Lang Resources & Evaluation (2013) 47:1163–1190
DOI 10.1007/s10579-013-9232-5

ORIGINAL PAPER

WHAD: Wikipedia historical attributes data

Historical structured data extraction and vandalism detection from the Wikipedia edit history

Enrique Alfonseca · Guillermo Garrido · Jean-Yves Delort · Anselmo Peñas

Published online: 28 May 2013

Abstract This paper describes the generation of temporally anchored infobox attribute data from the Wikipedia history of revisions. By mining (attribute, value) pairs from the revision history of the English Wikipedia we are able to collect a comprehensive knowledge base that contains data on how attributes change over time. When dealing with the Wikipedia edit history, vandalic and erroneous edits are a concern for data quality. We present a study of vandalism identification in Wikipedia edits that uses only features from the infoboxes, and show that we can obtain, on this dataset, an accuracy comparable to a state-of-the-art vandalism identification method that is based on the whole article. Finally, we discuss different characteristics of the extracted dataset, which we make available for further study.

Keywords Wikipedia · Infobox · Attributes · Temporal data

This work was partially done while the second author was visiting Google Switzerland GmbH.

E. Alfonseca (✉) · J.-Y. Delort
Google Research Zurich, Zurich, Switzerland
e-mail: ealfonseca@google.com

J.-Y. Delort
e-mail: jydelort@google.com

G. Garrido · A. Peñas
NLP & IR Group, UNED, Madrid, Spain

G. Garrido
e-mail: ggarrido@lsi.uned.es

A. Peñas
e-mail: anselmo@lsi.uned.es

1 Introduction

Wikipedia, the world's largest online encyclopedia, offers free access to millions of articles, with the goal of being a comprehensive and up-to-date reference work.[1] Aside from its value as a general-purpose encyclopedia, Wikipedia has also become one of the most widely used resources to acquire, either automatically or semi-automatically, knowledge bases of structured data. Much research has been devoted to automatically building lexical resources, taxonomies, parallel corpora and structured knowledge from it.

Many Wikipedia entries contain so-called *infoboxes*: tabular information encoded as (attribute, value) pairs that summarize key information about a given article. It has been reported that roughly 30 % of the articles in the English Wikipedia contain an infobox (Lange et al. 2010). Parsing infoboxes has yielded useful knowledge bases such as DBPedia (Auer and Lehmann 2007; Auer et al. 2007) and Freebase (Bollacker et al. 2008). Data extracted automatically from infoboxes has been applied to various NLP tasks such as document summarization (Ye et al. 2009; Xu et al. 2010) and relation extraction (Wu and Weld 2007, 2010; Hoffmann et al. 2010). Infoboxes have been successfully used for *distant supervision*, i.e. using data obtained from infoboxes to semi-automatically annotate a dataset that can be used in training a supervised machine learning (ML) algorithm (Mintz et al. 2009; Hoffmann et al. 2010).

Every editable page in Wikipedia has an associated page history, where users can view past versions and, if necessary, revert the current state of the entry to one of them. By now Wikipedia has accumulated a wealth of historical information about the last decade, encoded in its revision history. To the best of our knowledge, existing work using infoboxes to extract lexical and relational knowledge bases only uses *snapshot* versions of Wikipedia, containing a single (frozen) version for each article. In contrast, historic values of attributes in infoboxes can also be exploited; for example, for distant supervision in temporally-aware information extraction systems, with the goal of extracting values for attributes that change over time (Zhang et al. 2008; Wang et al. 2010). Probably, one of the reasons why the revision history has not been used before is the large size and the format of the dataset; these features render its processing a very difficult task.

In this paper, we describe the collection of a large, structured dataset of temporally anchored attributes and values, obtained from the revision history of Wikipedia, including the different steps involved in its construction, and analyze several properties of the obtained data. We call the generated dataset WHAD (Wikipedia historical attributes data).

We are releasing this dataset through Wikimedia Deutschland,[2] which proposed to distribute it from its Wikimedia Toolserver download page,[3] under the Creative Commons license that covers Wikipedia.

[1] As of March 2012, there were more than 85,000 active contributors working on more than 21,000,000 articles in more than 280 languages. The English Wikipedia contained more than 3.9 million articles. Ref: http://en.wikipedia.org/wiki/Wikipedia:About.

[2] Wikimedia Deutschland—Gesellschaft zur Förderung Freien Wissens e.V.

[3] Wikimedia Toolserver, http://toolserver.org. The dataset is available for download at http://toolserver.org/~RENDER/toolkit/downloads/. Additional information can be obtained at http://alfonseca.org/eng/research/whad.html.

When the edit history is used as data source, erroneous and vandalic edits are a concern for data quality. We present a study of vandalism identification in Wikipedia edits that uses only features from the infoboxes, and show that, for the subset of Wikipedia articles that contain infoboxes, we attain results comparable to a state-of-the-art vandalism identification method that is based on the whole article.

We believe that the released corpus will be particularly useful for training distant supervision classifiers to extract temporally-anchored attribute values. If we know that a given person X was president of a country during a period of time Z (as indicated by the updates to the Wikipedia infoboxes) we should be able to identify sentences containing the person, the country and a time inside that interval, from which a classifier can be trained. Working along this line is part of our immediate future work plans.

The paper is structured as follows: first, Sect. 2 discusses related work. Section 3 presents an overview of our approach, outlining our design and implementation. The components of our system are described in detail: in Sect. 3.1, we describe how we gathered the full edit history log of Wikipedia; in Sect. 3.2 we describe how we processed this information to extract updates to relational data; and Sect. 3.3 discusses the problem of identifying vandalic edits to demonstrate how vandalism can be filtered out using the available data.

Section 4 analyzes and discusses the dataset. Details on the released dataset's format and structure are provided in Sect. 4.1. Section 4.2 explores the generation of temporally anchored relational data from the infobox attribute updates. It shows, with two experiments, evidence that WHAD data can be used as proxy for real world temporal data, with a coverage and an accuracy that has increased over time. Section 4.3 describes the results of the evaluation of vandalism detection.

Finally, our conclusions and future lines of research are reported in Sect. 5.

2 Related work

Information Extraction (IE) is the task of acquiring structured information from unrestricted text or semi-structured sources such as Wikipedia. Data from infoboxes, lists, categories, and disambiguation pages has proven useful to gather semantic information (Suchanek et al. 2007; Auer and Lehmann 2007; Nguyen et al. 2007; Bollacker et al. 2008), and for many other tasks: text classification (Gabrilovich and Markovitch 2007); semantic similarity (Gabrilovich and Markovitch 2007); document clustering (Hu et al. 2009); ontology generation (Suchanek et al. 2007; Ponzetto and Strube 2007); entity linking (Milne and Witten 2008); or question answering (Ahn et al. 2004).

This paper is concerned in particular with Wikipedia infoboxes, whose semi-structured layout hints at their usefulness for knowledge extraction. Our approach is similar in spirit to Auer and Lehman's DBPedia (Auer et al. 2007; Auer and Lehmann 2007), who proposed parsing infoboxes as a way of automatically obtaining knowledge.

The first contribution of this paper is to extend this approach by using the full Wikipedia edit history, and not just a particular snapshot. Wikipedia infoboxes are

commonly designed to display the current value of attributes. We hypothesize that values that were valid in the past lie hidden among the older revisions. Not only the *surface* of Wikipedia, but also the underlying layers of *historical strata* can be mined for knowledge.

Recent research has started tapping into the Wikipedia edit history for a variety of tasks. We can establish two categories amongst them:

(a) Analyses of Wikipedia itself, the quality of its articles and the collaborative edit process. Examples of this include analyzing user and edit patterns (Voss 2005); measuring article quality (Zeng et al. 2006; Wilkinson and Huberman 2007); and detecting vandalism (Potthast et al. 2008; Chin et al. 2010).
(b) Research exploiting Wikipedia content as a resource or corpus for machine learning and natural language processing (NLP) tasks: sentence compression (Yamangil and Nelken 2008); textual entailment corpus expansion (Zanzotto and Pennacchiotti 2010); or unsupervised learning of lexical simplifications (Yatskar et al. 2010), exploiting the availability of an edition of Wikipedia in *simple English*.

To the best of our knowledge, the edit history has not been used yet to extract relational knowledge. Recently, an open software library that implements delta-compression of Wikipedia's edit history, and access through a Java API might ease the burden of processing the data, and facilitate future research (Ferschke et al. 2011). present an in-depth survey of methods and applications that exploit Wikipedia's dynamic and collaborative nature, materialized both in its edit history and its discussion pages.

The information we extract is not only interesting by itself, but also because it can be applied to other tasks. *Distant supervision* consists in semi-automatically annotating a dataset that can be used as training set for a supervised ML algorithm (Banko et al. 2007; Mintz et al. 2009; Hoffmann et al. 2010). Wu and Weld's system, KYLIN (Wu and Weld 2007), uses attributes extracted from Wikipedia infoboxes to bootstrap a distantly supervised learning system, in order to extract new attribute pairs. In this way, two drawbacks of supervised machine learning methods are tackled: the need for labour intensive labeling of training data, and for specifying the full set of semantic relations to be extracted as an input.

Our system extracts attribute updates that are temporally anchored, which would allow expanding previous approaches with this additional temporal information. The extraction of temporal information is an important open challenge for Information Extraction. Significant research, particularly around the TempEval community (Verhagen et al. 2009), has focused on the classification of the temporal links between events and temporal expressions, exploiting supervised machine learning techniques enabled by the release of the TimeBank temporally annotated corpus (Boguraev et al. 2007). The 2011 edition of the Knowledge Base Population track[4] at the Text Analysis Conference-2011[5] included the acquisition of temporally anchored attribute values. Recent research has extracted temporal facts from infoboxes, categories and lists, to be integrated with a pre-existing ontology

[4] http://nlp.cs.qc.cuny.edu/kbp/2011/KBP2011_TaskDefinition.pdf.

[5] http://www.nist.gov/tac/2011/.

(Wang et al. 2010; Zhang et al. 2008). These works do not use the edit history, so the facts they can extract are those present in a particular snapshot.

Working with the full edit history of Wikipedia comes not only at the cost of processing many edits to articles, but also of dealing with many *erroneous edits*. Some of these errors are the result of *vandalism*. Other, non-vandalic, errors within Wikipedia content are out of the scope of this work. An open line of research focuses on those *quality flaws* (Anderka and Stein 2012). The pervasive nature of vandalism in Wikipedia compromises its value as a resource for ML and NLP tasks, particularly when using the edit history rather than a single snapshot. Past research trying to leverage Wikipedia edit history often overlooks this issue, or leaves it for future work (Yamangil and Nelken 2008; Yatskar et al. 2010). Manually annotating a corpus might be feasible for small datasets (Zanzotto and Pennacchiotti 2010), but not for large-scale knowledge acquisition, and therefore it is unsuitable for our purposes.

We address the issue of vandalism in this paper, demonstrating how the information contained in the dataset can be exploited to filter out vandalic edits without using data extrinsic to the dataset. A contribution of this paper is to analyze how vandalic edits *to infoboxes* can be detected. As we have described above, infobox attributes are fed into other systems, so the scenario of having to decide whether a modification of an infobox is vandalic, without relying on full-page features, is realistic. Previous detection systems deal with full-page edits, while we are interested in edits to infoboxes.

Although vandalism in Wikipedia has been observed from its inception, relevant research on this topic is quite recent. The first systems to address the issue were automated scripts, or *bots*, based on heuristic rules, with an eye on high precision but very poor recall; see Geiger and Ribes (2010) for a historical analysis of these bots. Much research has been encouraged by the release of the manually annotated PAN-WVC-10 English Wikipedia vandalism corpus (Potthast 2010), extended later to German and Spanish (Potthast and Holfeld 2011), and the first two editions of a vandalism detection competition, PAN 2010 (Potthast et al. 2010), and PAN 2011 (Potthast and Holfeld 2011).[6] The proposed systems can be grouped by the kind of features that they focus on: the article *text content* and its revision history (Potthast et al. 2008; Smets et al. 2008; Chin et al. 2010). A related idea is to use the *compression rate* of edits (Itakura and Clarke 2009), although such methods tend to overlook small-sized vandalic edits. The best participant of the PAN 2010 competition (Mola-Velasco 2010), used textual and *linguistic features*. *Reputation*, particularly of users, was used by Adler et al. (2010) and West et al. (2010). This last paper also exploited *metadata* features. Adler et al. (2011) used the lessons learned in the previous work to implement a classifier by merging features from previous systems. Their work also compares the relative merit of features of different nature. We will compare the performance of our own vandalism detector to that of their revised, state-of-the art system. In the second edition of the competition, the best participant system (West and Lee 2011) demonstrated that a significant improvement is possible using features that exploit *a posteriori* knowledge, that is, taking into account later revisions to the one to be classified.

[6] The corpus is freely available at http://www.uni-weimar.de/cms/medien/webis/research/corpora/pan-wvc-11.html.

3 System overview

We hypothesize that Wikipedia is a useful resource for temporally-anchored knowledge. Relational facts that were valid in the past but have been overridden with more up-to-date information are hidden in the edit history. Our aim is to uncover this knowledge, tied to the time when it was valid. The reliability of this resource is affected by the presence of erroneous and vandalic edits in the edit history. In this paper, we demonstrate how information intrinsic to the dataset can be exploited in order to filter out vandalic edits without relying on extrinsic information.

The research question central to this work is: how can we generate historical data from the Wikipedia edit history in a robust way, mitigating the presence of vandalism? An outline of our approach is graphically depicted in Fig. 1; methodologically, it involves the following main steps:

(A) Data gathering: obtain the revision history from Wikipedia. We have implemented accessors to two sources of such data: the Wikipedia database dumps and a tailored crawler of Wikipedia to keep an up-to-date log of revisions. This process is detailed in Sect. 3.1.

(B) Harvesting infobox attribute updates. The edit history has to be processed to extract the relevant information; in our case, updates to the infoboxes in successive revisions (Sect. 3.2).

(C) Detecting vandalic edits. Vandalic and erroneous data are a burden for data quality. We show how we automatically filter out vandalic edits in Sect. 3.3.[7] data is out of the scope of this work; some lines of development that we are investigating to address this open research question are discussed in Sect. 5. The evaluation and analysis of this approach is described in Sect. 4.3.

(D) Generating temporal anchoring from selected edits. The relational facts we extract and store have an additional temporal dimension; it is possible to anchor facts to the time when they were introduced in Wikipedia. In Sect. 3.4, we introduce this representation of temporally anchored relational information, and in Sect. 4.2 we empirically explore whether such an approach can produce accurate and timely relational data.

The implementation of each of these steps is detailed in the following subsections, where we also discuss technical and scientific challenges that we have encountered during our development, experimentation and evaluation.

3.1 Data gathering

The Wikimedia foundation makes the Wikipedia edit history available for download at http://download.wikipedia.org/.[8] For this research, we have focused on the English edition

[7] Notice that the issue of detecting other kinds of *incorrect* data is out of the scope of this work; some lines of development that we are investigating to address this open research question are discussed in Sect. 5.

[8] Wikipedia makes database downloads available, including those of the full edit history of every article. All text content is released under a double license: the Creative Commons Attribution-ShareAlike 3.0

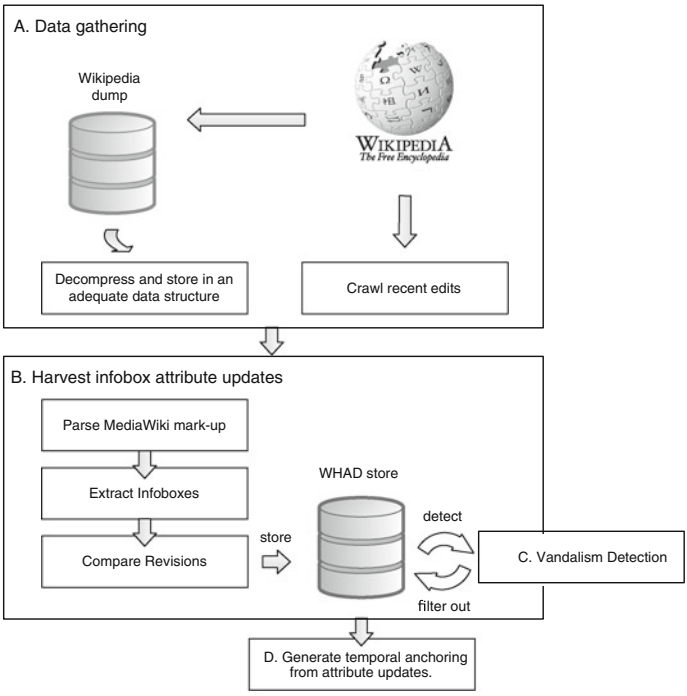

Fig. 1 System overview diagram

of Wikipedia, although the techniques we employ could be readily adapted to other languages. The downloaded file[9] contains all the articles that exist in the on-line version, together with the full sequence of edits for those articles. For each article, and for each revision, these data includes not only the full text, but also the discussion page, infoboxes and category annotations. Note that deleted articles are excluded from this archive.

The revision history is generated periodically, although on a somewhat irregular basis. Throughout the development and experimental phases of this work, we used a *dump* from January 30, 2010 of the English Wikipedia containing the full edit history of the articles. This is the compressed representation of an extremely large xml file, so storing and processing the file is not trivial. The .bz2 file that we downloaded is 280.3 GB in size. It was necessary to use a program that, as it decompresses the dump, distributes the different Wikipedia entries in many smaller files to be stored in a distributed file system in order to make it usable. We processed only content articles, and rejected all disambiguation, redirect and discussion articles. All further processing of these data was performed using an implementation of the MapReduce paradigm (Dean and Ghemawat 2008) inside a computing

Footnote 8 continued
License (CC-BY-SA) and the GNU Free Documentation License (GFDL). For details on the different download options, see: http://en.wikipedia.org/wiki/Wikipedia:Database_download.

[9] Specifically, the download of the English Wikipedia with its full edit history that we have used for this research, and newer versions available later, is distributed at http://dumps.wikimedia.org/enwiki.

cluster. The experimental results reported in Sect. 4.2 are performed on this dataset, which we will refer to as WHAD2010.

The availability of up-to-date revision data would be limited if the processing depended on the schedule of Wikipedia database dumps. To keep our dataset up-to-date, it is necessary to be able to crawl current revisions from Wikipedia. We have worked in this direction, and have now the infrastructure capable of crawling newer revisions in place, incrementally updating the obtained dataset using the MediaWiki API. The dataset described here, which constitutes the first release, is the most up-to-date at the time of writing, dated March 23, 2012. We will refer to this dataset as WHAD2012, and we will describe it in detail in Sect. 4.1 We plan to produce data refreshes periodically.

3.2 Infobox update extraction

In order to extract infobox updates, we follow a similar approach to Auer and Lehmann (2007), which is outlined as follows:

1. Parse the MediaWiki mark-up to identify infoboxes in all revisions for each entry. To do this, we have have employed our own parser, a Flex-based[10] lexical analyzer, tailored specially for MediaWiki template extraction.
2. Get the infobox type and all the (attribute name, attribute values) pairs contained in it. The main difficultly at this stage is to parse and interpret the MediaWiki mark-up language which is subject to frequent changes, and combines semantics and visualization information (Völkel et al. 2006; Wu and Weld, 2010). A notable source of complexity is the possible nesting of templates and lists; another is the behaviour of the MediaWiki parser, responsible of rendering the output (X)HTML, in the presence of errors. As stated in MediaWiki's documentation: "every input string should derive to the most-likely result, even if it contains syntax errors".[11] Our own parsing procedure is slightly more strict, skipping some unparseable edits.[12] Furthermore, MediaWiki allows for the use of *templates* to embed content inside a page. Templates are created and curated in the same way as any other page, and are subject to change at any time.
3. Some of the mark-up, such as hyperlinks to other entities in Wikipedia (e.g. if the value of an attribute is the title of a different entry) is also kept, together with the canonical name of the landing page. If the link pointed to a redirect page, the canonicalized landing page is obtained from resolving the redirect.

The differences with respect to previous approaches to infobox parsing are:

- For each entry and revision-timestamp we store an infobox instance, containing tuples of the following form:

$$(attribute, value_{prev}, value_{current}, timestamp)$$

[10] See http://flex.sourceforge.net/.

[11] MediaWiki, *Markup spec* http://www.mediawiki.org/wiki/Markup_spec, retrieved February 1, 2012.

[12] The number of edits skipped because of parse failures is negligible: 119.

where we extract from the revision the name of the attribute, the value that was edited out in that revision, and the value standing after the revision.[13] The relational fact is in this way augmented with an anchoring timestamp; we will study in Sect. 4.2 how this temporal information can be exploited. For newly added attributes, $value_{prev}$ is empty, and for attribute names that are removed from an infobox, $value_{current}$ is empty.

- Most of the changes to Wikipedia are edited by other users who have established *alerts* on certain pages. A page that is vandalized often can also be blocked to prevent further edits. These procedures are in place to ensure that vandalic edits have a short life until they are reverted. This means that the amount of vandalism on a given frozen version of Wikipedia is expected to be low. On the other hand, by looking at the whole edit history, all vandalic edits are available at some point in time. Likewise, there may be revision inconsistencies or markup errors in temporary versions of the pages that were edited afterwards and may make one particular revision impossible to parse. This means that vandalism is going to be a greater problem for our knowledge base than it is for a system analyzing a given frozen snapshot. We address this problem in the following subsection.

3.3 Vandalism detection

Working with the edit history as a data source, erroneous and vandalic edits are a concern for data quality. Simple heuristic filtering can be used to weed out the most obviously vandalic content,[14] but some kinds of vandalism will still not be detected. In this section, we show that it is possible to filter out vandalic edits to infoboxes in Wikipedia, and therefore to maintain a reasonable quality in the data.

Since the dataset described in the previous section includes every single modification to an infobox performed by a single user, it contains malicious, vandalic edits. Following previous research, we adopt Wikimedia's definition of a vandalic edit (http://en.wikipedia.org/wiki/Wikipedia:Vandalism): *any addition, removal, or change of content in a deliberate attempt to compromise the integrity of Wikipedia*. The issue of non malicious factual errors is not the focus of our investigation; in Sect. 5, we point out possible lines of research to address it.

As an example, Fig. 2 shows the number of edits for the attribute *president* in the infobox of the entry *France*. As in most countries, the president in France is elected every few years, and therefore this should be a fairly stable attribute. On the other hand, the figure shows that between 2006 and 2010 there have been 116 edits of the name of the French president. Some of them are accessory but legitimate changes, such as adding or modifying the name of the political party to which the president is

[13] Some other metadata is kept, see Sect. 4.1 for more details.

[14] Removing, for instance, edits which textual content is too long or too short, or edits that were rapidly reverted.

affiliated. Many of the spikes in the number of changes per month, though, were due to users adding, apparently deliberately, incorrect values. Fortunately, for popular Wikipedia entries, such as this one, vandal edits are usually reverted quickly. One particular source of extra revisions for this attribute comes from the interval of time between Sarkozy's election and office assumption (around May 2007), during which contributors did not agree whether the value should be already updated or not. The next larger spike (on July 2009) comes from a single (vandalic) user insisting over and over again that the president of France is Philippe Petain. All of these edits were reverted in a matter of minutes.

3.3.1 Procedure

Our aim is to show that it is possible to detect vandalic revisions using only the infobox update information that we collected. We think that this is a reasonable scenario, because of the following reasons:

- The infobox-only corpus is much easier to handle than the full revision history, both in terms of computational requirements (size and processing cost) and in terms of structure, as we can consider each changed attribute independently in order to obtain a more accurate representation of what is changing from version to version.
- Our focus is on relation extraction, and in particular on using infoboxes for this purpose. We are therefore not interested in being able to classify vandalism for edits that did not modify infoboxes, as these will not be reflected in our dataset.

We describe in the following the features included in our model. Some of them are similar to those used in standard full-article vandalism detection work (Mola-Velasco 2010; Potthast et al., 2010), but most are specific to our problem (updated infoboxes). We did not perform any manual feature engineering aside from putting together all the features that intuitively seemed useful for this task and were not too correlated with each other.

- *Lexical features*: whether the revision is adding new sex words (e.g. *sex* or *porn*) or vulgar or offensive words (e.g. insults, *nazi*, etc.) in the value of an attribute, or in the name of the infobox that is being edited. These features are language dependent, but similar lists of words can be compiled for other languages.
- Whether the revision has been tagged as "minor", or the comment indicates that the editor was actually a bot.
- Whether the contributor is identified with a user ID or an IP address.
- The number of infobox attributes added in this revision.
- The number of infobox attributes removed in this revision.
- The number of infobox attributes whose value changed in this revision.
- Statistics about how long it took for the changed attributes from this revision to be changed again: the average, minimum and maximum number of seconds for attributes changed in this revision until they are changed again by a later revision. These statistics are collected separately for added attributes, deleted attributes and changed attributes.

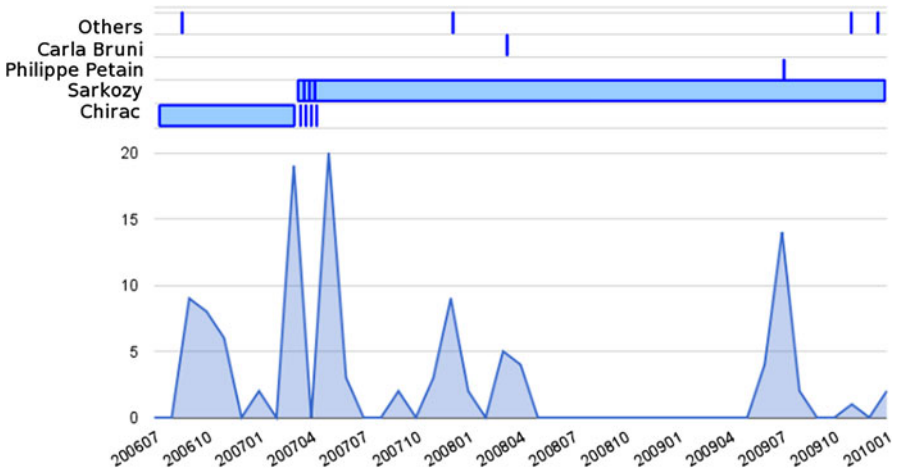

Fig. 2 Number of monthly revisions for the attribute *leader_name1* (corresponding to the president) of the entry *France*. Except for May 2007, all the other spikes come from vandalic edits

- The number of attribute values added, deleted or changed in this revision that were still valid by the time the edit history dump of Wikipedia was collected (January 30, 2010).
- Whether a whole infobox is being created or removed in this revision. Additionally, whether it existed in the past before being re-created now, or whether it was re-added later if it is being deleted now.

For classification we used an AdaBoost classifier: during our development work the choice of learning algorithm did not seem to affect the results on our development set much, and AdaBoost gives reasonably good results according to Wu et al. (2010). During early development we observed that, because of the highly imbalanced dataset, it was often the case that most ML algorithms simply learned to tag everything as non-vandalic, so we used a cost matrix penalizing false negatives ten times more than false positives. No other parameters were tuned on the development set.

3.3.2 Development and test sets

PAN-WVC-10 (Potthast 2010), which was developed by means of *crowdsourcing*, is probably the most comprehensive English Wikipedia vandalism corpus publicly available. The dataset contains 32,439 manually annotated Wikipedia edits, out of which 2,394 are classified as vandalic (around 7.5 %).

The sampling procedure performed by Potthast to select the revisions that were to be annotated weighted each article with the number of times that it was edited, so as to give more importance to documents that attracted more attention from Wikipedia contributors. To use it as test set, we adapted this gold standard by removing all revisions that were not modifying any infobox (because these are not present in our dataset). After removing those, the dataset obtained has 2,839 revisions, out of which 128 are labelled as vandalic (4.5 %), a ratio slightly lower than that of the full set.

For development purposes, we created a new development set in the same spirit of PAN-WVC-10. To do this, we randomly sampled 1,000 Wikipedia edits that modify at least one attribute in an infobox. To annotate the edits, we use a proprietary *crowdsourcing* approach similar to, but independent from services such as Crowdflower[15] or Amazon Mechanical Turk.[16] Non-expert annotators were assigned the task of deciding whether a particular revision in our dataset was vandalic. They were provided with the following information:

- The entry that was being changed, together with a pointer to the current version of the entry.
- The Wikipedia *diff* page[17] between the revision that we would like annotated, and the previous revision.
- The date of the revision.
- The infobox attribute that changed, the previous value and the new value for that attribute.

The task of the raters was to annotate the revision with one of the following options:

(a) It is a regular, legitimate revision.
(b) It is a vandalic revision.
(c) I don't know.

An example of the annotation form raters were provided is shown in "Appendix".

Three annotations were collected for each item, and each rater was set a limit of at most 30 ratings, to avoid bad raters having a large effect on the whole annotation. A total of 233 annotators participated in the task. The revisions without majority agreement or where *"don't know"* was the majority label were discarded. The final development dataset contains 74 items marked as vandalic edits (9.6 %), and 770 items marked as legitimate edits. Observe that the percentage of vandalism present here is somewhat larger than in the test set.

In our *crowdsourcing* evaluation setting, three ratings are assigned to each evaluation item, and the items are annotated by different raters, from the total of 233. A suitable statistical measure of inter-annotator agreement under these conditions is Fleiss' kappa (Fleiss et al. 2004). While Cohen's kappa is a good estimate of agreement between 2 raters, Fleiss's kappa is useful if there are more than 2 raters and/or if the ratings have been issued by different raters. This coefficient quantifies the extent to which the observed amount of agreement among raters exceeds what would be expected if ratings were completely random:

$$\kappa = \frac{P_o - P_e}{1 - P_e}$$

[15] http://crowdflower.com/.

[16] http://aws.amazon.com/mturk/.

[17] A Wikipedia diff page shows the difference between two versions of a page.

where P_o is the proportion of pair-wise agreements and P_e is the expected proportion of such agreements under a random assignment. Considering that our task is detecting vandalic edits, we chose to aggregate the responses of the categories "I don't know" and "It is a regular, legitimate revision". In this setting the Fleiss' Kappa value is: $\kappa = 0.30436$, which can be considered a "fair agreement".

Furthermore, there was a majority agreement (that is, at least 2 out of the 3 annotators did agree) in 937 of the examples (93.7 %). In 488 examples (48.8 %), 2 out of three annotators agreed, while in 449 (44.9 %) examples, all three annotators agreed.

Inspecting the comments in the 63 examples where a majority agreement was not reached, we observe that one of the major causes for disagreement (25 % of the cases) was that the raters tried to verify the factual correctness of a modified value and reached different conclusions. The other 25 % of the disagreements was due to the limitations of our extractions and annotation interface, as it was sometimes difficult for the annotators to decide whether an edit is vandalic when the infobox type is changed (maybe correctly), when it is removed or recovered altogether, or when the edit introduced a syntactic error in the MediaWiki code.

Also, when an image is modified, the Wikipedia diff page that is shown to the annotators provides only the name of the image,[18] so their decision based on that was difficult. This accounts for 15 % of the disagreements.

Roughly 10 % of the disagreement was caused by rater errors, which could be detected from comparing their decision to the comments they provided. The rest of the disagreements are caused by a variety of reasons, such as the edit changing an already wrong value, disagreement over the relevance of a piece of data, or over possible spelling mistakes.

We consider this level of agreement reasonable given the nature of the task, and the observed results are consistent with those reported in Potthast (2010). The corpus described in that work, Webis-WVC-07, consists of 940 human-annotated edits of which 301 are vandalism(Potthast et al. 2008). Webis-WVC-07 was annotated using Amazon Mechanical Turk; with a three-raters per example setting, and the author reports three-out-of-three agreement in 58 % of the cases, and two-out-of-three in 42 %. Note that, as opposed to our setting, here the raters' judgements are binary.

3.4 Temporally anchored relational data

One of the main potential uses of revisions of Wikipedia infoboxes is to recover historical values of infobox attributes: it is possible to anchor facts to the time when they were introduced in Wikipedia. The attribute updates in WHAD are temporally anchored, and can be represented as tuples:

$$(attribute, value_{prev}, value_{current}, timestamp)$$

For instance, in order to know the GDP of the United-States in 2005, one would ideally look at the value of a revision of the attribute late in 2005 or early in 2006.

[18] There exists a file history for image files, but it is not immediately available from the diff page.

The WHAD dataset can be used to enhance existing knowledge bases, such as DBPedia or Freebase, with historical values. Another possible application would be to date documents according to the age of information mentioned in their content.

In Sect. 4.2, we empirically explore whether such an approach can produce accurate and timely relational data

4 Evaluation, analysis and discussion

This section describes the results of an analysis of the data produced, and discusses its usefulness. We provide a more detailed description of the released data and a study of the timeliness of manual updates to Wikipedia that affect attribute values. Finally, we report the experimental results of the automatic vandalism detection experiment.

4.1 Dataset analysis

As a contribution of this work, we are releasing the full, up-to-date, dataset of Wikipedia infobox attribute updates, WHAD2012, for further research. In this section we start by providing general descriptive statistics about the dataset and then concisely describe the format and structure of the data.

Our aim is to distribute the most recent dataset possible; as we are able to process recent versions obtained by crawling regularly Wikipedia, we have augmented the data from the 2010 dump on which we performed our experimental analyses (WHAD2010) with more recent updates. This release dataset, updated to March 23, 2012, WHAD2012, is the one described in this section.

4.1.1 Descriptive statistics

From our store of Wikipedia's full edit history updated to March 23, 2012, we filter out non-content, disambiguation, redirect and discussion articles. The total number of revisions we parse is 291,601,701. From them, we extract a total of 510,102,778 individual infobox attribute updates (IAU), that correspond to 2,040,181 articles.

Table 1 collects some relevant statistics on the extracted data. For information purposes, the table has another column where a very conservative sanity-check filtering of the dataset has been applied: we remove every revision that introduced a string value of more than 10,000 characters (being this most certainly a mistake or a vandalic edit), and those edits that were reverted within a minute from being saved.

At the dataset level, we can see that in the period 2003–2012, 510,889,795 infobox attribute updates (IAU) have been made to 2,040,181 different entries by over 7 million users (identified by unique username if available, or by IP address otherwise). Roughly half of the wikipedia entries have an infobox.[19]

[19] As of March, 2011, the total number of Wikipedia pages is over 3.9 million articles. Source: http://en.wikipedia.org/wiki/Special:Statistics.

Table 1 WHAD2012 dataset statistics about infobox attribute updates (IAU) extracted from the Wikipedia edit history

	Full dataset	After clean-up
WHAD2012 dataset level statistics		
Infobox attribute updates (IAU)	510,102,778	38,979,871
Entries with at least one IAU	2,040,181	1,845,172
Users responsible for at least one IAU		
Identified by username	1,242,787	572,349
Identified by IP address	6,033,308	1,752,089
Different infobox templates	24,028	12,727

The clean-up removes edits that introduce a value more than 10,000 characters long or that was reverted within one minute of being saved

Infoboxes are classified according to the type of information they contain, indicating, for example, that they refer to a company or to a country. In our dataset there is a total of 24,028 different infobox type names.[20] The ones that appear in most entries are *settlement*, *album*, *french commune*,[21] *film* and *musical artist*.

At the attribute level, we report statistics on the detected typed values within attribute values. Many attribute values contain a mention of a typed value, such as a location or a date. Also, some attributes contain several typed-value mentions, of the same or different types.

We used a combination of gazetteer and regular expression-based Named Entity recognizers with manual heuristics, developed in-house, in order to normalize the values and characterize their types. Taking as input the 38,979,871 attribute updates after simple clean-up, we computed the number of updated values that contain one among a set of potentially interesting value types: numbers, hyperlinks, geographical locations, dates, measurements, currency values, and also time expressions and temporal intervals. Table 2 reports these statistics of the detected types. Note that the list of types is not necessarily comprehensive and that a type has not been detected for every attribute value. Also, more than one value can be identified and counted for a single attribute update. For instance, the value might be a list, and we detect an entity for each of its elements: the "developer" field of the article *Unix* has value: "Ken Thompson, Dennis Ritchie, Brian Kernighan, Douglas McIlroy, and Joe Ossanna at Bell Labs"; we detect and count six hyperlinks in such a value.

[20] Observe that not all of them are valid infobox names, as many are in fact editors errors, or vandalism.

[21] The high frequency of the "french commune" infobox might be surprising, but has a simple explanation. The *commune* is the lowest level of administrative division in France, and can range from a large city to a small village. As of January 9, 2008, there were 36,781 communes in France, and through the collaborative effort of a group of editors, most of them have an article, following a common template that defines the specific "french commune" infobox. See http://en.wikipedia.org/wiki/Communes_of_France and http://en.wikipedia.org/wiki/Wikipedia:WikiProject_French_communes.Similar reasons make "settlement" the top frequency infobox.

Table 2 WHAD2012 statistics about the values detected within attribute updates, after the clean-up that removes edits that introduce a value more than 10,000 characters long or that was reverted within one minute of being saved

Note that a type has not been detected for every attribute value, and that an attribute might simultaneously contain values of several types (e.g., a number and a date)

WHAD2012 attribute-level statistics	
Numbers in values	16,768,355
Hyperlinks in values	
To external pages	319,788
Within wikipedia	9,960,073
Locations in values	3,053,105
Dates in values	2,425,012
Measurements in values	425,576
Times in values	246,429
Currencies in values	206,483
Time intervals in values	79,299
Attribute updates with no type detected	19,494,842
Total attribute updates	38,979,871

4.1.2 Data generated, structure and format

The generation of the full dataset is indeed costly, as it involves the following steps:

- Either download a Wikipedia database dump to obtain the full edit history up to the moment of its creation, or crawl Wikipedia to obtain up-to-date versions of the article history.
- Transform this data into a usable database format, decompressing it and storing it in a suitable data repository structure.
- Extract and collect the updates to Wikipedia infoboxes, which involves the parsing of each of the revisions to every article (except those non-content articles that we filter out). This last step took 18 hours in a 128-core cluster.

The size of the dataset which will be released to the community is 5.5 GB; compared to the original edit-history dump released by Wikimedia, this new corpus is much easier to handle for everyone thus facilitating research in this area. The format of the data is JSON. Each text line corresponds to one Wikipedia entry. It has as fields:

- article_title: the name of the entry
- attribute: a list of attribute updates, each of which has:
  - timestamp: the time the attribute was changed.
  - contributor: the contributor ID or, if it is unavailable, the contributor's IP address.
  - infobox_name: the name of the infobox that had this attribute changed.
  - oldvalue: the value of the attribute prior to the update.
  - newvalue: the new value of the attribute (not present if the attribute was removed).

As an illustration, Listing 1 shows a sample from the actual dataset.

Listing 1 A sample from the JSON of the dataset.

```
{
  # [... excerpted ...]
  "article_title": "France",
  "attribute": [ {
    "title": "France",
    "timestamp": 1129500148,
    "contributor_ip": "Golbez",
    "key": "GDP_PPP",
    "newvalue": "$1.744 trillion",
    "oldvalue": "$1.661 trillion",
    "infobox_name": "Infobox_Country",
    "id": 25688496,
    "comment": "substing infobox"
  }, {
    "title": "France",
    "timestamp": 1129500148,
    "contributor_ip": "Golbez",
    "key": "GDP_PPP_year",
    "newvalue": "2004",
    "oldvalue": "2003",
    "infobox_name": "Infobox_Country",
    "id": 25688496,
    "comment": "substing infobox"
  },
  # [... excerpted ...]
  {
    "title": "France",
    "timestamp": 1142610920,
    "contributor_ip": "MJCdetroit",
    "key": "GDP_PPP_year",
    "newvalue": "2005",
    "oldvalue": "2004",
    "infobox_name": "Infobox_Country",
    "id": 44224442,
    "comment": "Reformated infobox  & updated it; also expanded
        the Geography section"
  }, {
    "title": "France",
    "timestamp": 1142610920,
    "contributor_ip": "MJCdetroit",
    "key": "GDP_PPP",
    "newvalue": "$1.816 trillion",
    "oldvalue": "$1.774 trillion",
    "infobox_name": "Infobox_Country",
    "id": 44224442,
    "comment": "Reformated infobox  & updated it; also expanded
        the Geography section"
  }
  # [...]
}
```

Listing 2 An excerpt of the MediaWiki source code of Alan Turing's article, as retrieved in February 2012. Observe the special purpose templates of the `birth_date` and `death_date` attributes.

```
{{Infobox scientist
| birth_name        = Alan Mathison Turing
...
| birth_date        = {{Birth date|1912|6|23|df=yes}}
| birth_place       = [[Maida Vale]], London, England,<br/> United
     Kingdom
| death_date        = {{Death date and age|1954|6|7|1912|6|23|df=
    yes}}
...
}}
```

It has to be noted that values of attributes are often special-purpose templates themselves. For instance, an excerpt of Alan Turing's article can be seen in Listing 2. As shown, the birth and death dates are encoded using a template, with some fields denoting the year, month and day of the date. Before processing the values of the attributes, specialized parsing of these value templates has to be performed, a procedure we followed prior to the experiments described in this paper. In our dataset release, on the other hand, we aim at providing the maximum coverage of attributes, and letting the users decide which parts are important to them. The produced dataset includes the verbatim string values for all attribute values, so clients of this resource can write the simple analysis tools needed to interpret the information they are interested in.

4.2 Accuracy and timeliness of temporally anchored relational data

As described above, the attribute updates in WHAD are a source of temporally anchored relational information, that can be represented as tuples:

$$(attribute, value_{prev}, value_{current}, timestamp)$$

In this section, we show to what extent this approach can produce accurate, timely relational data, and to demonstrate the kinds of analyses that the WHAD dataset enables. We focus on the WHAD2010 compilation of Wikipedia updates we obtained by processing the *dump* from January 30, 2010 of the English Wikipedia, as it is described in Sect. 3.1.

We address the following practical research questions: To what extent is the information contained in past revisions to Wikipedia infoboxes useful for knowledge extraction purposes? Is it reliable data? If the delay between an event occurring and the Wikipedia infobox being updated is short enough, it could indeed be used for event detection, so how often was the data updated? Our method of investigation is to analyze empirically these two different aspects of the quality of Wikipedia historical information: accuracy and delay.

Previous work has dealt with quality assessment of Wikipedia entries, but most has focused on a particular time, or snapshot (Stvilia et al. 2005; Arazy and

Table 3 WHAD2012 top and highest changing company attributes

Top attributes	Frequency	Most updated attributes	Revisions
Foundation	23,907	Revenue	6.03
Industry	22,674	Company logo	5.84
Homepage	22,324	Net income	5.48
Company name	20,422	Market cap	5.43
Location	18,218	Company type	5.31
Company type	17,826	Company slogan	5.15
Key people	16,296	Key people	5.10
Products	15,108	Number of employees	5.03
Company logo	13,721	Operating income	4.95
Number of employees	10,257	Products	4.24

Nov 2010).[22] In contrast, the analysis proposed in this paper focuses on time. The problem addressed here is to assess the quality of information *over a time period*.

We use the infobox types as the different categories in order to explore which attributes are more common in each type of infobox, and which ones change most often. The infobox type *company* is an interesting example because it has a number of attributes whose value changes over time, such as revenue or number of employees, allowing us to investigate the availability of data that was correct in the past but was later substituted by more recent values. Table 3 shows the analysis for the infobox type *company*. The left part of the table contains the attributes that appear most often in infoboxes of this type, and the right part contains the average number of times that the value of an attribute changed in any entry. As can be seen, the highest ranking attributes in number of revisions refer to *transient* properties of companies.

4.2.1 Accuracy analysis case study

We define transient attributes as those whose value changes in real life, either periodically (e.g. the GDP of a country, which is generally computed on a quarterly or yearly basis) or irregularly (e.g. the number of Grand Slam victories for a tennis player). We present here a case study on a periodic attribute, the population estimate for countries, whose real value is updated every year for most countries. We used the World Bank website to acquire actual population sizes for a large number of countries from 2005 to 2009. This website provides information for all of these years for 91 countries. By evaluating how accurately Wikipedia reflects the population of these countries we can get a measure of the reliability of the information.

Country population estimates are often updated when new estimates are published by sources such as the World Bank, the International Monetary Fund, or an official authority responsible for carrying out national census. Furthermore,

[22] A notable exception is volatility, which is defined in Stvilia et al. (2005) as the median revert time.

these sources can typically issue several population estimates for a given year as more and more data becomes available. Therefore, to estimate the population size for a given year, we define our Wikipedia-based proxy estimation as the value of the attribute at the time of the last revision in that year. Our proxy discounts revisions marked as vandalism and revisions whose values deviate by more than 1 standard deviation from the mean to avoid outliers. For example, an incorrect parse might be caused by a user using dots instead of commas to separate three-digit groups. These incorrect readings are easily discarded.

We can then compare the estimated values with the actual values for the 91 available countries. There were 400 entries in Wikipedia with a country infobox, which is more than the number of countries listed by the United Nations. The difference is due to the fact that the infobox has also been applied to regions (e.g. Ile de France). For all the 91 countries with population data from the World Bank there was a corresponding entry in Wikipedia with the country infobox.

Table 4 shows the median, mean and standard deviation of the difference between the observed and actual values, given as a percentage of the real value. For instance, in the case of the median, we see that the population included in Wikipedia differs from the real population value of these countries by roughly 2 % of its value.

To assess coverage, we study the proportion of countries for which the country infobox provides an estimate of the population size. Figure 3 indicates that from 2005 to 2009, the proportion of countries with a population attribute has steadily grown from 58 to 100 %. In the same time period, we can see that the proportion of countries providing an estimate that was updated in the previous year has also grown from 58 to 90 %.

These statistics suggest that the estimates for the population size are generally accurate and have a good coverage, which is improving over time.

4.2.2 Delay analysis

Popular Wikipedia entries are generally updated almost instantly when new information becomes publicly available. For example, Wikipedia entries reporting game scores of famous soccer teams are typically updated within seconds after the end of the game. However, update delays for less popular entries may be significantly longer. Thus, to avoid selecting obsolete values, a Wikipedia-based

Table 4 Accuracy of the Wikipedia values for population size measured against World Bank population estimates for 91 countries: median, mean and standard deviation of the difference between the observed and actual values, as a percentage of the real value

Year	Median	Mean	SD
2005	1.65	3.40	1.80
2006	1.62	2.92	1.81
2007	2.19	3.75	2.40
2008	1.83	3.52	2.33
2009	1.59	2.28	1.89

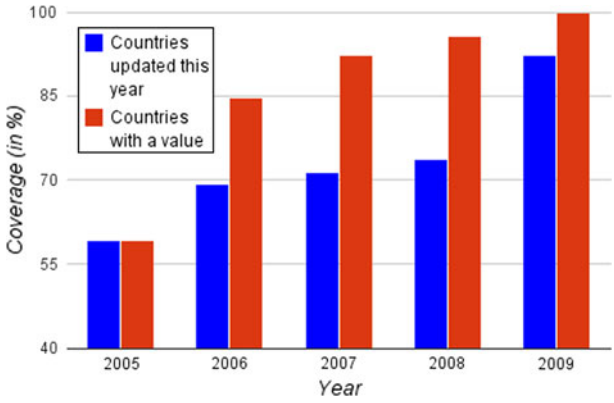

Fig. 3 Evolution of proportions of countries having a population-size attribute and having a new value provided in the year

proxy for temporal attribute values needs to take into account that some delays may happen before the real value is introduced in the dataset. This section describes an analysis of the latency of infobox date attributes.

Let us define the *latency* of an attribute update as the difference between the moment a piece of information was first available and when it was included as an attribute in the relevant Wikipedia infobox. Measuring the latency of an attribute is generally problematic because the exact time when the information item became available might be difficult to obtain. In particular, sources of historical data seldom report when the information was first available. In the previous experiment on population sizes, the World Bank data gives historical values but it does not indicate when exactly these values were released. We shall look for other attributes, that are more adequate for measuring latency.

Infobox attributes whose value is the *date* in which an event took place offer an additional temporal reference, which we can use for our purposes: we can focus on infobox attributes whose value is a date that has to be previous to the update to the corresponding infobox attribute. For instance, the infobox *person* has the attribute death-date; if for the entry of a person the attribute death-date was first filled in date d_u with value d, we can automatically compute how long it took to update the entry from the time the event happened in the real world: $d_u - d$.

In other words, to estimate the latency of a revision of a date attribute we compute the number of days between the revision date and the new value (assumed to contain a year, month, and day). Obviously, we remove from this study old date values. In an extreme case, if we had considered the date of death of Voltaire (30 May 1778), the update delay would be due to the fact that Wikipedia did not exist back then. As a general rule, the latency is only computed for revisions whose value is later than the earliest date when the attribute was first introduced across all entries.

The distribution of delays for various date attributes as computed using this method is reported in Table 5, organized by deciles. For example, in the case of the date of a military conflict, more than 40 % of updates are reported in <2 days. The last time a television show or episode was aired is typically updated much faster

Table 5 Delay (in days) for different infobox *date* attributes, between the value of the date and the timestamp when the attribute was added for the first time

Infobox	Attribute	Decile									
		1	2	3	4	5	6	7	8	9	*N*
Person	Death_date	0	0	1	1	2	5	18	64	296	1,045
Scientist	Death_date	0	1	3	8	31	151	323	396	680	1,093
Military conflict	Date	0	1	1	2	6	14	53	174	381	1,103
Television	Last_aired	0	1	2	4	9	28	77	180	474	1,514
Television	First_aired	1	7	22	49	106	205	352	512	847	1,763
Software	Latest_release_date	1	6	13	23	38	59	97	159	285	1,328
Company	Revenue	3	46	56	85	118	170	214	253	362	1,236
Book	Pub_date	11	31	70	105	143	189	262	366	490	810
All infoboxes		0	0	6	40	120	248	356	541	849	2,043

Each of the nine deciles separates the *N* values in the sample into 10 equal parts, so that each part represents 1/10 of the attributes

than the 'first time aired' attribute. Considering the aggregate values for all date attributes in all infoboxes (the last row of Table 5), 20 % of these date attributes are updated within the day that the event actually happened (the second decile is 0 days).

4.3 Vandalism detection experimental results

This section describes the results of the automatic evaluation of vandalic edits using only infobox data. As mentioned in Sect. 3.3, an AdaBoost classifier was trained on the development set that we have collected, and the test set used is PAN-WVC-10 (Potthast, 2010). As a baseline for comparison we decided to use WikiTrust (Adler et al. 2010), a state-of-the-art vandalism detection system. It has a public API available at http://www.wikitrust.net/vandalism-api, to which it is possible to send a document title and revision ID, getting as a response from the system backend a confidence value, between 0 and 1, of the article revision being vandalic. It is still necessary to define a threshold so that if the confidence value exceeds the threshold the revision will be considered to be vandalic. We have used our own development set to find the confidence threshold that maximizes the F-score for vandalic edits. This system ranked second in a recent vandalism detection competition (Potthast et al. 2010).

Figure 4 shows the ROC curves for our vandalism detection and the system described by Adler et al. (2010). The area under the curve reported by Adler et al. (2010) is 0.90, slightly higher than for our approach (0.88).

Something that can be noted is that the results are in the low end of those reported in the PAN evaluation for the system that we are using as baseline. One possible explanation is that we are using a different test set, including only the edits that affect infoboxes. To verify this, we ran the WikiTrust API on the original PAN-WVC-10 dataset, including changes that did not affect infoboxes. With this set, the obtained area under the ROC curve is again comparable with the results obtained here, 0.88. After a personal communication with one of the authors he indicated that

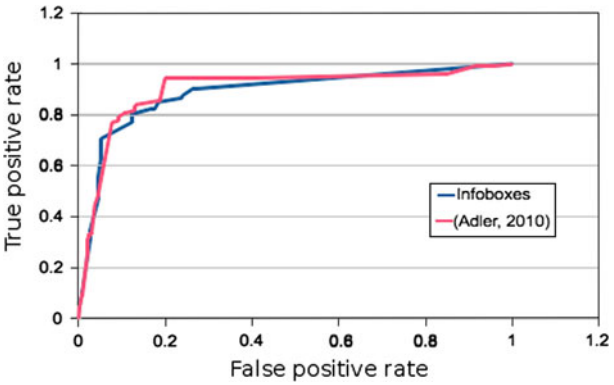

Fig. 4 ROC curve. True positive rate versus false positive rate for the vandalism detection

there may be many reasons for this to happen, such as that user reputation and other parameters also considered for classification may have changed since the PAN-WVC-10 evaluation was performed.

We conclude that in the task of vandalism identification, for those articles that contain infoboxes, using only infobox-dependent features it is possible to attain a performance comparable to that obtained using full entry features.

5 Conclusions and future work

In this paper we have described our approach to extract, compile and make available, from the revision history of Wikipedia, a dataset of temporally anchored relational data. The immediate contributions of this work are the built architecture, capable of extracting infobox attribute updates from the Wikipedia historical data (from both a historical dump and single page updates), and the dataset released.

We are releasing the full, up-to-date, dataset of Wikipedia infobox attribute updates, WHAD2012, under the Creative Commons license that covers Wikipedia. For each entry and revision-timestamp we store an infobox instance extracted for that revision, containing tuples of the following form: (*attribute*, *previous value*, *current value*, *timestamp*), and metadata pertaining to that revision. The full dataset, containing over 510 million such revisions, has been made available. With a total size of 5.5 GB, compared to the original edit-history dump released by Wikimedia, this new corpus is much easier to handle, thus facilitating further research.

We have presented several analyses performed on the dataset, including a case study on the population attribute for countries, showing that the accuracy of the values matches the real values reported by the World Bank within an error of around 2 %, and a study on the delay with which date attributes are encoded in Wikipedia, showing that 20 % of them are updated within a day, and 50 % within 4 months.

One particular characteristic of the revision history is that vandalic content is pervasive; even though most vandalic edits are typically short-lived, working as we do with the full edit history requires us to deal with them, and vandalism

identification becomes necessary. We have described a vandalism classifier using primarily features obtained from infoboxes, without relying on full-page features at all, and showed that we can attain results that are comparable to a state-of-the-art system trained on the full entries in Wikipedia.

In future work, we plan to investigate the following lines of research:

1. Extending the vandalism classifier to include more structured information about the infoboxes. For example, some attributes require different types of named entities as their value and numeric attributes typically only allow their values to belong to a certain range. By collecting statistics about the types of each attribute across the dataset we can probably discover more subtle vandalism that goes undetected without this kind of features.
2. Extending our accuracy analysis to more attribute types.
3. As our approach is language independent, we are considering processing and releasing datasets from languages other than English.
4. Assessing the factual correctness of the information contained in Wikipedia has not been the focus of our investigation, and we leave it for future work. As the breadth of the resource imposes us to consider automated approaches, a possible way for deciding on factual correctness would be to compare the values in different language versions of Wikipedia. This task is non-trivial. First, matching the infobox schemas of different language versions is not direct, and the schemas have to be translated (see for instance Nguyen et al. 2011). Then, as there is no argument type system for Wikipedia infoboxes, the values of the attributes would have to be parsed prior to comparison. Last, the difficulties introduced by the different coverage of different language editions of Wikipedia have to be investigated. Similarly, and inside a single language edition, discussions in the so called talk pages might be exploited to identify erroneous content in past revisions, a line of research that remains open and is also challenging.

Acknowledgments The research leading to these results has received funding from the European Unions Seventh Framework Programme (FP7/2007-2013) under grant agreement number 257790; the Spanish Ministry of Science and Innovation project Holopedia (TIN2010-21128-C02); and the Regional Government of Madrid MA2VICMR (S2009/TIC1542)

Appendix: Manual rating instructions

Instructions

Wikipedia is an on-line encyclopedia to which many users contribute editing the entries. Wikipedia entries sometimes contain one or several small boxes with structured data called Infoboxes. For example, the Wikipedia entry for United States has a small box at the right hand side containing the name of the country, its flag and seal, motto, anthem, capital, and other facts about the country. We'll call each of these lines in the infobox *attributes*.

If you want to read more about Wikipedia Infoboxes, you can see this page.

Wikipedia keeps logs of all the edits done by each contributor during the past many years. This allows us to explore the past changes for each entry. For example, this page shows a particular edit that was done to the entry "Articles of Confederation". In this example, the contributor modified the value of the attribute "writer". This attribute is the one that is used in the infobox line specifying who the authors were. This particular contributor edited the value of the writer from just "Continental Congress" to a new value of an insulting nature. This is a clear case of vandalism. For the purposes of this evaluation, we consider that a contribution is vandalic if either:

- It is adding insulting or obscene content.
- It is plainly false.

If a page contained a correct value and a user replaces it with an incorrect value, we assume that the edit is vandalism. For example, look at this page. The value of the origin (birth place) of Lil Jon was changed from Montreal to Atlanta. The correct value for this attribute is Atlanta. You can click on the "Previous edit" link to see that Montreal was added in replacement of the correct value Atlanta. For these reasons, we'll say that the page was initially correct, Montreal was added in a vandal edit, and the change in the shown page is fixing the vandalism by reverting the value to the previous correct value Atlanta.

You will be shown below the name of an entry, the time when it was changed, name of the attribute in the infobox, the old value of the attribute, and the new value of the attribute. The task is to reply to the questions below to identify possible cases of incorrect values or vandalic actions.

Wikipedia Entry:	African trypanosomiasis
Time when the entry was edited:	Fri, 02 Jan 2009 17:27:57 GMT
Infobox attribute that was edited:	Name
Previous value for this attribute:	Sleeping sickness
New value for this attribute:	African trypanosomiasis
Wikipedia edit diff for this change:	Click here

Questions

Please look at the old and new values of the attribute. To find out whether they were legitimate or not, it is useful to click on the "Previous edit" and "Next edit" links in the Wikipedia diff page, and also to see the value in the current version of the page in Wikipedia nowadays by clicking on the "Article" tab at the top.

This edit was...

- ◯ **it is a vandalic revision** (The new value African trypanosomiasis is a vandalic edit).
- ◯ **it is a regular, legitimate revision**
- ◯ **I don't know**

Comments (required):

References

Adler, B. T., De Alfaro, L., & Pye, I. (2010). Detecting Wikipedia vandalism using WikiTrust—Lab report for PAN at CLEF 2010. In *Notebook Papers of CLEF 2010 Labs and Workshops*.

Adler, B. T., De Alfaro, L., Mola-Velasco, S. M., Rosso, P., & West, A. G. (2011). Wikipedia vandalism detection: Combining natural language, metadata, and reputation features. In A. Gelbukh (Ed.), *Computational linguistics and intelligent text processing*, *Lecture Notes in Computer Science*, Vol. 6609, Berlin: Springer, pp. 277–288.

Ahn, D., Jijkoun, V., Mishne, G., Müller, K., de Rijke, M., & Schlobach, S. (2004). Using Wikipedia at the TREC QA track. In *Proceedings of TREC 2004*.

Anderka, M., & Stein, B. (2012). Overview of the 1st international competition on quality flaw prediction in Wikipedia. In P. Forner, J. Karlgren, & C. Womser-Hacker (Eds.), *CLEF 2012 Evaluation Labs and Workshop—Working Notes Papers*.

Arazy, O., & Nov, O. (2010). Determinants of Wikipedia quality: The roles of global and local contribution inequality. In *Proceedings of the 2010 ACM conference on computer supported cooperative work, CSCW '10*, ACM, New York, NY, USA, pp. 233–236.

Auer, S., & Lehmann, J. (2007). What have Innsbruck and Leipzig in common? Extracting semantics from Wiki content. In *Proceedings of the 4th European conference on the semantic web: Research and applications, ESWC '07*, pp. 503–517.

Auer, S., Bizer, C., Kobilarov, G., Lehmann, J., & Ives, Z. (2007). DBpedia: A nucleus for a web of open data. In *The semantic web, 6th international semantic web conference, ISWC '07*, Springer, pp. 722–735.

Banko, M., Cafarella, M. J., Soderland, S., Broadhead, M., & Etzioni, O. (2007). Open information extraction from the web. In *Proceedings of the international joint conference on artificial intelligence, IJCAI '07*.

Boguraev, B., Pustejovsky, J., Ando, R., Verhagen, M. (2007). *TimeBank evolution as a community resource for TimeML parsing. Language Resources and Evaluation 41*, 91–115.

Bollacker, K., Evans, C., Paritosh, P., Sturge, T., & Taylor, J. (2008). Freebase: A collaboratively created graph database for structuring human knowledge. In *Proceedings of the 2008 ACM SIGMOD international conference on management of data*, New York, NY, USA, pp. 1247–1250.

Chin, S. C., Street, W. N., Srinivasan, P., & Eichmann, D. (2010). Detecting Wikipedia vandalism with active learning and statistical language models. In *Proceedings of the 4th workshop on information credibility, WICOW '10*, ACM, New York, NY, USA, pp. 3–10.

Dean, J., Ghemawat, S. (2008). MapReduce: Simplified data processing on large clusters. *Communications of the ACM 51*, 107–113.

Ferschke, O., Zesch, T., & Gurevych, I. (2011). Wikipedia revision toolkit: Efficiently accessing wikipedia's edit history. In *Proceedings of the 49th annual meeting of the Association for Computational Linguistics: Human language technologies. System demonstrations*, Portland, OR, USA, pp. 97–102.

Fleiss, J. L., Levin, B., & Paik, M. C. (2004). *The measurement of interrater agreement* (pp. 598–626). New York: Wiley.

Gabrilovich, E., & Markovitch, S. (2007). Computing semantic relatedness using Wikipedia-based explicit semantic analysis. In *Proceedings of the 20th international joint conference on artificial intelligence, IJCAI '07*, pp. 1606–1611.

Geiger, R. S., & Ribes, D. (2010). The work of sustaining order in Wikipedia: The banning of a vandal. In *Proceedings of the 2010 ACM conference on computer supported cooperative work, CSCW '10*, ACM, New York, NY, USA, pp. 117–126.

Hoffmann, R., Zhang, C., & Weld, D. S. (2010). Learning 5,000 relational extractors. In *Proceedings of the 48th annual meeting of the Association for Computational Linguistics, ACL '10*, Association for Computational Linguistics, Stroudsburg, PA, USA, pp. 286–295.

Hu, X., Zhang, X., Lu, C., Park, E. K., & Zhou, X. (2009). Exploiting Wikipedia as external knowledge for document clustering. In *Proceedings of the 15th ACM SIGKDD international conference on knowledge discovery and data mining, KDD '09*, ACM, New York, NY, USA, pp. 389–396.

Itakura, K. Y., & Clarke, C. L. A. (2009). Using dynamic markov compression to detect vandalism in the Wikipedia. In *Proceedings of the 32nd international ACM SIGIR conference on research and development in information retrieval, SIGIR '09*, ACM, New York, NY, USA, pp. 822–823.

Lange, D., Böhm, C., & Naumann, F. (2010). Extracting structured information from Wikipedia articles to populate infoboxes. In *Proceedings of the 19th ACM international conference on information and knowledge management, CIKM '10*, pp. 1661–1664.

Milne, D., & Witten, I. H. (2008). Learning to link with Wikipedia. In *Proceedings of the 17th ACM conference on information and knowledge management, CIKM '08*, ACM, New York, NY, USA, pp. 509–518.

Mintz, M., Bills, S., Snow, R., & Jurafsky, D. (2009). Distant supervision for relation extraction without labeled data. In *Proceedings of the joint conference of the 47th annual meeting of the ACL and the 4th international joint conference on natural language processing of the AFNLP: Volume 2—Volume 2*, Association for Computational Linguistics, ACL '09, Stroudsburg, PA, USA, pp. 1003–1011.

Mola-Velasco, S. (2010). Wikipedia vandalism detection through machine learning: Feature review and new proposals. Notebook papers of CLEF 2010 labs and workshops .

Nguyen, D. P. T., Matsuo, Y., & Ishizuka, M. (2007). Exploiting syntactic and semantic information for relation extraction from Wikipedia. In *IJCAI workshop on Text-Mining & Link-Analysis*, TextLink '07.

Nguyen, T., Moreira, V., Nguyen, H., Nguyen, H., Freire, J. (2011). Multilingual schema matching for wikipedia infoboxes. *Proceedings of the VLDB Endowment 5*(2), 133–144.

Ponzetto, S. P., & Strube, M. (2007). Deriving a large scale taxonomy from Wikipedia. In *Proceedings of the 22nd national conference on artificial intelligence* (Vol. 2), AAAI Press, pp. 1440–1445.

Potthast, M. (2010). Crowdsourcing a Wikipedia vandalism corpus. In *Proceeding of the 33rd international ACM SIGIR conference on research and development in information retrieval, SIGIR '10*, ACM, New York, NY, USA, pp. 789–790.

Potthast, M., & Holfeld, T. (2011). Overview of the 2nd international competition on Wikipedia vandalism detection. In V. Petras, P. Forner & P. Clough (Eds.), *Notebook papers of CLEF 11 labs and workshops*.

Potthast, M., Stein, B., & Gerling, R. (2008). Automatic vandalism detection in Wikipedia. In *Proceedings of the IR research, 30th European conference on advances in information retrieval, ECIR'08*, Springer, Berlin, pp. 663–668.

Potthast, M., Stein, B., & Holfeld, T. (2010). Overview of the 1st international competition on Wikipedia vandalism detection. In *Notebook papers of CLEF 2010 labs and workshops*.

Smets, K., Goethals, B., & Verdonk, B. (2008). Automatic vandalism detection in Wikipedia: Towards a machine learning approach. In *WikiAI'08: Proceedings of the workshop on Wikipedia and Artificial Intelligence: An evolving synergy*.

Stvilia, B., Twidale, M. B., Smith, L. C., & Gasser, L. (2005). Assessing information quality of a community-based encyclopedia. In *Proceedings of the international conference on information quality, ICIQ 2005*, pp. 442–454.

Suchanek, F. M., Kasneci, G., & Weikum, G. (2007). YAGO: A core of semantic knowledge. In *Proceedings of the 16th international conference on world wide web, WWW '07*, ACM, New York, NY, USA, pp. 697–706.

Verhagen, M., Gaizauskas, R., Schilder, F., Hepple, M., Moszkowicz, J., & Pustejovsky, J. (2009). The TempEval challenge: Identifying temporal relations in text. *Language Resources and Evaluation 43*, 161–179.

Völkel, M., Krötzsch, M., Vrandecic, D., Haller, H., & Studer, R. (2006). Semantic Wikipedia. In *Proceedings of the 15th international conference on world wide web, WWW '06*, ACM, New York, NY, USA, pp. 585–594.

Voss, J. (2005). Measuring Wikipedia. In *Proceedings of the international conference of the international society for scientometrics and informetrics (ISSI)*, Vol. 10, Stockholm.

Wang, Y., Zhu, M., Qu, L., Spaniol, M., & Weikum, G. (2010). Timely YAGO: harvesting, querying, and visualizing temporal knowledge from Wikipedia. In *Proceedings of the 13th international conference on extending database technology, EDBT '10*, ACM, New York, NY, USA, pp. 697–700.

West, A. G., & Lee, I. (2011). Multilingual vandalism detection using language-independent and ex post facto evidence—Notebook for pan at clef 2011. In *CLEF (Notebook papers/labs/workshop)*.

West, A. G., Kannan, S., & Lee, I. (2010). Detecting Wikipedia vandalism via spatio-temporal analysis of revision metadata? Tech. rep., University of Pennsylvania, New York, NY, USA.

Wilkinson, D. M., & Huberman, B. A. (2007). Cooperation and quality in Wikipedia. In *Proceedings of the 2007 international symposium on Wikis, WikiSym '07*, ACM, New York, NY, USA, pp. 157–164.

Wu, F., Weld, D.S. (2007). Autonomously semantifying Wikipedia. In *Proceedings of the sixteenth ACM conference on conference on information and knowledge management, CIKM '07*, ACM, New York, NY, USA, pp. 41–50.

Wu, F., & Weld, D. S. (2010). Open information extraction using Wikipedia. In *Proceedings of the 48th annual meeting of the Association for Computational Linguistics, ACL '10*, Association for Computational Linguistics, Stroudsburg, PA, USA, pp. 118–127.

Wu, Q., Irani, D., Pu, C., & Ramaswamy, L. (2010). Elusive vandalism detection in Wikipedia: a text stability-based approach. In *Proceedings of the 19th ACM international conference on information and knowledge management*, ACM, pp. 1797–1800.

Xu, S., Yang, S., & Lau, F. C. M. (2010). Keyword extraction and headline generation using novel word features. In *Proceedings of the twenty-fourth AAAI conference on artificial intelligence, AAAI 2010*, AAAI Press.

Yamangil, E., & Nelken, R. (2008). Mining Wikipedia revision histories for improving sentence compression. In *ACL 2008, Proceedings of the 46th annual meeting of the Association for Computational Linguistics*, June 15–20, 2008, Columbus, Ohio, USA, Short Papers, pp. 137–140.

Yatskar, M., Pang, B., Danescu-Niculescu-Mizil, C., & Lee, L. (2010). For the sake of simplicity: Unsupervised extraction of lexical simplifications from Wikipedia. In *Proceedings of the conference of the north American chapter of the Association for Computational Linguistics, NAACL*, pp. 365–368.

Ye, S., Chua, T. S., & Lu, J. (2009). Summarizing definition from Wikipedia. In *Proceedings of the joint conference of the 47th annual meeting of the ACL and the 4th international joint conference on natural language processing of the AFNLP: Volume 1—Volume 1*, Association for Computational Linguistics, Stroudsburg, PA, USA, ACL '09, pp. 199–207.

Zanzotto, F. M., & Pennacchiotti, M. (2010). Expanding textual entailment corpora from Wikipedia using co-training. In *Proceedings of the COLING-Workshop on the peoples web meets NLP: collaboratively constructed semantic resources*.

Zeng, H., Alhossaini, M. A., Ding, L., Fikes, R., & McGuinness, D. L. (2006). Computing trust from revision history. In *Proceedings of the 2006 international conference on privacy, security and trust: Bridge the gap between PST technologies and business services, PST '06*, ACM, New York, NY, USA.

Zhang, Q., Suchanek, F. M., Yue, L., & Weikum, G. (2008). TOB: Timely ontologies for business relations. In *11th international workshop on the web and databases*, WebDB.