TrueWeb: A Proposal for Scalable Semantically-Guided Data Management and Truth Finding in Heterogeneous Web Sources

Amgad Madkour
Purdue University
West Lafayette, USA
amgad@cs.purdue.edu

Walid G. Aref
Purdue University
West Lafayette, USA
aref@cs.purdue.edu

Sunil Prabhakar
Purdue University
West Lafayette, USA
sunil@cs.purdue.edu

Mohamed Ali
University of Washington
Tacoma, USA
mhali@uw.edu

Siarhei Bykau
Bloomberg L.P.
New York, USA
sbykau@bloomberg.net

ABSTRACT

We envision a responsible web environment, termed TrueWeb, where a user should be able to find out whether any sentence he or she encounters in the web is true or false. The user should be able to track the provenance of any sentence or paragraph in the web. The target of TrueWeb is to compose factual knowledge from Internet resources about any subject of interest and present the collected knowledge in chronological order and distribute facts spatially and temporally as well as assign some belief factor for each fact. Another important target of TrueWeb is to be able to identify whether a statement in the Internet is true or false. The aim is to create an Internet infrastructure that, for each piece of published information, will be able to identify the truthfulness (or the degree of truthfulness) of that piece of information.

CCS CONCEPTS

• **Information systems → Data management systems**; **Semi-structured data**; **Uncertainty**; **Collaborative and social computing systems and tools**; *Data provenance*; *Incomplete data*; *Temporal data*; *Inconsistent data*;

KEYWORDS

Data Management, RDF, Linked Data, Truth detection

ACM Reference Format:
Amgad Madkour, Walid G. Aref, Sunil Prabhakar, Mohamed Ali, and Siarhei Bykau. 2018. TrueWeb: A Proposal for Scalable Semantically-Guided Data Management and Truth Finding in Heterogeneous Web Sources. In *SBD'18: Semantic Big Data , June 10, 2018, Houston, TX, USA.* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3208352.3208357

1 INTRODUCTION

Lot of data is being published in the Internet over various publication and media venues. The truthfulness of this published data cannot be established easily, leaving space for rumor spreading, and propagation of false information that may negatively impact businesses, political regimes, public health, etc. As people interact with each other and with their surrounding environment on a daily basis, their behaviors, beliefs and actions are highly impacted. Similarly, as users manipulate (read, share and comment on) posts on social media, the users' mindsets, attitudes and responses shape, crystallize or even change. Social media is an immense source of news and reports. From a trustworthiness perspective, social media contains both trustworthy news sources and rumor spreaders [2]. Consequently, the credibility of various posts on social media is hard to evaluate. It is a challenge to extract credible pieces of news out of a mixture of news coming from sources with variable degrees of trustworthiness.

In building TrueWeb, we focus on the following objectives: (1) Realizing a semantically-guided system for knowledge graphs and (2) Truth Finding Techniques. The results of TrueWeb are expected to have a potential impact in society. While our use cases are mostly drawn from validating sentences, they can still be extended for other cases such as prediction of crimes,climate change, accidents or applications that require background knowledge for decision making. We anticipate that the new technologies developed inside TrueWeb will be readily usable by many news agencies, activists, enthusiasts or scientists. We believe that news agencies can leverage such a system to help them validate news, tending topics or rumors as they spread instantaneously.

2 PROPOSED ARCHITECTURE

We envision TrueWeb, a responsible web environment that will offer organizations or entities the capability of verifying statements based on factual information and truthfulness and credibility of the entities that are known in the system. Building TrueWeb focuses on the following objectives: (1) realizing a semantically-guided system prototype for knowledge graphs, and (2) truth finder techniques.

Figure 1 illustrates the TrueWeb high-level architecture. It is composed of two main components: (1) A Semantically-Guided System Prototype For Knowledge Graphs,namely Knowledge Cubes [31]

SBD'18, June 10, 2018, Houston, TX, USA
© 2018 Association for Computing Machinery.
ACM ISBN 978-1-4503-5779-1/18/06…$15.00
https://doi.org/10.1145/3208352.3208357

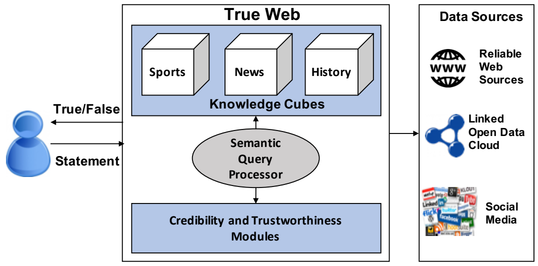

Figure 1: TrueWeb Architecture

and (2) Credibility and Trustworthiness modules. Both the Knowledge Cubes and Credibility And Trustworthiness components are loosely coupled and act as complement to each other in the TrueWeb architecture. The Knowledge Cubes rely on reliable data sources in order to create high quality factual information. The Knowledge Cubes also relies on the Linked Open Data cloud [1] in order to enrich its semantics. On the other hand, the Credibility and Trustworthiness modules relies on social media data. The Semantic Query Processor lies at the heart of TrueWeb, where it collaborates with the user, the Knowledge Cubes and the Credibility and Trustworthiness modules. It is responsible for understanding the semantics of the user query and in turn queries the Knowledge Cubes and Credibility and Trustworthiness modules when validating statements.

2.1 Targets

2.1.1 Social Media News. Social media provides a vast amount of information which contains some important facts or observations. News agencies aim at monitoring the mainstream social networks (such as Facebook or Twitter) and extracting valuable messages, posts or tweets which can be used as a source for its news articles. The main challenge in this context is to distinguish real facts and gossips or intentionally false evidences (e.g. vandalism). Moreover, the above task requires a data processing at a large scale of millions of social media messages, hundreds of thousands of news articles and billions of web pages.

2.1.2 User Reported News. Some news agencies, e.g., Al Jazeera, may ask their readers to report news and then uses those reports to provide up-to-date coverage of events and accidents. The main problem is that reports may contain inaccurate or even false information. For example, a user reports a traffic accident at Grant St. First, we need to verify if there is an accident at all or the user reports some false information. Second, we need to assess the quality of his/her report accuracy, e.g., whether there is an accident at Grant St. or maybe it is at Salisbury St.

3 SEMANTICALLY-GUIDED SYSTEM FOR KNOWLEDGE GRAPHS

We propose to develop and prototype for a semantically guided system for the management of knowledge graphs. We adopt the notion of Knowledge cubes (KC, for short) for the prototype [31]. Each KC is responsible for a certain semantic topic, e.g., sports, US presidents, or certain geographical regions. Data extracted from the Internet will be directed towards the relevant KCs for further investigation and scrutiny, and then for integration with the rest of the knowledge already existing in the KCs. A KC is an unsupervised and adaptive database instance of knowledge capable of storing, analyzing, and searching linked-data components in the form of RDF triplets. These RDF's will link to other KCs to provide further insight to the data.

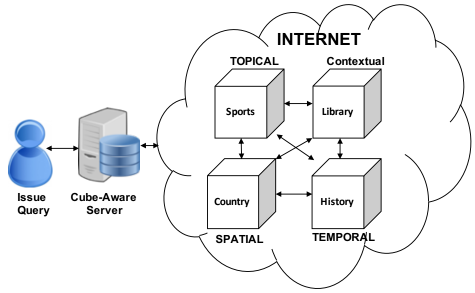

Figure 2: Knowledge Cubes High Level Architecture

As in Figure 2, a knowledge cube captures the notion of specialization, be it spatial, temporal, topical, contextual, or otherwise. Its granularity level adapts to represent finer grained knowledge cubes. For example, the "Sports" knowledge cube can be broken into multiple cubes, e.g., "Basketball" and "Baseball". These two sub-cubes would still be inter-linked together. This interlinking is important as it can provide more insightful information. Cube-awareness provides structural or data-level updates to the hosting cube. This component guarantees to the hosting instance the most updated information that other cubes possess.

Based on the nature of the data in the knowledge cube, it then dictates what other information should be aggregated, after coordinating with the remaining knowledge cubes. This communication is made possible as linked data can use HTTP as a pointer system for accessing the negotiated representation of resource/entity descriptions. The data sources for knowledge cubes are either self-initiated or distributed through a central source. In both cases, KC's coordinate with each other the content acquired in order to maximize decoupling of knowledge and hence increase their knowledge specialization.

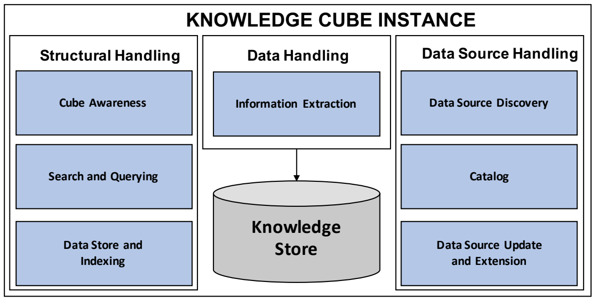

Figure 3: Knowledge Cubes Instance Architecture

Figure 3 illustrates the architecture for a knowledge cube instance. The catalog maintains all the information related to the data

[1] http://linkeddata.org/

sources it fetches. The catalog maintains a list of the data elements that each source captures. The information extraction component employs text analysis techniques in order to extract and learn from structured and unstructured sources [8]. The search and query component provides a rich set of constructs that understand the semantics of the Knowledge Cube instance. The search and querying component supports search-based methods that include keyword, textual, spatial, temporal, and semantic-based methods that rely on similarity functions and database-level methods. The Semantic Query Processing component of TrueWeb collaborates with the search and querying component of the Knowledge Cubes in order to infer if a certain statement is true or false. The Data Store and Indexing component provides the highly efficient and scalable storage and indexing mechanisms that operates over billions of RDF model triples. Knowledge Cubes is agnostic the type of Knowledge Store (whether its a graph database, triple-store, RDBMS) and the choice depends on providing scalability and performance guarantees. The Discovery of Data sources component provides a discovery mechanism for the list of data sources relevant to the Knowledge Cube semantic aspects. Finally, the Data Sources Update and Extension component create a time-oriented snapshot of the current knowledge store data. The time-roadmap guarantees "freshness" of the results by providing or linking information with the latest updates found from sources.

3.1 Research Challenges

3.1.1 Knowledge Cube Construction: This task will investigate techniques for the realization of knowledge cubes. We will employ text analysis techniques [8] in order to extract RDF data (i.e. subject-predicate-objects) from textual resources that will be used to construct the knowledge cubes. We will create an ontology to represent the domains of interest to the user. The ontology will allow us to understand the semantics of the underlying knowledge cube and the users queries. The co-occurrence of entities within textual resources exposes interesting implicit relations [33]. Modeling these relations can reveal valuable correlations among the topics [32]. We plan to associate a spatial dimension to non-spatial topics that in turn can help answer investigative queries that are not possible to answer otherwise. The relationships among spatial and non-spatial topics will be discovered automatically from textual resources and aggregated to the knowledge cube.

3.1.2 Semantic Query Processor: Semantic query processing [29, 30] is the main engine for TrueWeb. It will take as input the information that needs to be verified and outputs whether it is true or not along with an associated confidence value. The target is to develop a semantic query processor that operates on top of the knowledge cube architecture given the user’s query or the web statement under investigation. The semantic query process will make heavy use of the credibility and trustworthiness modules. Query processing will employ a set of predicates that operate on the knowledge cube attributes such as its topical, spatial, temporal, and contextual aspects to validate a given statement under investigation or respond to a user’s query. Challenges in realizing the semantic query processor include deciding on the order of execution of the query predicates that may touch [25].

4 TRUTH-FINDING TECHNIQUES

To establish the TrueWeb, every single entity, e.g., user, news reporter, and organization, is tagged with a dynamically changing trustworthiness score. Also, every post is tagged with a dynamically changing credibility score [49, 50] to reflect how far this post is believed to be true. The trustworthiness and the credibility scores that assess the truthfulness of an entity and the credibility of a post, respectively, cannot be single scalar values. The number of posts (either true or false) that an entity deliberately initiates relative to the number of posts it shares, edits, or comments on is another dimension. Social media can be modeled as a network graph, where each entity is a node in the graph. Edges between nodes in the graph represent the "follower of" or "friend of" relationship between entities, as in Twitter and Facebook, respectively. Based on how entities respond to a post, our proposed framework for a TrueWeb adjusts the credibility scores of the post. Meanwhile, as entities respond, share, and comment on posts with various degrees of credibility, the proposed framework adjusts the trustworthiness scores of these entities. These adjustments are a continuous process as the posts hop from one entity to another in the social media graph. Realizing a factual knowledge cube will enable us to validate pieces of information that need to be investigated for truthfulness [50]. The Knowledge cube can only be meaningful if it maintains a set of reliable and well established facts. We aim to employ a set of data quality techniques that ingests the extracted information and produce high quality knowledge.

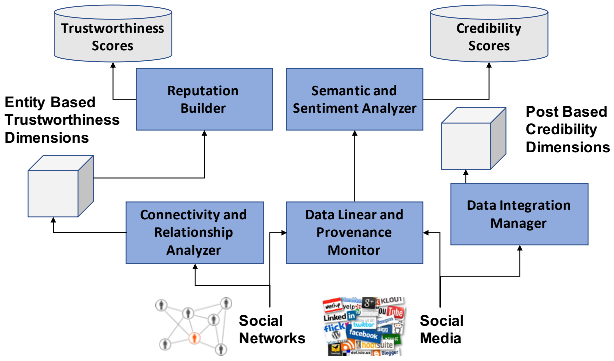
Trustworthiness Scores
Credibility Scores
Reputation Builder
Semantic and Sentiment Analyzer
Entity Based Trustworthiness Dimensions
Post Based Credibility Dimensions
Connectivity and Relationship Analyzer
Data Linear and Provenance Monitor
Data Integration Manager
Social Networks
Social Media

Figure 4: Credibility and Trustworthiness Architecture

Figure 4 illustrates the architecture of the Credibility and Trustworthiness module. The objective of the system is to build and continuously maintain credibility and trustworthiness scores for posts and entities, respectively, in a social network. The connectivity and relationship analyzer takes as input the social network graph and analyzes the connectivity and relationships among entities. Without looking into the posts that are initiated or shared by these entities, the connectivity and relationship analyzer builds several dimensions of the trustworthiness score that are computed solely based on the graph connectivity, e.g., the number of followers of an entity is computed based on the graph connectivity regardless of the entity’s activities. The data integration manager considers all posts (or the subset of posts under consideration) in the social network, correlates these posts together and decides on several dimensions in the credibility score that are computed solely on the post regardless of how it was handled by entities. For example, a semantic analysis of the post and related posts reveals whether the post is a fact or an opinion, past or future event. The data lineage and provenance monitor tracks posts as they hop from one entity to another and how they are handled by entities [11, 12]. The "semantic and sentiment analyzer" helps decide on the reaction an entity showed in response to a post. Hence, the system continuously updates the credibility cue of a post on the entities response pattern. Meanwhile, and as time proves the correctness or incorrectness of a post, the Reputation Builder revisits the entire trustworthiness cube if the entities that shared that post to elevate or de-elevate their trustworthiness score across all dimensions.

4.1 Research Challenges

4.1.1 Semantic Interpretation and Conflict Detection Techniques Under Heterogeneity: A truth finder deals with a large number of sources that provide the semantic data about same or similar objects in the form of RDF triples. A number of issues arise within this data. First, often the data about the same object can be represented in different ways depending on who the author of that data, when that data was produced, what is the context of it, etc [11]. Another example of heterogeneity is the different ranges of values (int vs double). Second, data can be seen as a semantic RDF graph where nodes are interconnected by associations and have primitive data type attributes. In this case, if some of the parts of the semantic RDF graph contains false information, then this affects the accuracy of its neighbors. This task will extend the existing truth finder algorithms [50] by modeling various representation choices as possible worlds and build algorithms that can find the most-likely state of conflicting data that maximizes the observations seen so far. We plan to study the correlations among concept attributes and investigate the ways to use these correlations to improve the accuracy of a truth finder and to detect semantics-based conflicts among the underlying data. Web sources may also provide conflicting data. We plan to develop new techniques for detecting truth in heterogeneous conflicting data sources by resolving the conflicts through the ground factual knowledge of the Knowledge Cube along with validating the meta-data of the web source using the credibility and trustworthiness modules.

4.1.2 Detection of Source Independence and Conflict of Interest: Rumor spreads easily across social networks [49]. However, a rumor tends to be initiated by few sources that are likely to be correlated. We plan to develop techniques to capture the conflict of interest among entities. The conflict of interest can be discovered through the network of relations of an entity or through the content of the post. Detecting conflict through the content requires creating a "profile" for each entity that would be representative of the entity's views. A conflict of interest would happen when an entity posts content that conflicts with his/her profile. Detecting conflict through the network of relations requires monitoring and assessing the similarity in behavior among the entire network of an entity. We will cluster the users using their social media features, and then identify the ones in the same cluster (with similar network or content) as having a conflict of interest, i.e., being dependent.

4.1.3 Assessment of User Proactiveness, Reactiveness, and False Proactiveness: This research task addresses the ability to classify users into three categories, (a) proactive sources: this category represents the original sources of the post, e.g., the news reporter who report the event while being at the field (b) reactive sources: this category represents the entities that compile their posts from other sources, yet, with proper citations, and (c) false proactive sources, this category represents the entities that initiates a post that is based on posts from other sources without proper citations. False proactiveness can be intentional to elevate one's trustworthiness score or unintentionally due to the lack of professionalism of the entity. This sub-classification is also relevant to the reputation building module.

4.1.4 Influence-based Entity Ranking: To increase the effectiveness of truth finding, we plan to leverage feedback provided by the users. For example, one can ask the user to confirm or reject some of the facts found by the truth finder. With the help of user feedback, the truth finder takes into account the updated information and improves its accuracy on other facts. Typically, the truth finder is expected to deal with a large number of facts that makes it feasible to collect feedback on only a very small fraction of the facts. In other words, having a limited budget on how many times we can ask the user to provide us with feedback the goal is to find such objects which would boost the effectiveness of the truth finder most of all. The above problem is difficult since we need to consider a large number of objects with complex dependencies among them (e.g. validating one object may lead to changes in others). To address the above problem, our task is to design and implement a framework and algorithms to solicit user feedback to improve the accuracy. We plan to take advantage of the voting relationships and dependencies among facts and sources and be able to efficiently determine accurate fact validation orders within a given limited budget.

4.1.5 Semantic-based Analysis and Classification Techniques: One of the important challenges in TrueWeb is the semantic-based analysis and classification of messages. We need classify a post based on its being a fact or an opinion, a prediction (forecast) or a fact that already took place, a Spam/advertisement, an opinion flip, a sentiment (like/dislike), a joke, similarity to another fact (describing the same info as another fact but using different keywords). These classifications are important as they can reflect on the credibility of a post as well as the trustworthiness of an entity in a variety of ways. For example, a post being an "opinion" is less damaging to the trustworthiness of the post's poster than a faulty fact. Facts can be past facts or future speculation. The classification of a user post as a past fact versus a future speculation is crucial in the reputation builder. Incorrect posts about events that happened in the past would mark the posting entity as less trustworthy. However, a degree of incorrectness in posts that refer to tentative events in the future may be acceptable without imposing a large penalty on the trustworthiness of the posting entity.

5 RELATED WORK

The problem of creating an efficient architecture to mediate knowledge linkage and abstraction [5, 6, 8, 18] contains various proposals ranging from loosely coupled to tightly coupled architectures. Franklin et al. [18] proposed an abstraction for databases that attempt to avoid data integration problems. Berners Lee et al. [5] proposed Linked Data which is about using the Web to connect related data that has not been previously linked, resulting in an architecture referred to as Web of Data. Other architectures are realized through learning such as NELL [8].

The problem of storage and indexing of distributed datastores was studied from multiple dimensions including Triplestores [15, 22], vertically-partitioned tables [1] [44] [16] and schema-specific systems [28]. Each category has its strengths and drawbacks based on the kind of domain of the data. All the datastores attempt to provide an efficient storage mechanism while maintaining a reasonable RAM and database ratio.

Up to our level of knowledge, the problem of implicit spatial learning has not be touched upon before in the literature. The task mainly involves semantic relational learning. Some of its applications include semantic search and querying and efficient query processing. Semantic relational learning has been tackled in many studies [24][21]. The methods range from knowledge engineering and deduction to data mining and induction. Reasoning focuses only on description logics and Horn rules which lacks the expressibility needed to reason over spatial relations. Semantic relationships can be automatically learned from various areas without the need of a manually annotated dataset. For example, relations could be learned from the search queries and the their results [40] or from textual patterns and tables [46]. Linked Data also provides a rich framework for establishing such relations but still need to accommodate for expressing such relations in terms of their expressiveness and uncertainty [39] [43] [20].

In terms of search and querying, it includes various studies discussing how to capture the query semantics or intent [7, 14, 29]. Most of the approaches propose mechanisms to capture the true query intent. Others rely on capturing the semantic similarity between items in order to provide better answers to semantic queries [4, 17, 23, 30, 33]. Optimizing the search process has been studied from a high level as well as from a system level. Reddy et al. [38] tackle the challenge of SPARQL queries optimization over the web of data. Tsialiamanis et al. [47] propose heuristics for a SPARQL query optimizer where they propose a Heuristic SPARQL Planner (HSP) that exploits the syntactic and the structural variations of RDF triples in a SPARQL query. On a system level optimization aspect, recent surveys [25, 34] present various classifications of RDF-based systems ranging from relational [42], No-SQL [10], to cloud-based systems [25]. Data partitioning across multiple machines has been extensively studied. Efficient Heuristic approaches [35, 41, 51] are employed in order to partition RDF data.

The problem of data quality has been studied extensively in the recent years. It encompasses different subproblems such as data veracity, vandalism detection, stability, reputation and trust and controversy detection. The problem of *veracity* (or conforming to truth) on the Internet in the presence of several sources providing conflicting values has been studied extensively in recent years [12, 19, 36, 49]. The problem pertains to identifying the true value for an object when a multitude of values are available. A naive method for this problem is majority voting that simply considers the most frequently provided value to be true. However, this is a flawed approach in a scenario where most of the sources report outdated values, sources copy from a source that initially provided a wrong value or simply because most of the sources thought a false value was true. To address the above problems, a number of different types of methods have been used: a historically first approach [27] was based on Web link analysis, then methods based on Bayesian analysis [11, 12, 50] emerged and found to be highly effective in truth detection.

Vandalism detection [9, 37, 45] focuses on identifying those edits which are performed in order to intentionally destroy the user-generated content (e.g. a Wikipedia page). Chin et al. [9] identified different kinds of vandalism based on the type of action (delete, insert, change, revert), the type of change (format, content) and the scale of editing (mass deletion or mass insertion). Machine learning techniques were used to classify edits as either blanking, large-scale editing, graffiti and misinformation, using text features such as the number of known words, perplexity value, number of known bigrams and so on.

For *stability prediction*, Druck et al. [13] examine how different features affect the longevity of an edit including whether an author is registered or not, the total number of reverts done by the author, the addition of links, etc. Based on this, they train a classifier to label edits in one of three categories: revert, 6-hour longevity and 1-day longevity.

The *reputation and trust* of user-generated content were studied in [3]. The authors proposed an algorithm to compute the reputation values of authors where each author increases her reputation if her edits are not changed by the subsequent edits. This is interpreted as a sign of the approval that the edit should remain on the page and thus the user who did it deserves that her reputation grows. In their subsequent work [2], the authors leverage the author reputations to compute trust levels for content, where the assumption is that a high reputation author is more likely to contribute trustable content. Trust is evaluated based on the algorithm's ability to predict stable content having longevity. This work led to an online system[2] which indicates the trust level of Wikipedia content using colors.

The problem of *controversy detection* has been only studied at the page level, i.e. ranking web pages as controversial or not. A machine learning approach was used by [26] in order to identify the amount of conflict on a page using the human-labeled controversy tags as a ground truth. They employed such features as the number of revisions, unique editors, page length and so on. A more statistical approach was proposed in [48]. It uses the statistics of deletions and insertions along with the Mutual Reinforcement Principle whereby frequently deleted content is considered more controversial if it appears on a page whose controversy level in the past was low, and if its authors were involved in fewer past controversies. Formulas to compute the level of controversy of a page are derived based on these assumptions.

[2] http://www.wikitrust.net/

6 CONCLUSION

We envision TrueWeb as an oracle for validating the truthfulness of sentences. We plan to study predictive queries based on the structured knowledge available in TrueWeb. These predictions can include queries about possible crime locations, climate changes, accidents and so forth. The predictive aspect can have a spatial and temporal dimensions.

We also plan to investigate extending the provenance of TrueWeb where we can indicate whether a sentence was true given a specific possible world scenario. This can give the user a form of "timeline" to understand how the sentence facts changed across time instead of denoting a statement as true or false only. This possible worlds lineage can also be used to investigate complicated rumor spreading behaviors that are not captured by the current TrueWeb prototype.

We also plan to investigate utilizing the current TrueWeb prototype in order to discover entities that can be masquerading as different individuals over the web. This can allow us to identify and remove false information that were in a way related to these masquerading entities and thus creating a more reliable TrueWeb.

REFERENCES

[1] Daniel J. Abadi, Adam Marcus, Samuel Madden, and Kate Hollenbach. 2009. SW-Store: a vertically partitioned DBMS for Semantic Web data management. *VLDB J.* 18, 2 (2009), 385–406.

[2] B. Thomas Adler, Krishnendu Chatterjee, Luca de Alfaro, Marco Faella, Ian Pye, and Vishwanath Raman. 2008. Assigning trust to Wikipedia content. In *WikiSym.* 26:1–26:12.

[3] B. Thomas Adler and Luca de Alfaro. 2007. A content-driven reputation system for the wikipedia. In *WWW.* 261–270.

[4] Andrea Ballatore, Michela Bertolotto, and David C. Wilson. 2012. Geographic knowledge extraction and semantic similarity in OpenStreetMap. *Knowledge and Information Systems* (Oct. 2012).

[5] Tim Berners-Lee. 2006. Linked Data - Design Issues. (2006).

[6] Christian Bizer, Tom Heath, and Tim Berners-Lee. 2009. Linked Data - The Story So Far. *Int. J. Semantic Web Information Systems* 5, 3 (2009), 1–22.

[7] Liliana Calderon-Benavides, Cristina Gonzalez-Caro, and Ricardo Baeza-Yates. 2010. Towards a Deeper Understanding of the User's Query Intent. *SIGIR* (2010).

[8] Andrew Carlson, Justin Betteridge, Bryan Kisiel, Burr Settles, Estevam Hruschka Jr., and Tom Mitchell. 2010. Toward an Architecture for Never-Ending Language Learning. In *AAAI.*

[9] Si-Chi Chin, W. Nick Street, Padmini Srinivasan, and David Eichmann. 2010. Detecting Wikipedia vandalism with active learning and statistical language models. In *WICOW.* 3–10.

[10] Philippe Cudre-Mauroux, Iliya Enchev, Sever Fundatureanu, Paul Groth, Albert Haque, Andreas Harth, Felix Leif Keppmann, Daniel Miranker, Juan F. Sequeda, and Marcin Wylot. 2013. NoSQL databases for RDF: An empirical evaluation. *ISWC* (2013), 310–325.

[11] XL Dong, L Berti-Equille, and Divesh Srivastava. 2013. Data Fusion: Resolving Conflicts from Multiple Sources. *Handbook of Data Quality* (2013).

[12] Xin Luna Dong, Laure Berti-Equille, and Divesh Srivastava. 2009. Integrating Conflicting Data: The Role of Source Dependence. *VLDB* 2, 1 (Aug. 2009), 550–561.

[13] Gregory Druck, Gerome Miklau, and Andrew McCallum. 2008. Learning to Predict the Quality of Contributions to Wikipedia. *AAAI* 8 (2008), 983–1001.

[14] MJ Egenhofer. 2002. Toward the semantic geospatial web. *SIGSPATIAL* (2002), 1–4.

[15] Orri Erling. 2010. Directions and Challenges of SemData. In *VLDB Industry position paper.*

[16] David C. Faye, Olivier Cure, and Guillaume Blin. 2012. A survey of RDF storage approaches. *ARIMA J.* 15 (2012), 11–35.

[17] Miriam Fernández, Iván Cantador, Vanesa López, David Vallet, Pablo Castells, and Enrico Motta. 2011. Semantically enhanced Information Retrieval: An ontology-based approach. *Web Semantics: Science, Services and Agents on the World Wide Web* 9, 4 (Dec. 2011), 434–452.

[18] Michael J. Franklin, Alon Y. Halevy, and David Maier. 2005. From databases to dataspaces: a new abstraction for information management. *SIGMOD Record* 34, 4 (2005), 27–33.

[19] Alban Galland, Serge Abiteboul, Amélie Marian, and Pierre Senellart. 2010. Corroborating Information from Disagreeing Views. In *WSDM.* ACM, New York, NY, USA, 131–140.

[20] Fausto Giunchiglia, Biswanath Dutta, Vincenzo Maltese, and Feroz Farazi. 2012. A Facet-Based Methodology for the Construction of a Large-Scale Geospatial Ontology. *Journal on Data Semantics* 1, 1 (March 2012), 57–73.

[21] Alon Halevy, Peter Norvig, and Fernando Pereira. 2009. The Unreasonable Effectiveness of Data. *IEEE Intelligent Systems* 24, 2 (2009), 8–12.

[22] Andreas Harth, Katja Hose, and Ralf Schenkel. 2012. Database techniques for linked data management. In *SIGMOD Conference.* 597–600.

[23] Krzysztof Janowicz, Carsten Keß ler, Mirco Schwarz, Marc Wilkes, Ilija Panov, Martin Espeter, and B Boris. 2009. Algorithm , Implementation and Application of the SIM-DL Similarity Server. *GEOS* (2009).

[24] Krzysztof Janowicz, Simon Scheider, Todd Pehle, and Glen Hart. 2012. Geospatial semantics and linked spatiotemporal dataâĂŞPast, present, and future. *Semantic Web* 0 (2012), 1–13.

[25] Zoi Kaoudi and Ioana Manolescu. 2015. RDF in the clouds: a survey. *VLDB Journal* 24, 1 (2015), 67–91.

[26] Aniket Kittur, Bongwon Suh, Bryan A. Pendleton, and Ed H. Chi. 2007. He says, she says: conflict and coordination in Wikipedia. In *CHI.* 453–462.

[27] Jon M. Kleinberg. 1999. Authoritative Sources in a Hyperlinked Environment. *J. ACM* 46, 5 (Sept. 1999), 604–632.

[28] Justin Levandoski and Mohamed Mokbel. 2009. RDF Data-Centric Storage. In *ICWS.* 911–918.

[29] Lipyeow Lim, H Wang, and Min Wang. 2009. Semantic queries in databases: problems and challenges. *CIKM* (2009).

[30] Lipyeow Lim, Haixun Wang, and Min Wang. 2013. Semantic Queries by Example Categories and Subject Descriptors. *EDBT* (2013), 347–358.

[31] Amgad Madkour, Walid G. Aref, and Saleh Basalamah. 2013. Knowledge cubes: A proposal for scalable and semantically-guided management of Big Data. In *2013 IEEE International Conference on Big Data.* 1–7.

[32] Amgad Madkour, Walid G. Aref, Mohamed Mokbel, and Saleh Basalamah. 2015. Geo-tagging Non-spatial Concepts. (2015), 31–39.

[33] Christoph Mülligann, Krzysztof Janowicz, Mao Ye, and WC Lee. 2011. Analyzing the spatial-semantic interaction of points of interest in volunteered geographic information. *Spatial information theory* (2011).

[34] M. Tamer Özsu. 2016. A Survey of RDF Data Management Systems. *arXiv* (2016).

[35] Nikolaos Papailiou, Ioannis Konstantinou, Dimitrios Tsoumakos, Panagiotis Karras, and Nectarios Koziris. 2013. H2RDF + : High-performance Distributed Joins over Large-scale RDF Graphs. *BigData* (2013).

[36] Jeff Pasternack and Dan Roth. 2012. Latent credibility analysis. In *WWW.*

[37] Martin Potthast, Benno Stein, and Robert Gerling. 2008. Automatic vandalism detection in Wikipedia. In *Advances in Information Retrieval.* Springer, 663–668.

[38] B R Kuldeep Reddy and P S Kumar. 2010. Optimizing SPARQL queries over the Web of Linked Data. *SemData@VLDB* (2010).

[39] Dave Reynolds. 2009. Uncertainty Reasoning for Linked Data. In *URSW.* 85–88.

[40] Stefan Riezler, Yi Liu, and Alexander Vasserman. 2008. Translating queries into snippets for improved query expansion. (Aug. 2008), 737–744.

[41] Kurt Rohloff and Richard E. Schantz. 2010. High-performance, massively scalable distributed systems using the MapReduce software framework. *PSIEtA* (2010), 1–5.

[42] Sherif Sakr and Ghazi Al-Naymat. 2009. Relational Processing of RDF Queries: A Survey. *SIGMOD Record* (6 2009), 23–28.

[43] Hamid Haidarian Shahri. 2010. Semantic Search in Linked Data: Opportunities and Challenges. *Twenty-Fourth AAAI Conference on Artificial …* (2010).

[44] Lefteris Sidirourgos, Romulo Goncalves, Martin L. Kersten, Niels Nes, and Stefan Manegold. 2008. Column-store support for RDF data management: not all swans are white. *PVLDB* 1, 2 (2008), 1553–1563.

[45] Koen Smets, Bart Goethals, and Brigitte Verdonk. 2008. Automatic vandalism detection in Wikipedia: Towards a machine learning approach. In *AAAI Workshop on Wikipedia and Artificial Intelligence: An Evolving Synergy.* 43–48.

[46] Partha Pratim Talukdar, Marie Jacob, Muhammad Salman Mehmood, Koby Crammer, Zachary G. Ives, Fernando Pereira, and Sudipto Guha. 2008. Learning to create data-integrating queries. *Proceedings of the VLDB Endowment* 1, 1 (Aug. 2008), 785–796.

[47] Petros Tsialiamanis, Lefteris Sidirourgos, Irini Fundulaki, Vassilis Christophides, and Peter Boncz. 2012. Heuristics-based query optimisation for SPARQL. In *EDBT.* 324.

[48] Ba-Quy Vuong, Ee-Peng Lim, Aixin Sun, Minh-Tam Le, and Hady Wirawan Lauw. 2008. On ranking controversies in wikipedia: models and evaluation. In *WSDM.*

[49] Dong Wang, Tarek F. Abdelzaher, Lance M. Kaplan, and Charu C. Aggarwal. 2013. Recursive Fact-Finding: A Streaming Approach to Truth Estimation in Crowdsourcing Applications.. In *ICDCS.* 530–539.

[50] Xiaoxin Yin, Jiawei Han, and Philip S. Yu. 2007. Truth Discovery with Multiple Conflicting Information Providers on the Web. In *KDD.* ACM, New York, NY, USA, 1048–1052.

[51] Kai Zeng, Jiacheng Yang, Haixun Wang, Bin Shao, and Zhongyuan Wang. 2013. A distributed graph engine for web scale RDF data. *VLDB* (2 2013), 265–276.