BetterRelations: Collecting Association Strengths for Linked Data Triples with a Game

Jörn Hees[1,2], Thomas Roth-Berghofer[2,3], Ralf Biedert[2], Benjamin Adrian[2], and Andreas Dengel[1,2]

[1] Computer Science Department, University of Kaiserslautern, Germany
[2] Knowledge Management Department, DFKI GmbH, Kaiserslautern, Germany
[3] School of Computing and Technology, University of West London, UK
{firstname.lastname}@dfki.de

Abstract. The simulation of human thinking is one of the long term goals of the Artificial Intelligence community. In recent years, the adoption of Semantic Web technologies and the ongoing sharing of Linked Data has generated one of the world's largest knowledge bases, bringing us closer to this dream than ever. Nevertheless, while associations in the human memory have different strengths, such explicit association strengths (edge weights) are missing in Linked Data. Hence, finding good heuristics which can estimate human-like association strengths for Linked Data facts (triples) is of major interest to us. In order to evaluate existing approaches with respect to human-like association strengths, we need a collection of such explicit edge weights for Linked Data triples.
In this chapter we first provide an overview of existing approaches to rate Linked Data triples which could be valuable candidates for good heuristics. We then present a web-game prototype which can help with the collection of a ground truth of edge weights for triples. We explain the game's concept, summarize Linked Data related implementation aspects, and include a detailed evaluation of the game.

1 Introduction

Since its introduction in 2001 the Semantic Web [1] has gained much attention. In recent years, especially the Linking Open Data (LOD) project contributed many large, interlinked and publicly accessible RDF datasets, generating one of the world's largest, decentralized knowledge bases [2]. The accumulated amount of Linked Data has many applications and can already be used to answer structured questions (e.g., the DBpedia [3] dataset can easily be used to compile a list of musicians who were born in Berlin).

Nevertheless, it currently is unclear how to rank result sets—not even those of simplistic (descriptive) queries—by importance as considered by an average human. For example, asked to describe (What/Who is ...?) a *topic* such as "Facebook", nearly all humans will explain that it "is an online social network", but only few will tell us that "Chris Hughes is one of its co-founders". In the remainder of this chapter, we will hence call the fact "Facebook has subject online social

networking" more *important* than "Facebook has key person Chris Hughes" wrt. the topic "Facebook". Despite the fact that this importance relation is surely user and context dependent, we want to focus on an average human's view, leaving the application of user and context models to future work. In terms of [4] our definition of importance balances formality, stability and sharing scope mostly by focusing on a wide sharing scope and being applicable to cross-domain datasets such as DBpedia.

In contrast to this human view, triples in Linked Data, which are (subject, predicate, object)-statements, also called *facts*, are facts in a logical sense. Like logical axioms, they all are of the same "importance", none being more valuable than another. Given a *topic* (e.g., `dbpedia:Facebook`) there is no easy way to order its more than 100 related facts in DBpedia by importance. This leads to problems, for example when a user requests a concise description[4] of a resource.

A collection of such importance information would allow us to ask machines not only to give us all known facts related to a resource in an arbitrary order, but also to rank this information by importance, allowing us to constrain the number of results to the most important ones (e.g., the top 10).

Aside from concise descriptions the applications of a method to rank facts about a given topic from Linked Data are manifold. With regard to Artificial Intelligence this would provide a basis for human-like reasoning on Linked Data (e.g., using spreading activation approaches [5] for semantic search [6] with meaningful edge weights) and enable us to drastically reduce the search space to only those concepts strongly associated with the current context by an average human. Another immediate benefit from annotating Linked Data triples with association strengths is the possibility of feedback for automated extraction processes, e.g., the one underlying DBpedia. One could investigate, which extraction rules yield high and which ones yield low strengths, facilitating a quality assurance process.

Besides these immediate benefits, such a collection of association strengths would also allow us to investigate whether currently used approaches to rank Linked Data (e.g., based on network analysis approaches, such as PageRank [7] and HITS [8], trying to model how much activation flows from one concept to another, or based on semantic similarities, such as estimated by word co-occurrences on websites) truly model how we associate thoughts. If this is the case, the heuristics could be used to bootstrap the acquisition of associations strengths for Linked Data triples, else such a dataset would be a valuable prerequisite to develop heuristics to estimate triple importances.

Despite all the benefits a collection of Linked Data triples rated by human association strengths would have, it suffers from the typical knowledge acquisition bottleneck. Collecting such strength values is prone to subjectivity, it is extremely monotonous and tedious, and it is difficult for humans to reliably and objectively assess the absolute strength value of a triple. Furthermore, the immense amount of Linked Data would cause great expenses if people were to be paid for rating even a small part thereof.

[4] Description as in SPARQL DESCRIBE queries.

In order to overcome the aforementioned problems, we sketched the idea for a web-game in [9] and briefly described our findings from developing a prototype called *BetterRelations* in [10] following the "Games With A Purpose" approach by von Ahn and Dabbish [11].

The rest of this chapter is structured as follows: We first give an overview of existing approaches to rank Linked Data (Section 2) and Games With A Purpose related to BetterRelations (Section 3). Afterwards we provide a detailed description of the game's concept as well as data acquisition and necessary preprocessing steps to present Linked Data triples to players in a comprehensible format (Section 4). Furthermore, we report on a detailed evaluation, consisting of statistics, the results of a user questionnaire and a comparison of the game results with manually generated ranked lists by a test group (Section 5), as well as a discussion of our findings, identifying possible improvements and future work (Section 6).

2 Existing Approaches to Rank Linked Data

The need for mechanisms to rank Linked Data grows with the ongoing adoption of Semantic Web technologies. In recent years, a variety of approaches have been developed. For an easier understanding we want to structure them into approaches which mainly analyze the graph structure of Linked Data itself and approaches which use Linked Data external information sources to rank Linked Data.

2.1 Approaches Using Graph Analyses

As Linked Data can be represented as a graph it is not surprising that many ranking approaches focus on the structural aspects of this graph. Most of these approaches try to apply well known ranking algorithms for the World Wide Web such as PageRank [7] or HITS [8] to the Semantic Web.

ObjectRank [12] was one of the first such approaches applying PageRank on databases modeled as labeled graphs. In order to reduce the Linked Data graph with different link types to a graph with just one link type on which PageRank can operate, ObjectRank requires domain experts to manually assign weights for each link type, which is impractical on large scale, evolving datasets such as Linked Data. As ObjectRank was developed with a single database system in mind, it does not track provenance information and hence is possibly vulnerable to spam.

Swoogle [13] was one of the first search engines for the Semantic Web, using OntoRank and TermRank for ranking. OntoRank ranks RDF documents with PageRank. TermRank ranks classes and properties by their popularity which is composed of their usage counts in other RDF documents and their OntoRank distributed over all classes and properties which are used. One main drawback of Swoogle is its inability to rank instances.

This shortcoming of Swoogle was addressed by the Naming Authority [14] approach. It ranks Linked Data resource and literals by calculating the PageRank on the interlinkage of source documents and then propagating the source rank to their resources and literals. The re-use of IDs (URIs) minted by other naming authorities (top level domains or pay level domains) increases their rank and provides spam resistance as it takes the provenance of RDF statements into account. Nevertheless, the same mechanism neglects dataset internal link structures, which are of importance w.r.t. big datasets such as DBpedia.

Hence, DING (Dataset Ranking) [15], which is currently used in Sindice [16] extends [14]. It uses two layers: the dataset graph and the entity graph. As in [14] the dataset graph consisting of links between datasets is used to compute the dataset ranks based on PageRank. The calculated dataset ranks are then combined with semantic-dependent entity rankings (which can be different for different datasets), such as PageRank or a simple in-degree. By this the approach has the ability to better model peculiarities of specific datasets.

In contrast to the aforementioned approaches which are based on PageRank, TripleRank [17] represents the RDF graph as 3D tensor and uses a tensor variant of HITS. By this TripleRank allows the identification of and grouping by similar properties. Despite its promising results, TripleRank is vulnerable to spamming as it does not track provenance information and includes a pruning step which removes properties that could potentially encode very useful information for semantic similarity (e.g., the DBpedia `dbo:wikiPageWikiLink`).

The last approach we want to mention is called SemRank [18] and is an information theoretic approach. Given two resources it ranks possible complex relationships (multi-step paths) between them based on information gain for the user. The user can configure the system to be rather conventional (low information gain) or use it in a discovery mode fashion (high information gain). For this SemRank combines three different components. Aside from providing a semantic keyword matching on the labels of involved properties, SemRank calculates the specificity of properties and refractions of a complex relation. The specificity of a property describes how unique it is w.r.t. the knowledge base and w.r.t. where it could be used due to domain and range restrictions. The refraction count measures how many different vocabularies a complex relation spans. A high specificity or high refraction count increase the rank in discovery mode but decreases the rank in conventional mode.

2.2 Approaches Using Graph External Features

The previously mentioned approaches all limit themselves to information which is available by analyzing Linked Data and especially its graph structure. We now want to focus on approaches which also use external information. Many of the following approaches are not originally devised to rank Linked Data, but instead focus on semantic similarity or semantic relatedness of terms, which are closely related to human association strengths. In order to apply such approaches to Linked Data, usually labels [19] are used to map between Linked Data instances and instances in external data sources.

In order to estimate the semantic relatedness of two concepts, many approaches are based on WordNet [20]. WordNet is a large lexical database of English words. Amongst others, WordNet groups words into synonym sets and provides hierarchical relations between them, such as hypernyms and holonyms. Most WordNet based relatedness measures use features of the hierarchical structure, such as the length of shortest paths between concepts or the overlap of synsets. An evaluation of WordNet-based semantic relatedness measures can be found in [21]. Despite its size and quality the main disadvantage of using WordNet is that it is far from complete and quickly becomes outdated (trend words such as "iPad" are still missing).

To overcome these issues other approaches are based on Wikipedia and typically focus on structural features of the corresponding articles in Wikipedia, such as the disambiguation pages, hierarchy of categories, listings, and WikiLinks (links between articles). For example, WikiRelate [22] uses the disambiguation pages, letting two concepts disambiguate each other, in combination with text overlap and category tree search for a lowest common category ancestor of Wikipedia articles to calculate the semantic relatedness of the concepts.

Another group of similarity measures focuses on distributional aspects of words and their co-occurrence in large text corpora (e.g., online documents) or social online platforms. Approaches in this group are typically based on the count of scopes in which both terms co-occur, as well as the counts of scopes in which they occur independently and then try to estimate the significance of the co-occurrences. Examples for such similarity measures include the Normalized Google Distance [23] (actually often applied to other search engines as well) and tag relatedness in social bookmarking systems [24]. Further such distributional systems can be found in [25], also including an approach which combines co-occurrence based measures ones based on WordNet-based.

Some of the aforementioned approaches, especially those depending on external datasets such as WordNet and Wikipedia, can actually be performed on Linked Data, as for most of such datasets mappings are existent, nowadays. Still such approaches typically use very specific knowledge about these datasets (and their mappings to Linked Data) in contrast to the methods presented in the previous section.

The last approach we want to mention is DBpediaRanking [26], an approach which makes use of such a mapping which maps Wikipedia to its Linked Data pendant DBpedia. DBpediaRanking finds semantically related terms for a given DBpedia resource. To some extent it can also be seen as a hybrid approach combining graph structural features and external information. DBpediaRanking exploits the graph-based nature for a limited depth-first search restricted to predefined properties (`skos:subject` and `skos:broader`). The discovered nodes are compared to the root node by a scoring mechanism which focuses on nodes that are encountered frequently during the discovery step (important nodes). The scoring includes similarity measures derived from co-occurrences of both `rdfs:label`s in web documents by querying search engines such as Google and Yahoo and online bookmarking services such as Delicious. The scoring mechanism also

ranks nodes higher which have bidirectional `dbo:wikiPageWikiLink`s with the root node (an idea which can also be found in [27]), and scores nodes higher which have bidirectional `dbo:abstract` inclusions of their `rdfs:label`s with the root node. The hybrid approach chosen by DBpediaRanking shows promising evaluation results.

The last approach indicates that a combination of procedures using structural features of the graph and techniques using information from external datasources might interesting for future research. As mentioned in the introduction to conduct such research, a collection of Linked Data triples rated by humans would be very helpful, especially considering the fact that in many of the presented approaches evaluations were limited to a small group of people and performed on small fractions of the datasets they should be able to rank.

3 Related Work

In terms of game design, BetterRelations is related to *Matchin* [28]. Matchin is a two player web-game, which confronts pairs of players with two pictures (taken from the WWW), asking them which one they prefer. If the preferences of both players match, the players are rewarded with points and an increasingly higher bonus. In case of a mismatch, they are not rewarded with points and the bonus is reset to 0. In this process, decisions which both players agree on are considered more valuable than mismatches. In the background Matchin records the pairwise user preferences and uses them to compute a global rating of the played images. In contrast, BetterRelations presents two textual facts corresponding to Linked Data triples about one topic to its players. Whereas Matchin creates a globally ranked list of images, BetterRelations computes a ranking for each topic and its related facts. Hence, the rating algorithm, which transforms the pairwise user preferences into the global ratings hence has to deal with significantly smaller lists. As detailed in Section 4.1, BetterRelations includes several additional features in order to make Linked Data issues such as noise or unknown facts tractable.

OntoGame [29] was the first and most prominent game with a purpose focusing on Linked Data. Nevertheless, it collects another type of information than BetterRelations: Players are asked to decide if a Wikipedia topic is a class or an instance, aiming at creating a taxonomy of Wikipedia.

WhoKnows? [30], 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 only uses a limited fraction of the DBpedia dataset and excludes triples not matched by a predefined domain ontology in a preprocessing step. This greatly reduces noise issues, but eliminates the possibility to collect user feedback about triple qualities and problems in the extraction process. Also, WhoKnows intends to rank triples by degree of familiarity. However, the used measurement only relies on the ratio of correctly recognized facts divided by number of times a fact was tested. The

quality of this ratio is doubtful as it does not distinguish whether a fact has been tested few or many times.

Other collaborative approaches to create large knowledge bases usable by machines exist, including the Open Mind Common Sense Project (OMCS) [31] or Freebase[5]. Freebase shows some input methods that resemble games, such as: *Typewriter*[6] or *Genderizer*[7]. Answers taken from users in these interfaces are directly converted into statements (e.g., "... is female.") issued by the user and added to the knowledge base, taking them out of the list of items which lack information. In contrast to BetterRelations, such input methods typically do not contain any means of filtering (possibly intentional) disruptive user input and do not provide edge weights.

4 The Game

A straightforward approach to collect association strengths for Linked Data triples is this: First, we select a Linked Data resource of interest (e.g., `dbpedia:Facebook` or `dbpedia:Wiki`). We call this a *topic of interest* or simply *topic*. We then show randomly shuffled lists of all related triples to test persons and ask them to order the triples by decreasing importance. In the context of this work, given a topic, we define *related triples* to be the collection of (subject, predicate, object)-triples where the topic is the subject.[8]

The aforementioned approach suffers from the problem that the outcome of each of these experiments, which is a user centric ranking, is not only highly subjective, but sometimes even unstable for one person over time. In order to overcome difficulties for humans when sorting lengthy lists, we could ask for the atomic relative comparisons of two facts about one topic and then use an objective rating algorithm to generate an absolute ranking of the topic's related facts. This leads us to the idea behind BetterRelations.

4.1 BetterRelations

BetterRelations[9] is a symmetric two player output (decision) agreement game in terms of von Ahn and Dabbish's design principles for Games With A Purpose [11]:

A player starting to play the game is randomly matched with some other player for a predefined timespan (e.g., 2 minutes). In every round (see Figure 1)

[5] `http://www.freebase.com/`
[6] `http://typewriter.freebaseapps.com/`
[7] `http://genderizer.freebaseapps.com/`
[8] Extending the list by triples where the topic is the object (incoming links) typically imports a large number of unimportant facts for the topic (e.g., in Wikipedia and thus in DBpedia one would expect to learn about Facebook by visiting the page about it, not by reading through all the pages linking to its page).
[9] BetterRelations can be played online: `http://lodgames.kl.dfki.de`

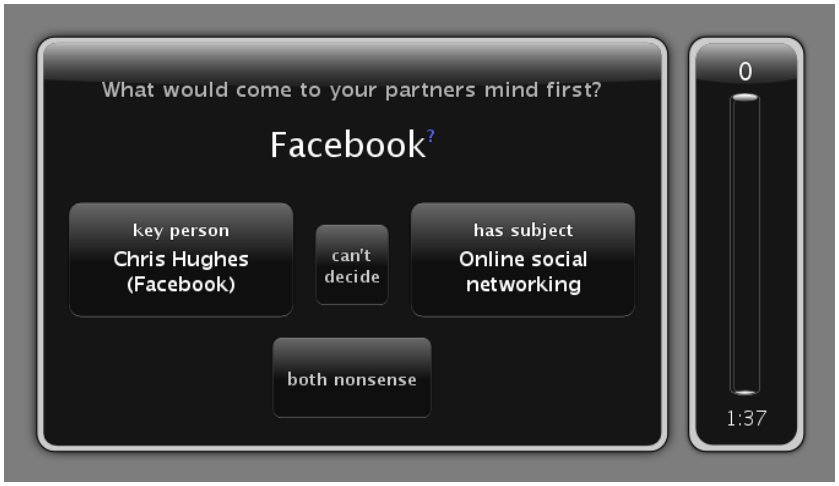

Fig. 1. In a game round, choosing phase.

both players are presented with a *topic*, which actually is a Linked Data resource's symbol (e.g., *Facebook*, the symbol for `dbpedia:Facebook`), and two *items*, which are symbolic forms of facts about the topic (e.g., *key person Chris Hughes (Facebook)* and *has subject Online social networking*). As in Matchin the facts are presented to the players in randomized order to counter easy cheating attempts.

Both players are asked to select the fact that their partner will have thought of first. In case a player does not know the topic, a quick info can be requested by clicking on the question mark appended to the topic. Doing so will internally mark the player's decision as influenced and the partner's as unvalidated. To decide, each player can either click on the more important fact's button or on two additional buttons in case the player can't decide between the alternatives or thinks that both alternatives are nonsense / noise.

As in Matchin, BetterRelations rewards agreements between both players with points and punishes disagreements without subtracting points, in order to increase game fun. The scoring function bases on the number of successive agreements in the current and preceding rounds: Players are rewarded with $0, 5, 10, 25, 50, 75, 90, 95, 98, 99, 99, 99, 100, \ldots, 100$ additional points for a streak of $0, 1, 2, 3, \ldots$ agreements. In contrast to Matchin (where the streak is reset to 0 on a mismatch), in BetterRelations a mismatch will only decrease the streak by 2 and does not reward the current round with additional points.

BetterRelations includes two more buttons: "can't decide" and "both nonsense" than Matchin. Hence, the scoring function was changed in order to counter easy cheating strategies such as always selecting the "can't decide" button. In terms of the scoring function the both middle buttons are the same button (it counts as an agreement if one player selects "can't decide" and the other "both nonsense") and an agreement on the middle buttons will not be rewarded with

additional point, but instead will sustain the accumulated streak. Furthermore, a player who requested a quick info will not be rewarded with points in the current round.

On the server side the game records a large amount of relative decisions between pairs of items, filtered by a partner and uses them to upgrade ratings in case of agreements. A both nonsense agreement will mark both items as nonsense and exclude them from future games. Generating an absolute ranking from such results can be compared to chess rating systems, where based on the outcomes of atomic competitions (player p_1 won against p_2), a global ranking is calculated, just that in this case it is not players competing, but facts [28]. In contrast to Matchin, BetterRelations uses a TrueSkill [32] based algorithm internally to update fact ratings after each agreement, selects next fact pairs for a topic in a way to minimize the overall needed amount of decisions and stops sorting lists with n facts after $n \cdot \log_2(n)$ updates, determined to be a good threshold by simulations.

After rewarding the players with points, the next round starts until the game runs out of time. The next topic is chosen by selecting the topic least often played by both players from a list of topics currently opened for playing, which is based on the topmost accessed Wikipedia articles. In the end, both players see a summary of their performance showing the amount of points gained in this game, the longest streak and their total game score in BetterRelations.

In case no partner can be found or the partner leaves the Game, BetterRelations also provides a single player mode, which will either replay rounds with unvalidated decisions or replay previous two player games if no unvalidated decisions are left. As the latter replays usually waste human decisions, the single player mode can also be configured to initiate two player games with a certain probability and fake the (dis-)agreement by chance, based on the player's historical rate of agreements. The results of such rounds again provide new unvalidated decisions used by other single players.

4.2 Game Data Acquisition and Preprocessing

As BetterRelations tries to rank multiple lists of triples related to one topic each, we first of all have to decide which topics we want to play. Topics should be well known to most players and be interesting, in order to receive valuable feedback and provide an entertaining game. Additionally each of the topics should have associated Linked Data triples. Hence, BetterRelations selects topics (Linked Data URIs) corresponding to the most often accessed Wikipedia pages[10], which include pages such as Wiki, United States, Facebook, Google. Every time the game needs a new game topic and its related triples (e.g., because an existing topic's facts were sorted), it loads the corresponding triples for the next topmost Wikipedia topic from a local DBpedia mirror, which also was pre-loaded with standard vocabularies such as `rdf`, `rdfs`, `foaf`.

[10] Stats aggregated from raw access logs, available at `http://dom.as/wikistats/`

As showing URIs to the end-users is of limited use, the users will always see `rdfs:labels` of such references. Hence, for each URI in the list of related triples of a topic, all English or non language tagged `rdfs:label`s are acquired. For URIs with multiple labels a best label is selected following a heuristic preferring language tagged literals and such which are similar to the URI's last part if still in doubt. Triples having the same labels are merged from a game's point of view and such with missing labels for predicate or object excluded from the game. We call this the *symbolic form* of a triple.

Finally labels and corresponding triples are excluded, which (due to long string length) don't fit into the game's window, end with suspicious file endings (e.g., .jpeg) or which have an object label equal to the topic's label ("Facebook label Facebook").

5 Evaluation

After the previous sections detailed the game's concept, data acquisition and preprocessing, we will now provide a detailed evaluation of the game itself and of the generated output.

5.1 The Game

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

Measurements and Estimates In the 18 day period from Jan. 12th until Jan. 30th, 2011, the game was played by 359 Users (re-identified by cookies if possible). In this timespan 1041 games were played, out of which 431 were two-player and 610 were single-player games.

The players played a total amount of $12K$ rounds submitting $14.7K$ decisions, out of which they selected $11.2K$ times an item, $2K$ times "can't decide" and $1.5K$ times "both nonsense". This led to an amount of $3.8K$ mismatches, $4.7K$ matches, including $3.8K$ item matches, and 840 non item matches.

The total amount of time all players together played the game was 42 hours (rounds without any decisions were not counted, they summed up to 5 hours, 46 minutes, e.g., idle tabs). With this, we can calculate the average time a decision takes to be 10.3 seconds. The *throughput*[11] of BetterRelations hence is 350 decisions per human hour of gaming. With the given numbers we can also find out the *average lifetime play*, so the time an average player plays the game, to be about 7 minutes. Multiplication of both numbers gives us an *expected contribution* of 41 decisions per human.

[11] For a definition of throughput, average lifetime play and expected contribution also see [11].

Repeating the above for matches instead of decisions yields a *throughput* of 112 matches per human hour of gaming, and an *expected contribution* of 13 matches per human.

Knowing that the top 1000 Wikipedia topics contain $56K$ game items, and taking into account the observed nonsense ratio of $\frac{1}{10}$, we can estimate that in order to sort the facts known about the top 1000 Wikipedia topic, we would need $313K$ matches. In terms of players, this means that with the current implementation and we would need about $23.9K$ players to sort the top 1000 Wikipedia topics, i.e., 24 players per topic.

Questionnaire Aside from these measurements and estimates, we wanted to know if the game was fun and wanted to collect feedback for possible future enhancements. For this, an online questionnaire survey was conducted among players of the game. The questionnaire was completed by 35 participants, mainly German (32) computer science students (23) or researchers (8), 31 male and 4 female.

Apart from background questions, the questionnaire consisted of a series of 5-point Likert scale items that are listed in Table 1 and comment fields asking what the participants liked, disliked and what they were missing. The summarized results in Table 1 show that most of the players were between 21 and 33 years old and had played online games before.

The main result from the conducted survey is that the game in its current version is of limited fun and that the majority of people do not plan to play it again. From the collected numerical data we can also see that in average the participants did not know all the topics and knew even less of the game items. At the same time most of the participants agreed that the game contained too much nonsense and too many irrelevant facts.

Apart from these numerical results, a view of the collected comments yields many common aspects. Many users mentioned that they liked the idea of creating a game to collect scientific data and the design of the game. In accordance with the numerical results, most users mentioned that they disliked the high amount of nonsense, consisting for example of unknown or cryptic abbreviations. Many

Statement	μ	σ
The gaming principle was easy	4.43	0.77
I knew all topics	3.11	1.04
I knew all items	2.54	0.91
Too much nonsense	3.68	1.23
Too many irrelevant facts	3.57	1.13
The game was fun	2.66	1.04
I will play it again	2.34	1.29
Played online games before	4.20	1.33
Age	27.68	6.76

Table 1. Results of an online survey answered by 35 game players. Except from *Age* users could select answers from a 5-point Likert scale: 1 (Strongly disagree), 2 (Disagree), 3 (Neutral), 4 (Agree), 5 (Strongly agree).

rank	sum	ns	predicate	object
0.0	14.0		has subject	Wikis
1.0	26.0		has subject	Social information processing
2.0	28.0		has subject	Self-organization
3.0	30.0		has subject	Hypertext
4.0	42.5		has subject	Human-computer interaction
5.0	47.5		has subject	Internet history
6.5	74.0	x	Jahr	2007
6.5	74.0	x	tag	10

rating	ns	predicate	object
19.41		has subject	Self-organization
18.33		has subject	Social information processing
15.78		has subject	Human-computer interaction
9.15		has subject	Wikis
5.34		has subject	Hypertext
-1.63		has subject	Internet history
4.24	x	Jahr	2007
4.21	x	tag	10

Table 2. Example: Gold Standard (left) and Game Output (right) lists for topic *Wiki*. In this case *predicate* and *object* are the symbolic forms of the corresponding triples from DBpedia.

participants also mentioned that they disliked the formatting of dates and often were confronted with facts they did not know anything about. Some of the participants also disliked the waiting period in the beginning of the game and complained about the mixture of German and English facts.

Many of the participants also mentioned that they were missing a button "I don't know any of these" or an initial selection of own interests, so they were not asked things they did not know that often. Many users requested a way to know who they were playing with and even suggested to make it possible to explicitly select a partner to play with. Some of the participants also suggested showing a highscore screen at the end of the game and including user accounts to save their own score and a recap phase after the game listing the questions and selected answers, showing their outcomes and providing more exploratory features.

5.2 Output Quality

Besides evaluating the game itself, the quality of generated results is of special interest in this work. As mentioned in the previous sections, the game calculates rating scores for the facts in each of the topics' related triples lists. The rating score can be used to order each of these lists, generating ordered output rankings. In the testing period, the game completed the generation of 12 such lists ordered by importance ratings.

In order to assess the quality of these lists, a Gold Standard list was generated for each of these 12 topics.

The Gold Standard lists were generated by a test group consisting of 11 people who had played the game before. Each candidate was asked to manually sort each of the 12 randomly shuffled lists of related facts by importance after excluding facts that the candidate identified as nonsense. For each of the topics the manually sorted lists were aggregated by summing up the ranks for each fact and afterwards sorting ascending by rank sum, forming the Gold Standard list. In this process nonsense facts were appended to each list's end and given a rank according to the barycenter of all nonsense items in that list. In the aggregated list a fact is said to be nonsense if the majority of test persons considered it as nonsense. An example of such a manually generated Gold Standard list can be seen in Table 2 (left).

Once a Gold Standard list is generated, the Mean Squared Errors (MSE) can be calculated for each of the individual manually generated ranked lists. The

MSE is computed as the average sum of squared rank differences of each fact in the list and can be seen as blue histogram bars in Figure 2.

Calculating the average of these MSEs (so the average error an individual human makes when compared to the Gold Standard) and the deviation thereof (seen as red dashed and dotted lines in Figure 2), we can compare the human results with the game's result (which is shown as green vertical bar).

Even though the statistics in Figure 2 should be handled carefully because of the low sample size, we can observe that the game's result are within the 1σ interval of manually created lists in 9 out of 12 cases. In 3 cases (*ISBN*, *Halloween* and *Harry Potter*), the game results are a bit worse than those generated by our test group, in 6 cases better than an average individual human.

After this description of the game's evaluation and its generated output rankings, the results will be discussed in the next section.

6 Discussion

One of the main concerns when designing BetterRelations was the desired high quality of its generated output ratings. This task was considerably complicated by the high amount of noise which occurs in the Linked Data triples acquired mainly from DBpedia. Nevertheless, the results of the evaluation show that the game's outputs are about as good as those of humans in 9 out of 12 cases and even better than an average human in half of the cases.

While a 75 % success rate is satisfactory, we were also interested in the problems of the 3 remaining lists, which correspond to topics *Harry Potter*, *ISBN* and *Halloween*.

An investigation of the topic *Harry Potter* revealed that while the game item ((p,o) pair) "image caption · Complete set of the seven books" was marked as nonsense in the Gold Standard list, it is ranked as top item by the game, indicating that many players preferred it over other game items. A possible explanation for this is that players of the game had limited time for their decisions and maybe overlooked the erroneous predicate label in a rush, and their association was likely dominated by the more prominent and very useful object label. At the same time, the participants of our Gold Standard test group had no time restriction to select items they regarded as nonsense. This single misplaced item accounts for a large amount of the game's calculated MSE ($\approx$ 15), probably making the result much worse than it is. In the results of *Halloween* we noticed that the facts "has subject · Irish folklore", "has subject · Irish culture" and "has subject · Scottish folklore" were marked to be nonsense in the game results. Nevertheless, these game items receive suspiciously high ratings for nonsense items which, if they were not reordered to the end of the list as done in each of the human-generated lists, would have caused a much lower MSE value. Hence, we suggest to trigger a review in cases of such discrepancies between current rating and nonsense flagging in future versions. In the third of these lists for topic *ISBN*, we could not identify an obvious reason for the discrepancy.

But even when taking these considerations into account, we are confident that the game—already in its current version—generates good output ratings from pairwise comparisons of items. Nevertheless, it remains part of future work to conduct a survey showing the game outcomes to a test group and asking for immediate feedback about the generated ranking.

Aside from the high quality of the generated ratings, we also evaluated the game itself. The the questionnaire reveals that game principle was easy and straightforward and the majority of topics was known. However, problems related to fun and replay-ability were also mentioned. An investigation of the given comments revealed that the primarily impairing factors were the presence of many cryptic abbreviations, *strange* formatting of numbers and dates, and the mixture of English and German facts. Since improvements of the game's fun factor would further decrease the amount of 24 players needed to sort the facts known about one Wikipedia topic, we performed an analysis on the reported problems. It turned out that many of them originated from errors in the DBpedia 3.5.1 dataset, e.g., German labels which had missing or incorrect language tags, and have been resolved in the more recent DBpedia 3.6. We expect that upcoming releases of the DBpedia dataset will address even more of these problems, as the extraction mappings are improving. Such an enhanced quality of the underlying datasets has the dual effect of reducing the amount of (erroneous) triples to sort and at the same time increasing the fun of the game.

7 Conclusion & Outlook

In this chapter we presented a survey of existing approaches to rank Linked Data and after identifying the need for a collection of Linked Data rated by humans, presented a game called BetterRelations as well as a detailed evaluation of our first implementation.

Our evaluation shows very promising results in terms of the desired quality of the generated collection of importance ratings. We believe that this approach can be used to successfully sort Linked Data triples. While the low average lifetime play indicates a problem with the game's motivation, this appears to be mainly caused by the high amount of noise in the underlying Linked Data triples. As even slight improvements of the average lifetime play could already drastically reduce the number of players needed to sort the facts known about a popular Wikipedia topic, our future work will focus on methods to detect noise and the way how the game deals with it. We also plan to provide the game's output (ranked lists with rating scores) as Linked Data, allowing others to rank result sets of queries by importance for humans, and implement other ways to increase the player's fun, such as user accounts and high scores.

This work was financed in part by the University of Kaiserslautern PhD scholarship program and the BMBF project Perspecting (Grant 01IW08002).

References

1. Berners-Lee, T., Hendler, J., Lassila, O.: The Semantic Web: A new form of Web content that is meaningful to computers will unleash a revolution of new possibilities. Scientific American **284**(5) (May 2001) 34–43
2. Bizer, C., Heath, T., Berners-Lee, T.: Linked Data - The Story So Far. International Journal on Semantic Web and Information Systems **5**(3) (January 2009) 1–22
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 World Wide Web **7**(3) (September 2009) 154–165
4. van Elst, L., Abecker, A.: Ontologies for information management: balancing formality, stability, and sharing scope. Expert Systems with Applications **23**(4) (November 2002) 357–366
5. Crestani, F.: Application of Spreading Activation Techniques in Information Retrieval. Artificial Intelligence Review **11**(6) (January 1997) 453–482
6. Schumacher, K., Sintek, M., Sauermann, L.: Combining Fact and Document Retrieval with Spreading Activation for Semantic Desktop Search. In Bechhofer, S., Hauswirth, M., Hoffman, J., Koubarakis, M., eds.: ESWC 2008. Volume LNCS 5021., Tenerife, Spain, Springer Berlin / Heidelberg (2008) 569–583
7. Brin, S., Page, L.: The Anatomy of a Large-Scale Hypertextual Web Search Engine. Computer Networks and ISDN Systems **30**(1-7) (April 1998) 107–117
8. Kleinberg, J.M.: Authoritative sources in a hyperlinked environment. Journal of the ACM **46**(5) (1999) 604–632
9. Hees, J., Roth-Berghofer, T., Dengel, A.: Linked Data Games: Simulating Human Association with Linked Data. In: LWA 2010, Kassel, Germany (2010)
10. Hees, J., Roth-berghofer, T., Biedert, R., Adrian, B., Dengel, A.: BetterRelations: Using a Game to Rate Linked Data Triples. In: KI 2011: Advances in Artificial Intelligence, Berlin, Springer Berlin / Heidelberg (2011) 5
11. von Ahn, L., Dabbish, L.: Designing games with a purpose. Communications of the ACM **51**(8) (August 2008) 58–67
12. Balmin, A., Hristidis, V., Papakonstantinou, Y.: ObjectRank: Authority-Based Keyword Search in Databases. In: Proc. of the 13th International Conference on Very Large Data Bases, VLDB Endowment. (2004) 564–575
13. Ding, L., Pan, R., Finin, T., Joshi, A., Peng, Y., Kolari, P.: Finding and Ranking Knowledge on the Semantic Web. In Gil, Y., Motta, E., Benjamins, V.R., Musen, M.A., eds.: Proc. of the ISWC 2005, Galway, Ireland, Springer Berlin / Heidelberg (2005) 156–170
14. Harth, A., Kinsella, S., Decker, S.: Using Naming Authority to Rank Data and Ontologies for Web Search. In Bernstein, A., Karger, D.R., Heath, T., Feigenbaum, L., Maynard, D., Motta, E., Thirunarayan, K., eds.: Proc. of the ISWC 2009. Volume 2., Chantilly, VA, USA, Springer Berlin / Heidelberg (2009) 277–292
15. Delbru, R., Toupikov, N., Catasta, M., Tummarello, G., Decker, S.: Hierarchical Link Analysis for Ranking Web Data. In Aroyo, L., Antoniou, G., Hyvönen, E., ten Teije, A., Stuckenschmidt, H., Cabral, L., Tudorache, T., eds.: Proc. of the ESWC 2010, Heraklion, Crete, Greece, Springer Berlin / Heidelberg (2010) 225–239
16. Tummarello, G., Delbru, R., Oren, E.: Sindice.com: Weaving the Open Linked Data. In: Proc. of the ISWC 2007, Springer Berlin / Heidelberg (2007) 552–565

17. Franz, T., Schultz, A., Sizov, S., Staab, S.: TripleRank: Ranking Semantic Web Data by Tensor Decomposition. In Bernstein, A., Karger, D.R., Heath, T., Feigenbaum, L., Maynard, D., Motta, E., Thirunarayan, K., eds.: Proc. of the ISWC 2009, Chantilly, VA, USA, Springer Berlin / Heidelberg (2009) 213–228
18. Anyanwu, K., Maduko, A., Sheth, A.P.: SemRank: Ranking Complex Relationship Search Results on the Semantic Web. In: Proc. of the WWW 2005, Chiba, Japan (2005)
19. Ell, B., Vrandečić, D., Simperl, E.: Labels in the Web of Data. In: Proc. of the ISWC 2011, Bonn, Germany, Springer Berlin / Heidelberg (2011) 162–176
20. Fellbaum, C., ed.: WordNet: An Electronic Lexical Database. MIT Press, Cambridge, MA, USA (1998)
21. Budanitsky, A., Hirst, G.: Evaluating WordNet-based Measures of Lexical Semantic Relatedness. Computational Linguistics **32**(1) (March 2006) 13–47
22. Strube, M., Ponzetto, S.P.: WikiRelate! Computing Semantic Relatedness Using Wikipedia. In: Proc. of the AAAI 2006. Number February, Boston, MA, USA, AAAI Press (2006) 1419–1424
23. Cilibrasi, R.L., Vitányi, P.M.B.: The Google Similarity Distance. IEEE Trans. Knowledge and Data Engineering **19**(3) (2007) 370–383
24. Cattuto, C., Benz, D., Hotho, A., Stumme, G.: Semantic Grounding of Tag Relatedness in Social Bookmarking Systems. In: Proc. of the ISWC 2008, Karlsruhe, Germany (2008) 615–631
25. Agirre, E., Alfonseca, E., Hall, K., Kravalova, J., Pasça, M., Soroa, A.: A Study on Similarity and Relatedness Using Distributional and WordNet-based Approaches. In: Proc. of the NAACL 2009. Number June, Boulder, Colorado, US, Association for Computational Linguistics (2009) 19–27
26. Mirizzi, R., Ragone, A., Noia, T.D., Sciascio, E.D.: Ranking the Linked Data: The Case of DBpedia. In Benatallah, B., Casati, F., Kappel, G., Rossi, G., eds.: Proc. of the ICWE 2010, Vienna, Austria, Springer Berlin / Heidelberg (2010) 337–354
27. Waitelonis, J., Sack, H.: Towards Exploratory Video Search Using Linked Data. In: Proc. of the IEEE International Symposium on Multimedia (ISM) 2009, San Diego, CA, USA, IEEE (2009) 540–545
28. Hacker, S., von Ahn, L.: Matchin: Eliciting User Preferences with an Online Game. In: Proc. of the SIGCHI Conference on Human Factors in Computing Systems, Boston, MA, USA, ACM (2009) 1207–1216
29. Siorpaes, K., Hepp, M.: OntoGame: Towards Overcoming the Incentive Bottleneck in Ontology Building. In: Proc. of the 3rd International IFIP Workshop On Semantic Web & Web Semantics (SWWS), OTM-WS 2007, Part II. LNCS, Vilamoura, Portugal, Springer, Heidelberg (2007) 1222–1232
30. Kny, E., Kölle, S., Töpper, G., Wittmers, E.: WhoKnows? (October 2010)
31. Singh, P.: The Open Mind Common Sense Project. KurzweilAI.net (January 2002)
32. Herbrich, R., Minka, T., Graepel, T.: TrueSkill(TM): A Bayesian Skill Rating System. In Schölkopf, B., Platt, J.Hoffmann, T., eds.: Advances in Neural Information Processing Systems. Volume 19., Cambridge, MA, USA, MIT Press (2007) 569–576

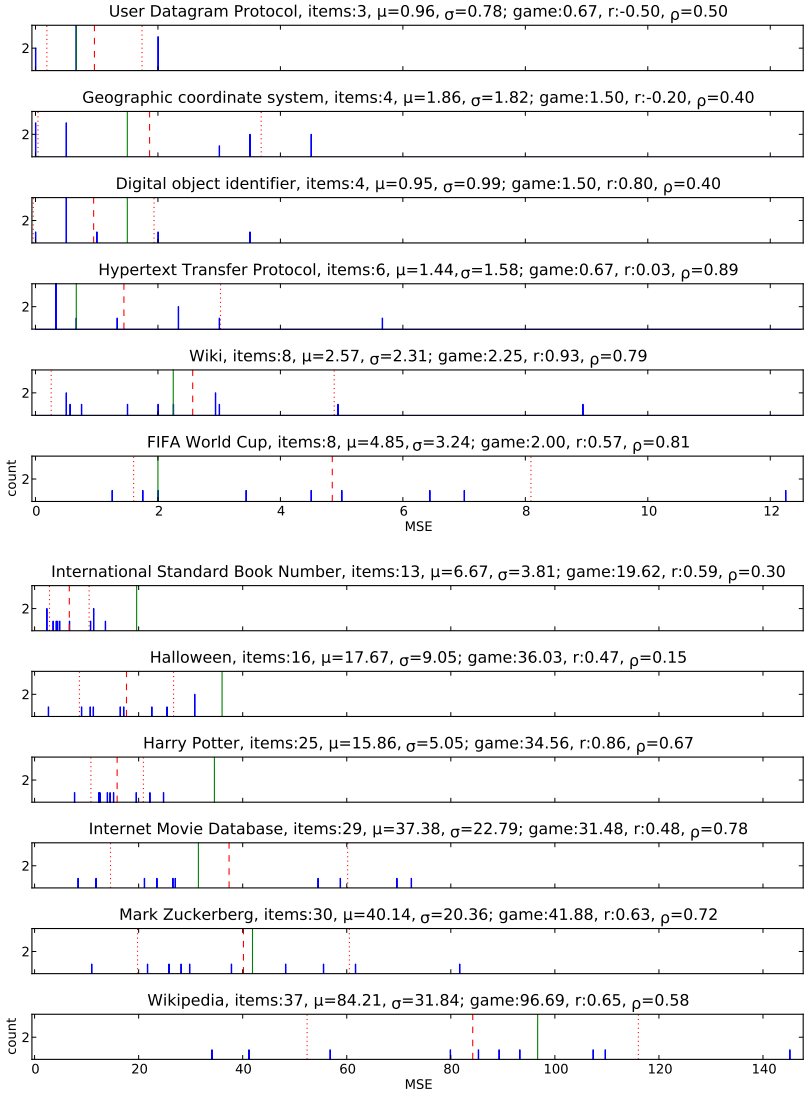

Fig. 2. Comparison of Gold Standard and game output on 12 topics' item lists. Blue histogram bars show the MSEs of each manually generated lists, their mean μ is shown as a red dashed line, their standard deviation σ as red dotted lines. The game's MSE error is shown as a green line. The titles also include the Pearson product-moment correlation coefficient r and Spearman's rank correlation coefficient ρ of the Gold Standard List and the game's output.