published as: N. Ludwig, J. Waitelonis, M. Knuth, H. Sack: WhoKnows? – Evaluating Linked Data Heuristics with a Quiz that Cleans Up DBpedia. In Proc. of the 8th Extended Semantic Web Conference (ESWC), Poster-Session, Springer LNCS 6643, 2011. (ISBN 978-3-642-21033-4)

WhoKnows? - Evaluating Linked Data Heuristics with a Quiz that Cleans Up DBpedia

Nadine Ludwig, Jörg Waitelonis, Magnus Knuth, and Harald Sack

Hasso-Plattner-Institut für Softwaresystemtechnik GmbH,
Prof.-Dr.-Helmert-Str. 2–3, 14482 Potsdam, Germany
{nadine.ludwig,joerg.waitelonis,magnus.knuth,harald.sack}@hpi.uni-potsdam.de
http://www.hpi.uni-potsdam.de

1 Introduction

With the Linking Open Data (LOD) [1] initiative large resources of publicly available structured data from various domains have been turned into interlinked RDF(S) facts to constitute the so-called "Web of Data". But, this Web of Data is by no means a perfect world of consistent and valid facts. One of the major problems that we experienced working with LOD was the lack of any relevance ranking for the represented facts. Therefore, we have developed fact ranking heuristics based on statistical, graph theoretic, and liguistic data [2]. On the other hand there is also no sound evaluation for property ranking heuristics. Thus, a qualitative evaluation had to be realized.

2 *WhoKnows?* - and how it is played

WhoKnows? is designed as game for in-between times[1]. The game is based on the principle to present questions to the user that have been generated out of DBpedia[3] RDF triples. A question originated from a RDF triple is composed by turning the order of a triple upside down: 'object is the property of subject'. In addition, false answers (subjects) are randomly selected meeting the requirement to hold the same property but are not related to the object the question is asking for. E. g., the triple 'dbp:Chile dbpprop:language dbp:Spanish_language .' added by wrong answers 'Iraq', 'Brazil', 'Italy' is turned into the question 'Spanish is the language of Chile, Iraq, Brazil or Italy?'.

After selecting the answer, the user receives immediate feedback about the correctness of her choice by showing happy resp. unhappy emotions of the game mascots. If the user has the opinion a question is somehow odd or strange a 'Dislike'-button enables to mark this question as an potentially inconsistent, ambiguous, confusing, or conflicting question.

[1] *WhoKnows?* can be played as stand-alone web game here: http://www.yovisto.com/whoknows

3 Evaluation

WhoKnows? was originally designed to evaluate the ranking heuristics proposed in [2]. By the addition of the 'Dislike'-button potential discrepancies according to the 'real' world within DBpedia can be revealed. These problems are caused either during automated data extraction from Wikipedia, or by false statements within Wikipedia articles. Typically, the discrepancies occur, if Wikipedia info boxes are not well structured, as, e. g. if several different entities are put in a single infobox field.

Our evaluation of the property ranking underlies the assumption that, the more often a question is answered correctly, the more likely it is that the fact seems to be of importance. Whenever a distinct RDF property for different subjects and objects occurs frequently among the correct answers, it can be regarded as an important RDF property. *WhoKnows?* keeps track for every triple how often it has been played and how often it has been matched correctly. Using this data we are able to rank DBpedia properties according to their ratio, how often a question with this property has been answered correctly. According to this heuristic we can recommend a selection and ordering of likely relevant RDF properties for an entity. As, e. g., for Greece[2] the properties capital, largestCity, anthem, language, leaderName, and governmentType are considered to be important.

For the detection of incorrect facts within DBpedia data we had to analyze all games that have been marked by using the 'Dislike'-button. The 'Dislike'-button was used 289 times. In some cases users have had the opinion that the question was too trivial or too simple. The remaining triples marked as with the dislike button refer to real discrepancies within DBpedia.

At present, the game has been played 523 times by 94 different users. In total 2,525 distinct triples have been played. Overall 12,663 rounds have been performed of which 9,494 have been answered correctly.

References

1. Bizer, C., Heath, T., Idehen, K., Berners-Lee, T.: Linked data on the web. In: Proc. of the 17th Int. Conf. on World Wide Web, ACM (2008) 1265–1266
2. Waitelonis, J., Sack, H.: Towards exploratory video search using linked data. Multimedia Tools and Applications (2011) 1–28 10.1007/s11042-011-0733-1.
3. Auer, S., Bizer, C., Kobilarov, G., Lehmann, J., Cyganiak, R., Ives, Z.: Dbpedia: A nucleus for a web of open data. In: Proceedings of 6th International Semantic Web Conference, 2nd Asian Semantic Web Conference (ISWC+ASWC 2007). (November 2008) 722–735

[2] http://dbpedia.org/resource/Greece