. In recent years, the ongoing adoption of Semantic Web technologies has lead to a large amount of Linked Data that has been generated. While in the early days of the Semantic Web we were fighting data scarcity, nowadays we suffer from an overflow of information. In many situations we want to restrict the amount of facts which is shown to an end-user or passed on to another system to just the most important ones.

In this paper we propose to rank facts in accordance to human association strengths between concepts. In order to collect a ground truth we developed a Family Feud like web-game called "Knowledge Test Game". Given a Linked Data entity it collects other associated Linked Data entities from its players. We explain the game's concept, its suggestion box which maps the players' text input back to Linked Data entities and include a detailed evaluation of the game showing promising results. The collected data is published and can be used to evaluate algorithms which rank facts.

1 Introduction

Since its introduction in 2001 the Semantic Web [1] has gained much attention. While in the early days of the Semantic Web only few large, interlinked and publicly accessible RDF datasets were available, especially the Linking Open Data (LOD) project has changed this situation over the last years, generating one of the world's largest, decentralized knowledge bases [2]. Extracted from Wikipedia, DBpedia [3] is the most central of these datasets as it provides information about entities from a large variety of domains, provides URIs for these entities and thereby provides a bridge between many other domain specific datasets in Linked Data.[1]

[1] Also see http://lod-cloud.net/ the Linking Open Data cloud diagram by Richard Cyganiak and Anja Jentzsch.

P. Cimiano et al. (Eds.): ESWC 2013, LNCS 7882, pp. 517–531, 2013.

Despite being a huge success for the Semantic Web, the increasing amount of available Linked Data creates new problems. While in the beginning there was not nearly enough data available to answer simple real-world queries, nowadays it often is easier to answer very specific queries. Simple queries lack specificity and it is not rare that they return thousands of facts. Widely known examples of such queries are SPARQL's DESCRIBE queries. For a given concept `:c` of interest on many SPARQL endpoints a DESCRIBE just returns the union of all outgoing `{ :c ?p ?o. }` and incoming `{ ?s ?p :c. }` triples. The same holds true for the majority of resolvable URIs. Sometimes, the often alphabetically sorted results are even truncated without any sanity to reduce bandwidth consumption.

While this behavior is acceptable for debugging, it most certainly is not what should be happening in productive systems which try to use the gathered information and in the end present the results to users. When simply asked about a URI, servers should return useful information opposed to all information they know, as mentioned in the Linked Data Design Issues by Berners-Lee [4].

The problem with this rule is that it is unclear which information is useful for a client. It depends on the context of the client. Nevertheless, we can observe that clients who are in a specific context typically have a specific information need and are able to formulate more specific SPARQL queries than DESCRIBE or resolving URIs. Hence, in this paper we focus on a general purpose information need, as often encountered in search engines.

As human associations play a key role in human thinking, leading us from one thought to the next, we propose to rank Linked Data facts according to human association strengths between entities. This means that for an entity such as `dbpedia:Steve_Jobs` which is strongly associated to `dbpedia:Apple_Inc.` we will rank facts between these two entities higher than facts connecting `dbpedia:Steve_Jobs` and `dbpedia:Toy_Story` entities.

Note that associations should be distinguished from semantic similarity. Two entities can be associated (see above), semantically similar (`dbpedia:Steve_Jobs`, `dbpedia:Brin_Sergey`), or both (`dbpedia:IPhone`, `dbpedia:IPad`).

To the best of our knowledge, currently no heuristic for or dataset of human association strengths between Linked Data entities is available. Furthermore, collecting such a dataset is prone to subjectivity, it is extremely monotonous and tedious, and the immense amount of Linked Data would cause great expenses if it was collected with a traditional experiment with paid participants.

In this paper we present a web-game called "Knowledge Test Game" to overcome the aforementioned problems, following the "Games With A Purpose (GWAP)" approach by von Ahn and Dabbish [5]. For a given Linked Data entity the game collects other associated Linked Data entities by outsourcing the problem to its players. The game is not intended to collect and rank associations for all Linked Data entities. Rather it is intended to build a ground truth that can be used to benchmark existing or new ranking techniques for Linked Data. As a next step, well performing ranking techniques could then be used to streamline the acquisition of associations between Linked Data entities, possibly allowing for a more human like exchange of knowledge between machines in the future.

The remainder of this paper is structured as follows. In Section 2 we list related GWAPs. In Section 3 we explain the game's concept, its suggestion box which maps the players' text input back to Linked Data entities, before presenting a detailed evaluation of the game showing promising results in Section 4. The results of this evaluation are discussed in Section 5 before our conclusion and future work in Section 6.

2 Related Work

While many approaches to rank Linked Data exist [6], we are not aware of any approach to collect or approximate human association strengths between Linked Data entities which also distinguishes them from semantic similarities. Hence, we will mainly focus on GWAPs which are related to our "Knowledge Test Game" in this section.

In terms of game design, the Knowledge Test Game is an output-agreement game [5] and a game with a purpose for the Semantic Web [7]. Its gaming principles are influenced by *Common Consensus*, another Family Feud like web-game which asks its players to name common sense goals (e.g., "What can you do to watch TV?"). In contrast to Common Consensus our approach focuses on all associations and does not only collect textual player inputs, but also maps the entered answers back to existing Linked Data entities with its suggestion-box.

The Knowledge Test Game can be seen as a successor of *Associator* [8] which was a pair-game to collect free-text associations for given topics. Associator as Common Consensus did not attempt to match the entered strings back to Linked Data entities during play time.

Other GWAPs to rate Linked Data exist. *BetterRelations* [9], a pair game asks its player which of two facts they consider more important. Aside from not collecting free associations between entities, BetterRelatons suffers from noise issues that our approach overcomes by using its suggestions-box.

WhoKnows? [10], a single player game, judges whether an existing Linked Data triple is known by testing players with (amongst others) a multiple choice test or a hangman game. In contrast to our approach, WhoKnows restricts itself to a limited fraction of the DBpedia dataset and excludes triples not matched by a predefined domain ontology in a preprocessing step. Similarly, *RISQ!* [11], a Jeopardy like single player game that generates questions from DBpedia, restricts itself to the domain of people after excluding non-sense facts in a preprocessing step. It then rates the remaining facts by using predefined templates to generate questions (clues) about subjects and tests if they are correctly recognized from a list of alternatives. This greatly reduces noise issues, but eliminates the possibility to collect user feedback about triple qualities and problems in the extraction process. Furthermore, unlike in the three aforementioned games, players of the Knowledge Test Game are not limited in their choices to previously existing connections of Linked Data entities, but instead can freely associate between them and even introduce new entities, should they be missing.

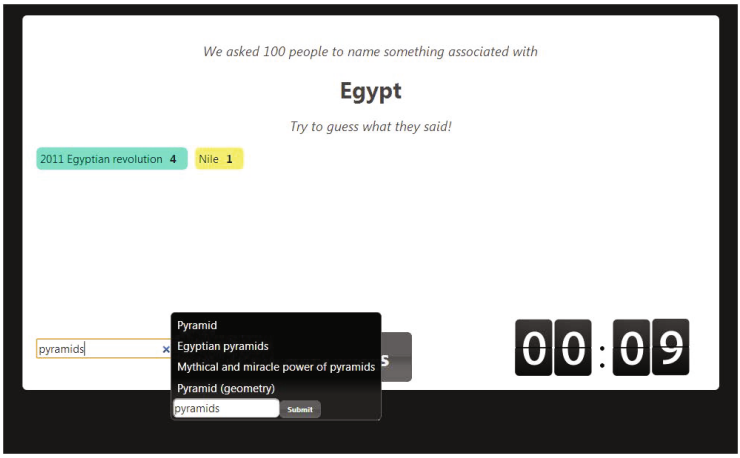

Fig. 1. A player has submitted 2 associations already for the topic "Egypt" (scored 4 points for the first one and 1 point for the other), and is now viewing the suggestions after guessing "pyramids"

3 The Knowledge Test Game

The Knowledge Test Game is a Game With A Purpose (GWAP), aiming at collecting and ranking associations. Players provide associations to Linked Data entities, whereas the associations themselves are Linked Data entities as well. The game is available on `http://www.knowledgetestgame.org` and through Facebook on `http://apps.facebook.com/knowledgetestgame/`.

3.1 Walkthrough

Each round of the Knowledge Test Game is referred to as a game instance, or simply a *game*. Each game has 2 to 10 players, all seeing the same *topic*, which is a Linked Data entity for which we would like to collect associations. Upon visiting the Knowledge Test Game homepage, players can choose to directly play a game or go through the *How to play* interactive tutorial. Furthermore, the players are able to authenticate themselves by logging in using their Google or Facebook accounts, or they can play anonymously as guests.

Joining a Game. When a player chooses to join a game, he either directly joins a random running game or creates a new one. A player can only join games that have less than 10 players, and have not been running for more than 70 % of its time. Additionally, the topic of the game being joined must be suitable according to the topic restrictions for that player (see Section 3.3).

Within a Game. Once a player joins a game (see Figure 1), he is presented with the statement: "We asked 100 people to name something associated with **Egypt** try to guess what they said!", where "Egypt" is the current game's topic. The mention of 100 people is a white lie in order to remind of the well known Family Feud TV show. This form of the question communicates to players that subjectivity should be avoided.

In a duration of 45 seconds, shown by a timer, the players are able to submit guesses resembling what they think is associated to the topic. For each submitted *guess*, the player gets a list of suggestions from which he can select the one most relevant to what he had in mind. The selected suggestion is then submitted as an *association* to that topic. If none of the suggestions were satisfactory, the player can still submit his guess as it is. The process of displaying relevant suggestions is managed by the suggestions-box, which is discussed in Section 3.2.

Throughout the game, each player can see the associations he submitted along with the score of each. The scores are increased dynamically when others have submitted the same association. This motivates players to enter associations that others would agree upon, consequently countering the subjective nature of the players' inputs.

The Recap Page. When the game's time is elapsed, the players are forwarded to the recap page, where they can see the associations submitted by all other players, as well as their scores. Players can then decide whether to join the next game in the series with the same players or join a new one.

3.2 The Suggestions-Box

The Knowledge Test Game offers a suggestions feature that enhances the data collection process, in addition to making the game more entertaining. The most important purpose of the suggestions-box, is to link the players' text input back to Linked Data. Each of the suggestions corresponds to a Linked Data entity. Since the topic is a Linked Data entity as well, linking the topic and an association results in connecting Linked Data entities.

The suggestions-box makes it easier to match submissions. Facilitating the matching process is in our interest as well as the players', since we will be getting more useful information, and the players will be getting more matches and consequently better scores.

The Knowledge Test Game does not rely on the submitted guess to find a match, but rather uses it as a clue to display relevant associations, and then collect the selected association afterwards. For example, if the current topic is "Egypt", and three different players submitted "pyramids", "the pyramids" and "Egyptian pyramids". It would be challenging to detect a match, although they could have meant the same thing. On the other hand, once the suggestions-box displays the suggestions for each of these guesses, the players would eventually pick the association that they meant, which could be "Egyptian pyramids", realizing that it best matches what they had in mind. Consequently, the three

players will get matches and therefore bonus points, and the game will give the association `dbpedia:Egyptian_pyramids` a higher rank.

Another immediate benefit of the suggestions-box is to distinguish ambiguities. When a player submits "pyramids" as a guess, he could have meant the geometric shape, the Egyptian pyramids, the Mayan pyramids, or anything else named pyramid. The suggestions-box clears these ambiguities, by allowing the player to further distinguish what he has meant by his guess.

The suggestions-box makes use of features from Google and Bing, which include auto-correction and being lenient towards different representations of the same word. Therefore the possible negative impact of using different dialects, or even languages, is absent. For example, submitting the British "organisation" and the American "organization" will result in two very similar, if not identical, suggestions lists. Players can even enter hints to the association instead of an exact association name. For example, a player can submit "c inventor" as a guess for "Deaths in 2011", and get a suggestions list that includes "Dennis Richie", who died in 2011, and who is also the inventor of C.

Furthermore, the suggestions-box can accept any language, including complex ones such as Arabic, or even transliteration[2] of Arabic words in English literals, and still yield relevant results. Nevertheless, regardless of the used language, the resulting suggestions always correspond to English Linked Data entities.

The Other Box. Players are also allowed to submit their guess as it is, by using the *other box* at the bottom of each suggestions list. Submitting a guess this way allows the player to come up with own associations which are not well represented or outside the scope of Wikipedia, at the expense of making it harder to match with other players. In order to get bonus points for an association submitted using the *other box*, other players have to submit the exact same string. In order to analyze the importance of such an association the game creates URIs of the form `ktg:<topic>/association/<association>`, creating new Linked Data entities (for a discussion of this effect see Section 5).

Approaches to Implement the Suggestions-Box. The goal was to present the players with associations relevant to the entered guess, in the context of the topic in question. Therefore, the retrieval method is a function of the player's *guess* and the game's *topic*.

The initial step was to manually collect associations for topics, to formulate a ground truth, with which we could benchmark different methods of collecting associations. We asked 9 participants to name associations to random topics, each coupled with one or more links to corresponding Wikipedia articles, ordered by relevance. We collected a total of 224 Wikipedia articles as associations to 32 different topics (full list is available at `http://goo.gl/hXhFt`).

[2] Transliterating Arabic words to English is common over the Internet in the Arab world. See `http://en.wikipedia.org/wiki/Arabic_Chat_Alphabet`.

Table 1. The mean *Recall@10* and mean *GamePlayability@10* achieved by each methods in attempt 1

	Mean Recall@10	Mean Game Playability@10
DBpedia Spotlight	26.17%	29.91%
Freebase	34.15%	39.28%
Bing	40.3%	48.6%
Google	49.69%	59.81%

After collecting the ground truth, we started testing different methods of retrieving these links in order to find a suitable one to be used for the suggestions-box. The first attempt to retrieve relevant links, was to query for the *Topic* and the entered *Guess*. We refer to this query as $T + G$.

To evaluate the results, we used *Recall@k* by calculating the percentage of the ground truth links retrieved out of the top k links obtained using the retrieval method. It was also significant to see if the retrieval method was able to retrieve any of the ground truth links at all. For this we defined a metric, called *GamePlayability@k*, which is 1 if any of the ground truth links exist within the first k retrieved links, and 0 otherwise.

In an effort to provide players with ten relevant suggestions for each guess, various APIs were evaluated to seek the highest *GamePlayability@10* and *Recall@10*. Among the tested APIs were DBpedia Lookup API, which was excluded for its strictness, as it expects a query string that is an exact substring of a URI's label. Wikipedia API had a very slow response rate for an interactive game, and was excluded accordingly. Finally, we tested the query using DBpedia Spotlight, Freebase, Bing and Google (see Table 1).

In the second attempt, we classified the results into three categories: those related to both the *Topic* and the *Guess* ($T + G$), those related to the *Guess* only (G), and those related to the *Topic* only (T). We reached a hypothesis that we can achieve better results by searching for $T + G$, in addition to promoting results common with G, and demoting those common with T. We refer to this merging process as `merge(T+G, G, T)`.

Google and Bing were preferred for this attempt because of their previous plausible results, and their quick response rate. Upon applying `merge`, there was a considerable increase in both the *Recall@10* and *GamePlayability@10*. Bing got a mean *Recall@10* and a mean *GamePlayability@10* of 71.34% and 77.57% respectively, while Google got 79.78% and 85.51%.

Google's results were better, while Bing had a faster response rate. We exploited this for the third attempt, by making three concurrent requests to each search engine. The final results are then passed to the merging algorithm again `merge(mergeGoogle, mergeBing, [])`, where `mergeGoogle` and `mergeBing` were the results of applying `merge(T+G, G, T)` on Google and Bing respectively.

This further increased the mean *Recall@10* and mean *GamePlayability@10* to 80.37% and 86.45% respectively, to reach the highest values we could achieve, without introducing any time overhead.

3.3 Topic Selection

Presenting players with topics that they are familiar with increases the fun factor of the game, as well as the validity of the results, since users with interest in a topic are more qualified to provide valid associations.

In order to focus on topics that are likely to be known, we collected the top most visited $10K$ Wikipedia articles in 2011[3]. Knowing that each of these articles corresponds to a Linked Data entity, the topics are randomly selected from their titles.

There are some restrictions in the context of topic selection that increase the validity of the players' submissions. These restrictions are shared by all the players within the same game. For example a topic cannot be played by the same player more than once, as we wanted to exclude possible influence from earlier games.

The Knowledge Test Game is also available on Facebook. By logging in using a Facebook account, the topic selection process is additionally influenced by the players' likes on Facebook, to make it more likely to get topics of interest.

If 50 unique players provided associations to a topic, the topic will be marked as *done*, and can be optionally prevented from appearing in future games. This gives the chance to analyze the collected associations, and to focus on other topics. The topic selection algorithm is biased towards closing topics as early as possible, meaning that if there are several topics available for a game, the one that was played most is preferred.

3.4 Generated Dataset

We keep track and log a lot of data based on the users' input. The data is made available online through `http://knowledgetestgame.org/export`. The main components of interest are the players' guesses. For every submission, the guess string provided by the player is stored along with the list of suggestions that he sees afterwards. Within the same record we also log the game's ID, the topic's name and URI, as well the player's ID and account type (the ID hides all potentially personal information about the player).

When a player selects an association from the list, the same record is updated to hold the association's URI and its index with respect to the suggestions list. The time of submitting the guess, and the time of choosing the associations are both stored. We also keep track of the time taken by the player, in milliseconds, to choose the association from the list. The number of occurrences and the score of the association across the game are also logged. Furthermore, each record holds "nth guess" and "nth association" which show the record's submission order as a guess and its order as an association by that player in the given game.

4 Evaluation

After the previous sections focused on the game, its suggestion box and topic selection, we will now provide a detailed evaluation of the game itself and of the generated output.

[3] Obtained from `http://dumps.wikimedia.org/`

4.1 The Game

First, the game's concept and its realization are evaluated by summarizing measurements and derived estimates. Afterwards, the outcomes of a questionnaire, which was presented to players of the game, are provided.

Measurements and Estimates. The game was run in several focused experiments, that added up to 26.6 hours of game-play time by humans. In these experiments the game was played by 267 different players who played a total of 1046 games together collecting 6882 ranked associations.

Using these numbers we can evaluate the game wrt. the *throughput*, *average lifetime play* and *expected contribution* metrics for Games with a Purpose defined by von Ahn and Dabbish [5].

The *throughput* is calculated by dividing the collected data (6882 ranked associations) by the total human game-play time (26.6 hours), resulting in ~ 259 ranked associations per human hour. At this rate if there were 50 players online for a day playing the game (a decent estimate for typical online games), we could collect about 310 800 ranked associations in a single day.

We can also compute the *average lifetime play* by dividing the total game-play time (26.6 hours) by the number of players (267), resulting in an average lifetime play of ~ 6 minutes per player, which is equivalent to the time needed for ~ 8 games.

Finally, we can calculate the *expected contribution* by multiplying the average lifetime play with the throughput, resulting in an expected contribution of $\sim$ 25.78 ranked associations per player.

Questionnaire. Apart from the metrics in the previous section, we conducted an online survey which was filled out by 21 players after playing the game. Most of the participants were students from Egypt and Germany, between 20 and 25 years old, had a computer science or engineering background, had played web games before and described their English skills as fluent. Besides these demographic questions, the survey consisted of 3 open and 13 5-point Likert scale questions. The 3 open questions were asked beforehand in order not to influence participants. The text of the questions was: "What did you like about the game?", "What did you dislike about the game?" and "What would you improve?".

Summarizing most players liked the idea of the game and described it as fun, mentally challenging and interesting to compare their own thoughts to those of others. Many participants mentioned that they enjoyed the topic mix and were surprised by the quality of the suggestions-box:

> It is very challenging, not only are you challenging the other players, but also your own knowledge The topics are very good. The recommended words are very good, Ex. I got the topic "Princess Diana" and I wanted to add the name of the man she was with in the car accident but I couldn't remember his name, I just know he was Egyptian, I wrote down "Egypt" and I found "Dodi Al Fayed".. very cool!:)

In the dislike section it was mentioned that some topics were too vague or unknown, that the suggestions-box sometimes was slow and that the 45 seconds per round were not sufficient to enter all your associations in some cases. Also some participants complained about the little information they got about other players which was in line with the improvements section.

Here we received a lot of feedback that can be grouped into the category enhancing the interaction with and information about other players. Many participants want to know more about the people they're playing with and suggested to introduce a chat after the game in the recap phase. Others want to be able to play with their friends. Also participants mentioned that they would want to see global high-scores after the round and live stats of other players in their game during the game, so they don't have to wait for the recap page to see their own performance. Furthermore, it was suggested to provide the ability to select categories of topics to play, to show photos for the topic or for vague topics to provide hints by showing some of the most often entered associations.

Table 2. Results of an online survey answered by 21 game players. Except for *Age*, users could select answers from a 5-point Likert scale. If not indicated otherwise the options were: 1 (Strongly disagree), 2 (Disagree), 3 (Neutral), 4 (Agree), 5 (Strongly agree).

Statement	μ	σ
"The game rules and concept were direct and straight forward."	4.5	0.8
"The How To Play tutorial was..." (useless ... useful)	4.8	0.4
"45 seconds for the game were..." (too short ... too long)	2.6	0.7
"The topics were clear and know to me."	4.0	0.8
"The suggestions were relevant to what I had in mind."	4.1	1.0
"The suggestions that I got for a guess influenced my following guesses."	4.0	0.9
"15 seconds for the recap page were..." (too short ... too long)	3.1	0.6
"I understood the recap page."	4.6	0.6
"I was interested in reading the scores in the recap page."	4.5	0.7
"Seeing my partner's answers influenced my guesses in the following games."	3.2	1.3
"I enjoyed the game."	4.5	0.7
"I would play it again"	4.3	1.2
"I played web games before."	4.0	1.2

The findings from the open questions were refined by 13 questions in which participants could select numerical values between 1 and 5 (5-point Likert scale). The results are summarized in Table 2. In general we can see that the game concept was easy to understand, people found the tutorial useful, knew the topics, found the suggestions relevant to what they had in mind, understood the recap page and were interested in it and that most people enjoyed the game and would play it again. The timing restrictions of 45 seconds per round was perceived as slightly too short, but 15 seconds for the recap page were just right.

The questionnaire identified a key problem, namely that many participants had the feeling the suggestions-box influenced their following guesses. This effect was later mitigated by reducing the suggestions from ten to four (see Section 5). The effect seems to be less pronounced for the recap page.

Before discussing these findings and possible solutions, we first want to present our evaluation of the data collected.

4.2 Data Quality

In order to assess the quality of the collected data, we aggregated the associations collected by the game for each topic. Focusing on topics for which the most associations were submitted by players, we counted the number of occurrences of each association and ordered them descending by counts. In this process we excluded associations which were submitted by less than two players as a provisional filter against noise.

After the first major experiment, the resulting ordered lists of associations for the 10 topics which were played most often were generated. With these lists we conducted another online questionnaire with 36 participants out of which 19 had played the game. The participants' demographics resembled those of the game players: they mainly were computer science students from Egypt and Germany, between 20 and 25 years old and described their own English skills as fluent. In the questionnaire for each of the topics we asked the participants to rate the ordering of the list of associations on a scale from 1 (Makes no sense at all) to 5 (Makes perfect sense). The histogram of the ratings can be found in Figure 2 and indicates that the majority of participants were very satisfied with the presented associations and their ordering. With $\mu = 4.2$ the average over all ratings ($\sigma = 0.9$) is close to its maximum of 5.

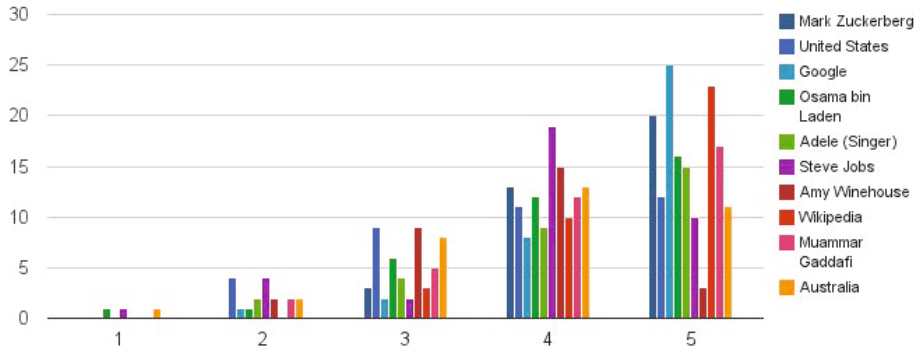

Fig. 2. Histogram of ratings for the ordered lists of associations. For each topic the participants could chose on a scale from 1 (Makes no sense at all) to 5 (Makes perfect sense).

After a second large experiment we chose another form to evaluate the generated association lists (an example can be seen in Table 3). We again conducted an online survey, this time with 17 participants, where they were asked to rank given randomized lists of the top-20 associations for the most often played topics. By then, we had 15 *done* topics (i.e. played by more than 50 players). Out of these 15 topics the 9 lists summarized in Table 4 were picked to form a ground truth, as they had been ordered manually by more than 5 participants. The ground truth was formed by averaging the individual ranks of the manually

Table 3. The most frequently submitted associations for the topic Mark Zuckerberg

Association	Times mentioned
Facebook	50
The Social Network	15
Chief Executive Officer	12
Rich	8
Millionaire	7
Social Network	6
Entrepreneur	5

Table 4. The 9 most often played topics. The associations are printed as titles here instead of the URIs of the corresponding DBpedia instances. Each topic's associations lists were presented in the questionnaire in a randomized order, where participants were asked to rank them. The resulting ranks were then compared with the nDCG to those generated with the Normalized Google Distance (NGD) and the game.

			nDCG	
Topic	Top-N Associations	Manual sorting	NGD	Game
Charlie Sheen	8	7 participants	0.860	0.969
Eminem	11	14 participants	0.870	0.931
Lady Gaga	18	9 participants	0.806	0.924
Mark Zuckerberg	7	15 participants	0.895	0.954
Osama bin Laden	12	7 participants	0.814	0.835
Transformers: Dark of the Moon	18	6 participants	0.768	0.926
United Kingdom	14	7 participants	0.806	0.873
World War II	17	17 participants	0.876	0.953
YouTube	10	17 participants	0.927	0.928
		μ	0.847	0.921
		σ	0.051	0.042

ordered lists of the participants and sorting the associations accordingly. Afterwards, the normalized Discounted Cumulative Gain (nDCG) was calculated to compare the manually ranked ground truth association lists with those retrieved by the game. As a relevance metric, we used a linear mapping of the top element to a relevance of 1 down to the last element with a relevance of $\frac{1}{n}$.

In order to differentiate our game's results from simple corpus based similarity metrics, we also re-ranked the ground truth lists according to the popular Normalized Google Distance (NGD) [12]. As the NGD calculates a similarity between pairs of entities only and cannot trivially be used to find the top candidates for a given topic we artificially enhanced the method by only focusing on the top-20 candidates in the ground truth. The nDCGs can be found in Table 4 as well. We discuss our results and findings in the next section.

5 Discussion

After detailing our evaluation in the previous section, we will now discuss our findings. In summary we were very satisfied with the results of our evaluations, as the game was well perceived and fun for the players and also collected associations of high quality.

We consider the achieved throughput of 259 associations per human hour quite satisfactory, as it means that on average less than 14 seconds were spent for typing in a guess string, waiting for the suggestions-box and selecting one of the alternatives. As many players complained that the suggestions-box was slow we investigated our server logs to find that under high load it seems our requests to Google were rate limited, resulting in an average response time of the suggestions-box of approx. 2.3 s. At the same time all 3 requests to Bing on average return within 250 ms. As we also got a lot of feedback that the quality of the suggestions-box is astonishing, we would like to keep using the merged results of Google and Bing. In order to decrease the delay we consider more aggressive caching. Also we plan to include incremental updates of the suggestions list to lower the waiting time and increase the throughput in future versions.

In order to solve ambiguity issues of the strings displayed in the suggestions list, we plan to display the `rdfs:comment` or a useful `rdf:type` from DBpedia in future versions. At the same time a `foaf:depiction` could be shown to make the suggestions visually recognizable. As queries to the online DBpedia will take additional time we again consider caching and asynchronous updates of the GUI.

The evaluation also revealed the problem that later guesses were likely to be influenced by the displayed suggestion lists for preceding guesses. Throughout the experiments we therefore collected the index (zero-based) of the association selected from the suggestions list. On average the second (1.04) suggestion was selected with a standard deviation of 1.7 in the first major test. Based on that, we recalculated the *Recall* and *GamePlayability* using different ks ranging from 1 to 10. *Recall*@4, which translates to showing 4 suggestions, was found to be a suitable compromise to mitigate the influence (see Figure 3). Another alternative we want to investigate in the future consists of further reducing the amount of suggestions and providing a "more" button.

We were very pleased with the evaluation of the data quality, as the game shows a high average nDCG of 0.921 (Table 4) in comparison to the ground truth. The comparison to a popular corpus based technique shows that even when enhanced with an oracle that only suggested the associations we consider correct, the corpus based technique still was not able to rank the associations as well as the game (average nDCG of 0.847).

Last but not least, we investigated a potential design issue of our approach, which links Linked Data entities to one another. Our approach thereby neglects the possibility that people could want to associate a Linked Data entity with one of its Literals. Hence, we studied the list of all associations which were submitted with the "other" option of the suggestions-box and all guesses for which no association was selected, coming to the conclusion that not a single one of them corresponded to a desired but missing literal value in the suggestion lists. Also we were surprised how seldom players seemed to have missed an association target. From this we conclude that even though theoretically possible it seems to be very rare that people would want to associate an entity with one of its literals or cannot find a desired association target in the domain of Wikipedia. Nevertheless, in future versions we plan to explicitly log events when one of the

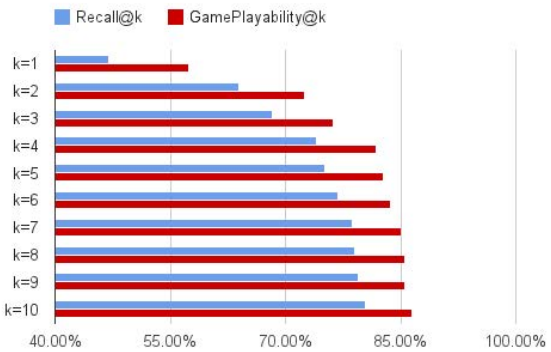

Fig. 3. The Recall@k and GamePlayability@k from k = 1 to k = 10

guesses matches one of the topic's literals and when newly created linked data entities are matched multiple times.

6 Conclusion and Outlook

In this paper we presented our idea to rank Linked Data facts according to human association strengths to cope with the increasing information overflow when performing simple queries on Linked Data entities. In order to collect a dataset of such association strengths between Linked Data entities we developed a game with a purpose called "Knowledge Test Game".

Our evaluations show good results wrt. throughput and perceived fun of the game, especially the quality of the suggestions box received a lot of positive feedback as it is even able to retrieve complex, clue based associations. Furthermore, collected data seems to be of very high quality.

Apart from the planned improvements mentioned in Section 5 our future work on the game will mainly focus on making it more desirable for players to stay in the game in order to collect more and more data, for example providing a chat on the recap page, global high-scores, an exponential scoring scheme, player ranks and permissions (such as reporting cheaters). We would also like to experiment with social gaming aspects such as team games by taking more advantage of the Facebook integration. Furthermore, we plan to provide a transparent single-player mode where players play against recorded sessions of other players in order to reduce waiting times, validate existing data and detect cheaters.

In terms of data quality we want to investigate other aggregation methods, for example taking the submission order of the associations into consideration. Also we would like to experiment with the thresholds to close topics as well as exclude noisy associations.

Last but not least we want to use the collected association data published at `http://knowledgetestgame.org/export` to evaluate existing or future methods to rank Linked Data according to human associations.

This work was financed by the University of Kaiserslautern PhD scholarship program and the BMBF project NEXUS (Grant 01IW11001).

References

1. Berners-Lee, T., Hendler, J., Lassila, O.: The Semantic Web. Scientific American 284(5), 34–43 (2001)
2. Bizer, C., Heath, T., Berners-Lee, T.: Linked Data - The Story So Far. International Journal on Semantic Web and Information Systems 5(3), 1–22 (2009)
3. Bizer, C., Lehmann, J., Kobilarov, G., Auer, S., Becker, C., Cyganiak, R., Hellmann, S.: DBpedia - A crystallization point for the Web of Data. Web Semantics: Science, Services and Agents on the WWW 7(3), 154–165 (2009)
4. Berners-Lee, T.: Linked Data - Design Issues (2009)
5. von Ahn, L., Dabbish, L.: Designing games with a purpose. Communications of the ACM 51(8), 58–67 (2008)
6. Hees, J., Roth-Berghofer, T., Biedert, R., Adrian, B., Dengel, A.: BetterRelations: Collecting Association Strengths for Linked Data Triples with a Game. In: Ceri, S., Brambilla, M. (eds.) Search Computing III. LNCS, vol. 7538, pp. 223–239. Springer, Heidelberg (2012)
7. Siorpaes, K., Hepp, M.: Games with a Purpose for the Semantic Web. IEEE Intelligent Systems 23(3), 50–60 (2008)
8. Hees, J., Roth-Berghofer, T., Dengel, A.: Linked Data Games: Simulating Human Association with Linked Data. In: LWA 2010, Kassel, Germany, pp. 255–260 (2010)
9. Hees, J., Roth-Berghofer, T., Biedert, R., Adrian, B., Dengel, A.: BetterRelations: Using a Game to Rate Linked Data Triples. In: Bach, J., Edelkamp, S. (eds.) KI 2011. LNCS, vol. 7006, pp. 134–138. Springer, Heidelberg (2011)
10. Kny, E., Kölle, S., Töpper, G., Wittmers, E.: WhoKnows? (October 2010)
11. Wolf, L., Knuth, M., Osterhoff, J., Sack, H.: RISQ! Renowned Individuals Semantic Quiz – A Jeopardy like Quiz Game for Ranking Facts. In: Proc. of the I-SEMANTICS, Graz, Austria, pp. 71–78. ACM (2011)
12. Cilibrasi, R.L., Vitányi, P.M.B.: The Google Similarity Distance. IEEE Trans. Knowledge and Data Engineering 19(3), 370–383 (2007)