Web Semantics: Science, Services and Agents on the World Wide Web 35 (2015) 85–101

Contents lists available at ScienceDirect

Web Semantics: Science, Services and Agents on the World Wide Web

journal homepage: www.elsevier.com/locate/websem

DeFacto—Temporal and multilingual Deep Fact Validation

Daniel Gerber, Diego Esteves *, Jens Lehmann *, Lorenz Bühmann, Ricardo Usbeck, Axel-Cyrille Ngonga Ngomo, René Speck

University Leipzig, Institute for Computer Science, Agile Knowledge Engineering and Semantic Web (AKSW), D-04009 Leipzig, Germany

ARTICLE INFO

Article history:
Received 16 February 2015
Received in revised form
17 June 2015
Accepted 3 August 2015
Available online 20 August 2015

Keywords:
Web of Data
Fact validation
NLP
Provenance

ABSTRACT

One of the main tasks when creating and maintaining knowledge bases is to validate facts and provide sources for them in order to ensure correctness and traceability of the provided knowledge. So far, this task is often addressed by human curators in a three-step process: issuing appropriate keyword queries for the statement to check using standard search engines, retrieving potentially relevant documents and screening those documents for relevant content. The drawbacks of this process are manifold. Most importantly, it is very time-consuming as the experts have to carry out several search processes and must often read several documents. In this article, we present DeFacto (Deep Fact Validation)—an algorithm able to validate facts by finding trustworthy sources for them on the Web. DeFacto aims to provide an effective way of validating facts by supplying the user with relevant excerpts of web pages as well as useful additional information including a score for the confidence DeFacto has in the correctness of the input fact. To achieve this goal, DeFacto collects and combines evidence from web pages written in several languages. In addition, DeFacto provides support for facts with a temporal scope, i.e., it can estimate in which time frame a fact was valid. Given that the automatic evaluation of facts has not been paid much attention to so far, generic benchmarks for evaluating these frameworks were not previously available. We thus also present a generic evaluation framework for fact checking and make it publicly available.

1. Introduction

The past decades have been marked by a change from an industrial society to an information and knowledge society. This change is particularly due to the uptake of the World Wide Web. Creating and managing knowledge successfully has been a key to success in various communities worldwide. Therefore, the quality of knowledge is of high importance. One aspect of knowledge quality is provenance [1]. In particular, the sources for facts should be well documented since this provides several benefits such as a better detection of errors, decisions based on the trustworthiness of sources etc. While provenance is an important aspect of data quality [2], to date only few knowledge bases actually provide provenance information. For instance, less than 10% of the more than 708.26 million RDF documents indexed by Sindice[1] contain metadata such as creator, creation date, source, modified or contributor.[2] This lack of provenance information makes the validation of the facts in such knowledge bases utterly tedious. In addition, it hinders the adoption of such data in business applications as the data is not trusted [2]. The main contribution of this paper is the provision of a fact validation approach and tool which can make use of one of the largest sources of information: the Web.

More specifically, our system DeFacto (Deep Fact Validation) implements algorithms for validating RDF triples by finding confirming sources for them on the Web.[3] It takes a statement as input (e.g., the one shown in Listing 1, page 13) and then tries to find evidence for the validity of that statement by searching for textual information in the Web. To this end, our approach combines two strategies by searching for textual occurrences of parts of the statements as well as trying to find web pages which contain the input statement expressed in natural language. DeFacto was conceived to exploit the multilinguality of the Web, as almost half of the content of the Web is written in a language other than English[4] (see

* Corresponding authors.
E-mail addresses: dgerber@informatik.uni-leipzig.de (D. Gerber), esteves@informatik.uni-leipzig.de (D. Esteves), lehmann@informatik.uni-leipzig.de (J. Lehmann), buehmann@informatik.uni-leipzig.de (L. Bühmann), usbeck@informatik.uni-leipzig.de (R. Usbeck), ngonga@informatik.uni-leipzig.de (A.-C. Ngonga Ngomo), speck@informatik.uni-leipzig.de (R. Speck).

[1] http://www.sindice.com.

[2] Data retrieved on February 13, 2015.

[3] Please note that we use *fact* as a synonym for a *RDF triple*.

[4] 45% non-English web pages according to http://w3techs.com/technologies/overview/content_language/all.

http://dx.doi.org/10.1016/j.websem.2015.08.001
1570-8268/© 2015 Elsevier B.V. All rights reserved.

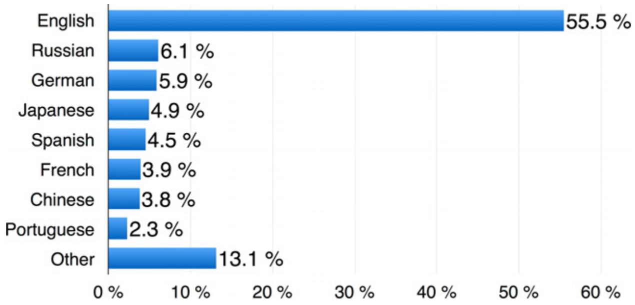

Fig. 1. Usage of content languages for web pages. (W3Techs.com, 21 November 2013.)

Fig. 1). To this end, our approach abstracts from a specific language and can combine evidence from multiple languages—currently English, German and French.

The output of our approach is a confidence score for the input statement as well as a set of excerpts of relevant web pages which allows the user to manually judge the presented evidence. Apart from the general confidence score, DeFacto also provides support for detecting the temporal scope of facts, i.e., estimates in which timeframe a fact is or was valid.

DeFacto has three major use cases: (1) Given an existing true statement, it can be used to find provenance information for it. For instance, the WikiData project[5] aims to create a collection of facts, in which sources should be provided for each fact. DeFacto could help to achieve this task. (2) It can check whether a statement is likely to be true and provide the user with a corresponding confidence score as well as evidence for the score assigned to the statement. (3) Given a fact, DeFacto can determine and present evidence for the time interval within which the said fact is to be considered valid. Our main contributions are thus as follows:

- We present an open-source approach that allows checking whether a web page confirms a fact, i.e., an RDF triple.
- We discuss an adaptation of existing approaches for determining indicators for trustworthiness of a web page.
- We present an automated approach to enhancing knowledge bases with RDF provenance data at triple level as well as.
- We provide a running prototype of DeFacto, the first system able to provide useful confidence values for an input RDF triple given the Web as background text corpus.

This article is an extension of the initial description of DeFacto in [3]. The main additions are as follows:

- A temporal extension detecting temporal scope of facts based on text understanding via pattern and frequency analysis.
- An extensive study of effect of the novel multilingual support in DeFacto, e.g., through the integration of search queries and temporal patterns in several languages.
- A freely available and full-fledged benchmark for fact validation which includes temporal scopes.

The rest of this paper is structured as follows: We give an overview of the state of the art of relevant scientific areas in Section 2. This part is followed by a description of the overall approach in a nutshell in Section 3. We show how we extended the BOA framework to enable it to detect facts contained in textual descriptions on web pages in Section 4. In Section 5, we describe how we calculate and include the trustworthiness of web pages into the DeFacto analysis. Section 6 combines the results from the previous chapters and describes the mathematical features we use to compute the confidence for a particular input fact. Subsequently, we describe the temporal extension of DeFacto in Section 7 and provide an overview of the FactBench benchmark in Section 8. We provide a discussion of the evaluation results in Section 9. Finally, we conclude in Section 10 and give pointers to future work.

2. Related work

The work presented in this paper is related to five main areas of research: Fact checking as known from NLP, the representation of provenance information in the Web of Data, temporal analysis, relation extraction and named entity disambiguation (also called entity linking).

2.1. Fact checking

Fact checking is a relatively new research area which focuses on computing which subset of a given set of statements can be trusted [4]. Several approaches have been developed to achieve this goal. Nakamura et al. [5] developed a prototype for enhancing the search results provided by a search engine based on trustworthiness analysis for those results. To this end, they conducted a survey in order to determine the frequency at which the users accesses search engines and how much they trust the content and ranking of search results. They defined several criteria for trustworthiness calculation of search results returned by the search engine, such as topic majority. We adapted their approach for DeFacto and included it as one of the features for our machine learning techniques. Another fact-finding approach is that presented in [6]. Here, the idea is to create a 3-partite network of web pages, facts and objects and apply a propagation algorithm to compute weights for facts as well as web pages. These weights can then be used to determine the degree to which a fact contained in a set of web pages can be trusted. Pasternack and Roth [7,8] present a generalized approach for computing the trustworthiness of web pages. To achieve this goal, the authors rely on a graph-based model similar to hubs and authorities [9]. This model allows computing the trustworthiness of facts and web pages by generating a *k*-partite network of pages and facts and propagating trustworthiness information across it. The approach returns a score for the trustworthiness of each fact. Moreover, the generalized fact-finding model that they present allows expressing other fact-finding algorithms such as TruthFinder [6], AccuVote [10] and 3-Estimates [11] within the same framework. The use of trustworthiness and uncertainty information on RDF data has been the subject of recent research (see e.g., [12,13]). Moreover, approaches such as random walks [14] have been used to measure the trustworthiness of graph data based on the topology of the underlying graph. Our approach differs from previous fact finding works as it focuses on validating the trustworthiness of RDF triples (and not that of facts expressed in natural language) against the Web (in contrast to approaches that rely on the RDF graph only). In addition, it can deal with the broad spectrum of relations found on the Data Web.

2.2. Provenance

The problem of data provenance is an issue of central importance for the uptake of the Web of Data. While data extracted by the means of tools such as Hazy[6] and KnowItAll[7] can be easily mapped to primary provenance information, most knowledge sources were extracted from non-textual source and are more difficult to link

[5] http://www.wikidata.org.

[6] http://hazy.cs.wisc.edu/hazy/.

[7] http://www.cs.washington.edu/research/knowitall/.