A Framework for Collecting and Managing Entity Identity Information from Social Media

Research in progress

Debanjan Mahata[1,a], John Talburt[1]

[1]University of Arkansas at Little Rock, United States

[a]dxmahata@ualr.edu

Abstract

An entity in general may be defined as an object that has a distinct, independent and self-contained existence. An entity could be a person, place, product, real-life event, or anything that has an individual identity. The identity of an entity is a set of attributes that distinctly characterizes it and differentiates it from all other entities. With the popularity of social media, there has been voluminous growth in the digital footprints of different types of entities in the Internet. The references to different types of entities in social media have the potential to provide extremely valuable information to researchers and organizations, which could be mined and analyzed for making major decisions. There are tremendous applications in the areas of real-life event analysis, opinion mining, reference tracking, online advertising, recommendation engines, cyber security, fraud detection, enterprise data integration, among others. Thus, there is a need of a generic framework that can collect different entity references, extract identity information of the entities from them and maintain the information persistently for resolving new references of entities. The presented research establishes the preliminary design of such a framework from the perspective of *Entity Identity Information Management* (EIIM) in the domain of social media. The paper introduces the problem of EIIM in social media, discusses the prevalent challenges and proposes the design of a framework capable of managing persistent identity information of pre-specified set of entities. We further explore the applications of the research and conclude with the future plan of action for implementing an EIIM system in the realm of social media.

Keywords: entity resolution, social media, entity identity management, entity identity information management, entity identity structure, information integration, social media data integration, enterprise data integration, ontology, semantic web.

Introduction

Social media has brought a paradigm shift in the way people communicate with each other. It has gone from being just a medium to a global medium of communication between people. It has provided a communication platform to the masses enabling them to post short real-time messages in the form of micro-blogs, status updates, photographs and videos, to write full length articles expressing their views in blogs. This has turned the information consumers to original information producers and curators. According to a recent survey reported by Pew Research about 46% of adult Internet users post original photos or videos online that they themselves have created [1]. The humungous volumes of dynamic user-generated real-time data from social media provide great opportunities to businesses, governments, and researchers to tap valuable meaningful information for further analysis.

The task of Information Extraction (IE) is to identify instances of a particular pre-specified class of entities in

natural language texts, and the extraction of the relevant properties of the identified entities [2]. An entity in general may be defined as an object that has a distinct, independent and self-contained existence, whether hypothetical or real. Thus, an entity could be a person (e.g `Barack Obama'), a place (e.g `Little Rock'), a product (e.g `Iphone6'), an event (eg `Egyptian Revolution') or anything from real-life that has an individual identity. The identity of an entity is a set of attribute values for that entity along with a set of distinct rules that allow that entity to be distinguished from all other entities of the same class in a given context [3].

The identity information of an entity depends upon the type of entity. For example, the geographical co-ordinates could be identity information for a particular place/location. Named entities, popular keywords, timespan of occurrence and place of occurrence could be identity information for a real-life event as shown in Figure 1. Similarly, name, age, gender, social media profiles, URL of the homepage, friend list, work history and location could be identity information for a person having a digital presence in social media. Any source of information that mentions about a concerned entity is known as a reference to that entity. The micro-blog posts shown in Figure 2 are all references to different entities. A major problem is to identify the references from social media and map it to a particular entity. The dynamic, diverse, colloquial and unstructured nature of the content produced in social media makes this a challenging task. This process is also known as entity resolution, or record linking, or reference tracking. The identity information of an entity when captured and stored could help in resolving the references and map them to the entity. One of the main aims of this research is to track the references of a given entity, and resolve new references to it by maintaining persistent and evolving data structures containing identity information of the entity.

Figure 1. Identity information for real-life event, 'Egyptian Revolution 2011'.

- #iPhone battery life driving you crazy? (YES!) Here are a few simple fixes. http://ow.ly/uNMHb
- #iphone An earlier iPhone 6 launch would help Apple avoid Samsung's threat: A few days back, DigiTimes publish... http://yhoo.it/1g5VQjn
- #sochi #sochi2014 - Warrior-athletes Bring Home Sochi Gold - Department of Defense: Department of DefenseWarri... http://bit.ly/1du9RTg
- @SarahPalinUSA rocked the house at #CPAC We led the "Run Sarah Run" chant https://www.youtube.com/watch?v=oO6C5UwwJTA#t=13 ... 20min #Palin2016 pic.twitter.com/0q6K8Bv2QD
- If u r not here, u r missing a lot. Down wt mobarak, Mohandesin #jan25
- Mubarak Mubarak! The plane is waiting for you
- Police just arrested everyone in cilantro. Took their ids and phone. We barely got out. #jan25
- In past 8 days at least 12 #Egyptians set themselves on fire out of desperation: unemployment, poverty, corruption. #Jan25 #Egypt protest

Figure 2. Entity mentions in micro-blog posts (the mentioned entities are underlined).

With the popularity of social media there has been proliferation of unstructured textual information about different types of entities, in the Internet. Given the valuable uses of tapping the opinions expressed by users in social media about company products [4], live reporting of events [5], discussions on socio-political issues [6] and analysis of user interests [7], there is a need to identify and track the references to an entity of interest

from unstructured content. A framework with capabilities of collecting, extracting and managing entity identity information from social media is proposed in this paper that solves the problem of persistently tracking entity references from social media. Such a framework could further help in real-life event analysis [8, 9], improving customer engagement [10], tracing terrorist attacks [11], analyze election campaigns [12], tracking the need of humanitarian aid during natural disasters [13], and predicting stock markets [14], among other possibilities.

An entity resolution system that creates and maintains persistent data structures representing the identities of entities in an information system have already been proposed by [16]. The process is known as Entity Identity Information Management (EIIM) and is an essential component for any system that provides persistent entity identifiers, i.e. entity identifiers that do not change over time. However, EIIM has been implemented over archived structured datasets. Given the popularity of social media and highly valuable uses of entity identity information from user generated content in different social media platforms, motivated the presented research. The research proposes a framework that builds upon the concept of EIIM and paves a path of its implementation in the realm of social media.

In the following sections, the paper describes the current EIIM process, which acts as the basic foundation of the research. Next, we define the problem of EIIM in social media and explore the challenges prevalent in the problem domain. The major contribution of the paper is a proposed generic framework for implementing EIIM in social media. Various applications of the research are discussed next. Work related to the research is presented highlighting the need and significance of our contributions in the backdrop of the current state-of-art. We conclude by stating the future directions and plans for the ongoing research.

Current Entity Identity Information Management

In this section we explain the current EIIM process that lays the foundation and acts as a background of the presented research.

The idea of ***Entity Identity Information Management*** (EIIM) as defined by [15] is the collection and management of identity information of real-world entities with the goal of sustaining entity identity integrity. Their model of EIIM was motivated by the problem of entity resolution in information systems, particularly in the domain of MDM (Master Data Management). They define *entity resolution* as the process of determining whether two references to real-world objects in an information system are referring to the same object, or to different object [16]. The EIIM life cycle as proposed by them is an iterative process that combines entity resolution and data structures representing entity identity into specific operational configurations (EIIM configurations, as shown in Figure 3), that when executed in concert, work to maintain the entity identity integrity of master data over time. The EIIM framework is implemented by developing open source software known as OYSTER[15].

[15] http://sourceforge.net/projects/oysterer/

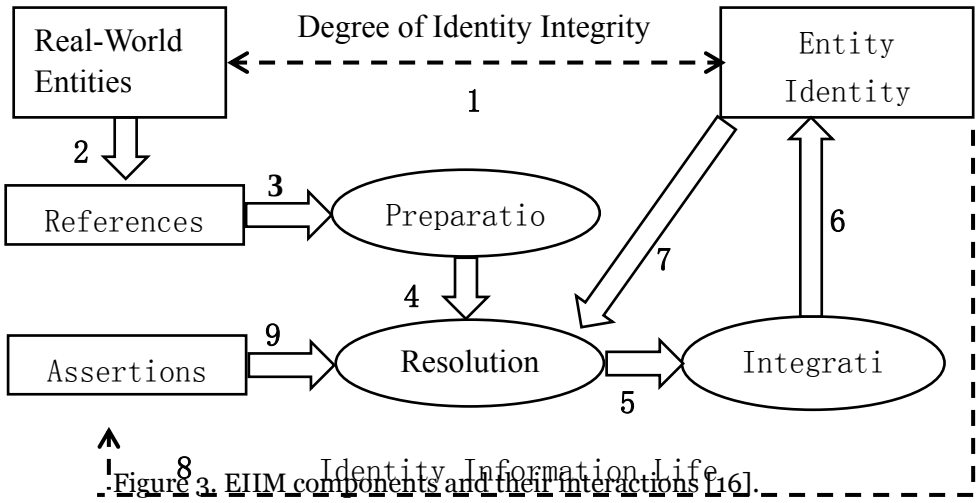

Figure 3. EIIM components and their interactions [16].

Some of the definitions as specified by the current EIIM model, which also holds for the model proposed in this research, are:

- **Definition 1.** *An **entity (e_i)** is defined as a real-life object that has a distinct identity.*
- **Definition 2.** ***Entity Identity Information** is defined as a set of attributes of a given entity that distinctly characterizes it and allows that entity to be distinguished from all the other entities maintained by the framework.*
- **Definition 3.** *An **Entity Identity Information Structure (EIIS) (s_i)**, is defined as a data structure that can persistently and efficiently store, retrieve, and manipulate entity identity information.*

One of the basic tenets of data quality that applies to the representation of a given domain of real-world entities in an information system is entity identity integrity [16]. Entity identity integrity requires that:

- Each real-world entity in the domain has one and only one representation in the information system.
- Distinct real-world entities have distinct representations in the information system.

Therefore, ideally in an information system, if **E** ($\{e_1, e_2, \ldots, e_n\}$) represents a finite set of entities, **R** ($\{r_1, r_2, \ldots, r_m\}$) represents a finite set of references to the entities, and **S** ($\{s_1, s_2, \ldots, s_n\}$) represents a finite set of EIIS maintaining identity information of the entities then there should be one-to-one correspondence between the real-life entities ($\in$ E) and the EIIS ($\in$ S) representing their identity information. Also, the references ($\in$ R) of a particular entity ($\in$ E) should always map to one and only one EIIS ($\in$ S) maintaining its identity information. This is shown in Figure 4. Such a situation ensures that the condition of entity identity integrity is satisfied by the information system. One of the main aims of EIIM is to satisfy the conditions of entity identity integrity along with persistently maintaining the entity identity information.

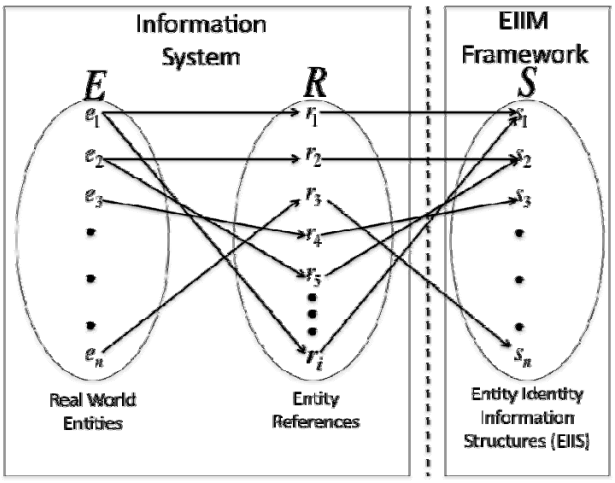

Figure 4. Entity Identity Integrity in EIIM process.

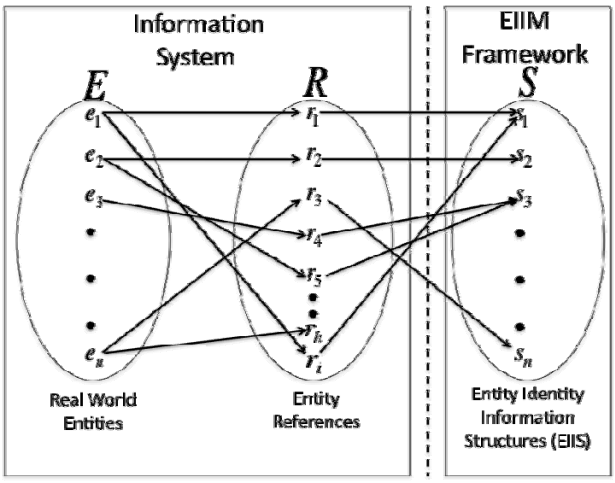

Figure 5. Misjudgments made by EIIM process.

The current EIIM model deals with a closed environment of an information system where there is fixed number of entities along with fixed number of references to them. In an ideal situation the EIIM process should always satisfy the conditions of entity identity integrity as shown in Figure 4, and previously explained. However, in practice, all the references to an entity in the information system might not get mapped to the EIIS maintained for that particular entity due to misjudgments made by the automated processes as shown in Figure 5. This might result in *false negative* and *false positive* errors. A *false negative* error arises when the system fails to map a reference of an entity to its corresponding EIIS. This is shown in Figure 5, where the system fails to map the reference r_h ($\in$ R) of entity e_n ($\in$ E) to an EIIS ($\in$ S). A *false positive* error arises when the system maps two references of different entities to a single EIIS. This is shown in Figure 5, where the system wrongly maps reference r_5 ($\in$ R) of entity e_2 ($\in$ E), to the EIIS s_3 ($\in$ S) being maintained for entity e_3 ($\in$ E). Such a situation creates dissonance between the actual identity of the real-world entities being stored in the information system and their identities interpreted by the automated processes, resulting in low entity identity integrity of the system. Asserted resolutions are introduced in order to deal with such problems (shown in Figure 3.).

The EIIM processes and life cycle is a step ahead of the basic *record linking* process that identifies references to same entities for a given dataset. The goal of EIIM is to consistently label references to the same entity with the same identifier across different datasets processed at different times. Through the management of persistent entity identity structures, EIIM provides an added functionality for an entity resolution system to create and assign persistent entity identifiers that do not change from process to process. The current EIIM can also be thought of as forming a nexus between ER and MDM by adding an explicit longitudinal dimension to the management of identity information. The EIIM model proposed in the presented research expands the current model into the unstructured domain of social media, bringing in new challenges and devising new techniques for solving them. The next section gives a detailed discussion and definition of the problem of extending the EIIM model to social media.

Problem Definition

Given the large number of social media websites producing variety of user-generated data in huge volumes and high velocity, different entities are being mentioned and discussed as shown in Figure 2. This produces multiple footprints of an entity at different venues. These footprints are considered as references to an entity

in social media. There is a need to collect these references, extract identity information from them and maintain the information persistently for resolving new references of the entities. The references could be further used for mining valuable information. All the references might not contain identity information. Also, the information being produced is dynamic in nature and evolves with time. The references might be different social media contents like micro-blogs, blog posts, photos, videos, etc. Moreover, it is also important to integrate newly extracted identity information related to an entity to the already existing identity information of that entity. This section outlines the differences of EIIM in social media with the traditional EIIM as explained in the last section, and defines the problem of developing an EIIM framework in the domain of social media.

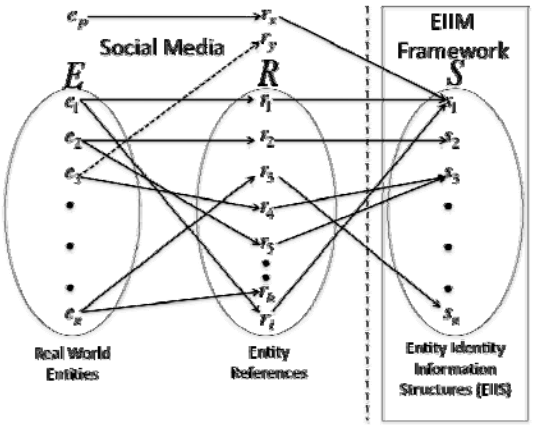

Figure 6. Relation between elements of E, R and S for EIIM in social media.

Although the basic premise for EIIM remains the same when extended to the domain of social media, but the operating environment changes. EIIM in social media does not work in a closed environment of an information system. Instead, it functions in an open and dynamic environment of social media as shown in Figure 6. Though the framework deals with pre-specified entities (E) of interest, yet there are other entities ($\notin$ E) and references ($\notin$ R). Therefore, the entities and the references to the entities are not fixed at a certain point of time. This is a challenging scenario, which might introduce new conditions as discussed next. Also, the current EIIM model has been implemented on structured data sets of relational databases. Social media data brings in new challenges associated with unstructured and real-time data as discussed in the Challenges section. Thus, it is necessary to introduce new techniques and expand the EIIM framework for making it suitable for handling the unstructured domain of social media.

In the real-world scenario, users of social media generate the references with an intention to represent information about a particular entity. For example the tweet *(#iPhone battery life driving you crazy ? {YES!} Here are a few simple fixes...)* primarily talks about the entity iPhone. There might be multiple references of the entity iPhone being generated in social media. One of the main aims of the proposed framework would be to consolidate all the references of a particular entity and map it to its corresponding EIIS maintained by the framework. However, there might be a mismatch between the way an entity is interpreted in its social media reference by the creator and the way the proposed framework interprets the same in an automated fashion. This would result in loss of data integrity and increase the chances of erroneous results. Such problems are prevalent in the vanilla flavor of the EIIM as discussed in the previous section, giving rise to *false negative* and *false positive* errors. Since the EIIM framework operates in an open environment of social media as discussed earlier, two new scenarios leading to erroneous conditions are observed:

- **Noisy Entity:** The first situation occurs when a reference gets mapped to an existing EIIS in the framework, although it refers to an entity, which is out of the pre-specified list of entities. Such a scenario is shown in Figure 6, where the entity e_p ($\notin$ E) generates the reference r_x ($\notin$ R), yet the external reference gets mapped to s_1 ($\in$ S).
- **Untracked reference:** The second situation occurs when there is a reference, which the framework is unable to track and map it to a pre-specified list of entities although it refers to it. Such a scenario is shown in Figure 6, where the framework should have tracked the reference r_y ($\notin$ R) and associate it with entity e_3 ($\in$ E), yet it is unable to do so and lose track of the reference in the EIIM process.

Problem: *Given a pre-specified finite set of real-life entities ($\boldsymbol{E = \{e_1, e_2, \ldots, e_n\}}$) generating a finite set of references ($\boldsymbol{R = \{r_1, r_2, \ldots, r_m\}}$) in social media, and a finite set of EIIS structures ($\boldsymbol{S = \{s_1, s_2, \ldots, s_n\}}$) corresponding to each real-life entity ($e_i \in E$), the problem is to resolve references of an entity (e_i), and to persistently extract, store and manage identity information of the entity in its corresponding EIIS (s_i).*

The relationships between E, R and S are shown in Figure 6. The problems that the proposed research intends to solve are:

- How to create the initial set of EIIS (S) for the entities of interest ($\in$ E)?
- How to collect reference to the entities ($\in$ E) containing identity information of the entity from different social media platforms?
- How to extract entity identity information from the social media references?
- How to integrate the entity identity information from the unstructured data platforms into the appropriate EIIS ($\in$ S)?
- How to determine that new references from social media relates to an EIIS ($\in$ S)?
- How to continuously integrate new information into the existing EIIS structures ($\in$ S)?

Challenges

Identifying, collecting and managing entity identity information from social media entails various challenges. Some of the major challenges are discussed below:

- **Information Overload:** A daily average of 58 million tweets is posted in Twitter[16]. On an average 60 million[17] photos are shared in Instagram daily. Facebook stores 300 petabytes[18] of data related to its users from all over the world. These are some compelling statistics that makes social media not only rich in volume of data, but also variety, and the velocity at which data is being generated. The search engines and content filtering algorithms often face the problem of information overload [17] due to the great pace at which data is produced in social media. They suffer from the dilemma of assessing the accuracy and quality of information content in the sources being produced over their freshness. Thus, collecting different types of references of entities from various social media platforms,

16 http://www.statisticbrain.com/twitter-statistics/

17 http://instagram.com/press/

18 http://expandedramblings.com/index.php/by-the-numbers-17-amazing-facebook-stats/

assessing their quality, resolving and extracting identity information of the entities poses great challenges in such a situation.

- **Veracity of Sources:** Judging the accuracy of the information and deciding relevant information content in social media references for the purpose of extracting entity identity attributes constitutes another challenging situation. For trending topics the search engines have started showing real-time feeds from social media websites in their search results. This has attracted spammers who post trending hash-tags or keywords along with their spam content in order to attract people to their websites offering products or services [18]. An alarming 355% growth of social spam has been reported in 2013[19]. Social media has also been instrumental in spreading misinformation and rumors. Spread of misinformation not only results in pandemonium among the users[20] but also result in extraction of completely wrong information about entities.
- **Informal Text:** Unlike sources of news media and edited documents on the web, the textual content of the social media sources are highly colloquial and pose great difficulties in extracting information out of them. One of the most important sources of information about entities, prevalent in the domain of social media are the micro-blogging platforms. Micro blogs pose additional challenges due to their brevity, noisiness, idiosyncratic language, unusual structure and ambiguous representation of discourse [22]. Variation in language, less grammatical structure of sentences, unconventional uses of capitalization, frequent use of emoticons, and abbreviations have to be dealt by any system processing social media content. Moreover, various signals of communications embedded in the text in the form of hash-tags (eg.#sochi), retweets (RT) and user mentions (@) should be understood by the system in order to extract the contextual information hidden in the text. Intentional misspellings sometimes demonstrate examples of intonation in written text [23]. For instance, expressions like, '*this is so cooool',* emphasizes stress on the emotions and conveys more information that should be captured. It has been shown that it is extremely challenging for the state-of-the art information extraction algorithms to perform efficiently and give accurate results for micro-blogs [24]. For example, named entity recognition methods typically show 85-90% accuracy on longer texts, but 30-50% on tweets [25]. Status messages in social networking websites, content in question answering websites, reviews, and discussions in blogs, and forums exhibit similar nature and present similar challenges to information extraction and text mining procedures.
- **Multiple Data Sources:** The APIs (Application Programming Interfaces) of the different social media websites returns data in different formats (JSON, XML) using different web standards (REST, HTTPS). Moreover, the information obtained from a social media website is dependent upon the type of content it produces. A video sharing website might return an entirely different set of information from a blogging website. Thus, integrating the data obtained from the various social media platforms for the purpose of extraction and tracking of entity identity information is also one of the challenges. Integrating the extracted identity information in the existing EIIS structures from time to time would be another challenging task.
- **Lack of Evaluation Framework:** Due to the lack of ground truth, new benchmark datasets needs to be created in order to evaluate the experiments. Measures of precision, recall and TWI [26] cannot

19 http://www.likeable.com/blog/2013/11/10-surprising-social-media-statistics/

20 http://www.theguardian.com/uk/interactive/2011/dec/07/london-riots-twitter

be used without modifications because of the dynamic nature of the framework and the presence of *noisy entities* and *untracked references* as explained in the previous section.

EIIM Framework for Social Media

A high level view of the EIIM components and processes for social media is shown in Figure 7. These components provide a generic framework on which any EIIM system based on social media references could be built. The various components of the framework go through cycles of interactions with each other over time, which is known as ***EIIM life cycle***. At the heart of the framework lies the *Entity Identity Information Structure* (**EIIS**), which manages the identity information related to a particular entity and also helps in resolving references to the entity. The labeled items in Figure 7 are described as follows:

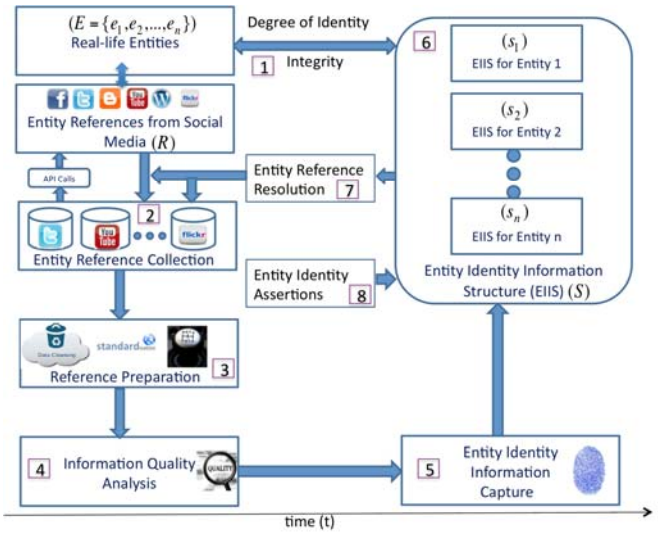

Figure 7. Entity Identity Information Management Framework for Social Media.

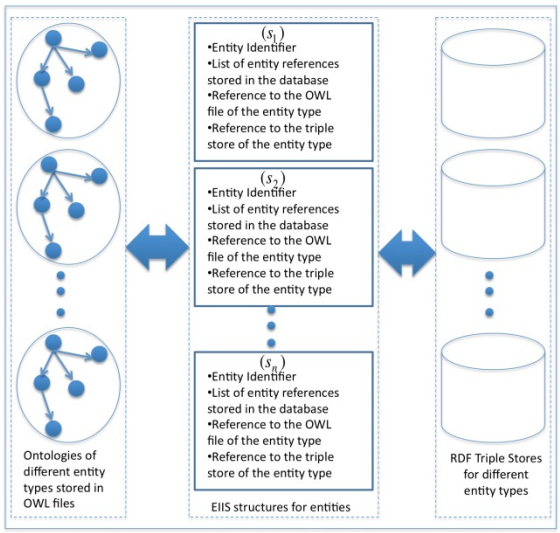

Figure 8. Close view of the EIIS component of EIIM for social media.

1. **Degree of Identity Integrity:** The fundamental goal of EIIM is to maintain a one-to-one correspondence between the real-world entities being monitored and the EIIS of the corresponding entities for ensuring entity identity integrity, as previously described. Thus, the framework maintains a separate EIIS corresponding to every individual entity it monitors. The evaluation process would be designed around the notion of measuring the degree of identity integrity of the implemented system. As new entities are introduced to the framework, a unique identifier is assigned to them along with the allocation of individual EIIS structures. The framework is expected to maintain the integrity throughout the EIIM life cycle, by consistently assigning the same identifier to the references of a tracked entity.
2. **Entity Reference Collection:** Collecting references related to an entity from social media is one of the essential components of the EIIM framework. This component allows the framework to collect entity references from different social media platforms using their respective APIs (Application Programming Interface), and store them in the database. All the fields of a reference as returned by the APIs are dumped into the corresponding table/collection for an entity in the database. Due to the semi structured nature of the files (JSON and XML) and variations in the number of fields returned by

the social media APIs, a NoSQL document oriented database management system (MongoDB[21]) is being used for storing the returned data.

3. **Reference Preparation:** Preprocessing the raw references and making it suitable for different future processing tasks is an important stage of any data intensive application. This component makes the raw entity references collected go through a number of preparation steps to improve the quality of the data. In the case of unstructured data from social media, these steps may include removal of unwanted stop words, stemming of words, standardization of meta information (like time), cleaning of embedded tags, spelling corrections, data enhancement etc. The effectiveness of these procedures can have profound impact on the success of the overall process.
4. **Information Quality Analysis:** This component deals with the quality of information present in the entity references that are collected. It segregates the high quality informative sources from the ones with lower quality. The notion of information quality might be subjective. Several approaches are taken in this stage to detect the references that might be spams and does not contain entity information. For example, tweets with occurrences of large number of links or variety of hashtags from different topics might not contain any interesting information. Similarly, blogs containing very less text, large number of links, and generic information not related to the concerned entity might be filtered out. Different approaches are taken for each social media platform depending upon its information sharing nature and structure.
5. **Entity Identity Information Capture:** This component extracts identity information of an entity from the filtered high quality informative sources and stores them in the entity's corresponding EIIS. One of the major decisions made by this module is to identify the right information that would act as an identity for the concerned entity. This might vary according to the type of entity. Depending upon the type of entity we create ontology or extend an already existing ontology (FOAF[22] for People, LODE[23] for real-life events), for their respective identity information and store them in OWL files. We plan to develop algorithms using Natural Language Processing techniques and external knowledge bases like DbPedia[24], gazetteer, Freebase[25], etc, in order to extract information from social media references. We are currently experimenting with existing information extraction tools like AlchemyAPI[26], OpenCalais[27] and Dbpedia Spotlight[28]. Although these tools work well for blog data, yet they fail to perform efficiently with references having short texts, as already discussed in the Challenges section. We are working towards devising new techniques for efficiently extracting entity identity information for different types of entities from various social media references.

[21] http://mongodb.org

[22] http://www.foaf-project.org/

[23] http://linkedevents.org/ontology/

[24] http://dbpedia.org

[25] http://freebase.com

[26] http://alchemyapi.com

[27] http://opencalais.com

[28] http://spotlight.dbpedia.org

6. **Entity Identity Information Structure:** This is the most important component of the framework, which refers to a dynamic data structure capable of persistently storing identity information about a specific entity throughout the EIIM life cycle and tracking its references. Each entity has its own EIIS, and is assigned a unique identifier. The structure is consistently updated with new identity information related to the entity. The identity information stored in an EIIS also acts as a knowledge base for the concerned entity and is used for resolving new references to the entity. Depending on the type of entity, we use the corresponding ontology in order to extract its identity information and persistently store them in a RDF triple store. The resultant knowledge bases can be queried, reasoned and updated from time to time. All the references of an entity would be assigned the identifier of the corresponding entity it refers to, and would be pointed to by its EIIS structure. This is shown in Figure 8.
7. **Entity Reference Resolution:** The process of entity reference resolution from social media takes place in this component. The resolution process might be implemented in different ways. Classifiers could be trained in order to resolve new entity references. Matching rules could be formulated and similarity scores could be calculated. For each case a set of features are needed against which the similarities would be measured. These features are extracted from the knowledge base stored in the EIIS structure of the particular entity. Once a reference is resolved to a particular entity, it is assigned its identifier and is tracked by its EIIS. The resolved reference might contain new identity information made available by the evolving social media. For example, while tracking an event, new incidents might occur and new people might get involved. This has to be captured and updated in the EIIS structure for the event. Such an evolutionary EIIS would be able to track the entire evolution of an event and resolve newly occurring references. The performance of the resolution process would play a major role in deciding the entity identity integrity of the EIIM framework.
8. **Entity Identity Assertions:** This component helps in asserting entity identity information in the EIIS of the concerned entity by manual intervention. It allows manipulation of EIIS in order to update information or to rectify incorrect information through an interactive mode. In many scenarios it can help in solving the cold start problem when a lot of information about an entity might not be available for bootstrapping the EIIM life cycle. This component could possibly be crowd-sourced. In order to collect and manage identity information of any entity from social media for tracking its references, we might bootstrap the process by feeding the system with handpicked references. For example in order to track the references related to a particular event we use the popular hashtags for the event in order to collect the real-time feeds from Twitter. These feeds pass through the different components of the EIIM life cycle and fill in the EIIS structure of the corresponding event. We further use the created knowledge base in order to resolve new references from other social media websites, and update the EIIS structures with new information.

The entire EIIM life cycle as explained above is a controlled process for each individual entity monitored by the framework. It is bootstrapped by component 8 (Entity Identity Assertions) during its initial phase. From there onwards, an entity goes through the EIIM life cycle and passes through its different components (1-8) as shown in Figure 7, until it is terminated by human intervention.

Applications

Extracting entities from social media and managing their identity information over time through EIIM process

has wide range of applications. Some of them are discussed below:

- **Event Monitoring and Analysis:** References related to real-life events are extremely abundant in social media. Right from natural disasters such as the 'Haiti Earthquake' to international sporting events like the 'Winter Olympics' to socio-political and socio-economical events that shook the world such as presidential elections, 'Egyptian Revolution', and recessions were covered, analyzed, extrapolated and informed by social media. This prolific event-specific content in social media makes it a promising ground for performing event analytics. Platforms like seen.co[29], TwitterStand[30], twitris[31], Truthy[32], and TweetTracker[33] have developed techniques to provide analytics related to different local and global real-life events. One of the main components of each of these applications is tracking references related to the events. EIIM could be an essential component of such a system. It would help in identifying, tracking and analyzing events and its related references in an organized manner over time.
- **Opinion Mining and Review Analysis:** Every day millions of people express their opinions in social media about products and companies they like and dislike. Their communications often include thoughts about good and bad experiences with the products and services. This provides a great opportunity for companies to understand its customers and to get unbiased valuable feedback from them about their product offerings without asking them to fill out time consuming outdated surveys. EIIM when used for monitoring references of products/services from social media could be an essential component of such a system [27]. Combined with sentiment analysis, EIIM could be a powerful tool for review analysis. This would provide actionable insights to a company about its products/services, and help in efficient customer relationship management, improve their services and product development. Mining opinions related to entities could be used in many other contexts like political campaigns, socio-political studies, market behavior analysis, ecommerce applications, etc.
- **Online Advertising:** Extraction of named entities and resolving their references are used in the world of online advertising in showing contextual and sponsored advertisements. Contextual advertising refers to the placement of commercial textual advertisements within the content of a generic web page, while Sponsored Search (SS) advertising consists in placing ads on result pages from a web search engine, with ads driven by the originating query [28]. Resolving named entities in a web page helps in understanding the context of the page, placing the right advertisements and efficient indexing of advertisements. The knowledge base of entities in an EIIM system and its functionalities could be instrumental in building online advertising engines for placing relevant advertisements in web pages. Different social media websites can also use EIIM for targeted and contextual advertising. A good EIIM system could prove to be crucial for generating large revenues through an efficient online advertising system.

[29] http://seen.co

[30] http://twitterstand.umiacs.umd.edu/

[31] http://twitris.knoesis.org/

[32] http://truthy.indiana.edu/

[33] http://tweettracker.fulton.asu.edu/

- **Social Media Data Integration:** Organizations have increasingly started integrating the data available in social media with the enterprise data[34]. Social media data is most powerful when it is combined with daily transactional data and the master data to give a comprehensive view of customers, products and business conditions. Customers and products are considered as real-world entities in context of *Master Data Management*. An EIIM system capable of operating in social media could go a long way in collecting the right information about the entities of concern maintained in the enterprise databases and integrate the collected information with the already existing ones. The entity resolution aspect would further help in managing the data quality issues related to data integration. In such conditions the EIIM model proposed could be used for integrating entity information from two distinct domains of enterprise system and social media in order to gain strategic intelligence related to business of an organization. This would further help an organization in marketing, corporate communications, public relations, customer support, product development, advertising, market research, product recommendations and gaining competitive intelligence.
- **Other Applications:** Monitoring social media has become one of the essential activities of national security agencies for predicting potential threats and mass protests [29]. Social media is being used for tracking terrorism activities [11], tracking foreign policies [38], and countering cyber-attack threats[35]. Managing entity identity information over time could be crucial in such applications. Credit card companies are also increasingly becoming dependent on social media in order to build and maintain identities of their customers. The identities constructed from social media are used for calculating credit scores of the customers[36]. Identity information from social media is also used for fraud investigations in insurance, retail and healthcare domains respectively. Online payment companies like Affirm[37] is also using social signals of a customer in order to assess risks in their transactions. Startup lending companies like Moven[38] and LendUp[39] are using the activities of the borrowers in social networks in order to make their decisions of lending money. EIIM from social media could be effectively used in each of these cases.

Literature Review

Entity resolution is one of the core topics that need attention in the literature behind the research. Some of the other topics that are emphasized in the research are the tasks of entity extraction, and identity management in social media. Entity resolution has been known for five decades as the record linkage or the record matching problem in the statistics community [31]. The term Entity Resolution (ER) was first proposed in the research

34 http://www.altimetergroup.com/research/reports/social-data-intelligence

35 http://www.foxnews.com/politics/2012/02/12/us-government-looks-to-mine-social-media-to-combat-terrorist-attacks-uprisings/

36 http://www.fico.com/en/about-us/newsroom/news-releases/fico-acquires-infoglide-a-leading-provider-of-entity-resolution-and-social-network-analysis/

37 http://affirm.com

38 https://www.moven.com/

39 https://www.lendup.com/

published by the Stanford InfoLab along with a generic model known as the Stanford Entity Resolution Framework (SERF) [32]. Historically, the focus of ER has been in developing processes and algorithms for determining if two references to an entity are equivalent. The Felligi Sunter model, SERF and the Talburt Wang model [16] are some of the prominent ones for conducting ER process. All the above models were based on matching algorithms working either at the record level or attribute level, and were tested with structured data in the relational databases. With the rise of big data, the modern trend is to perform entity resolution process in humongous volumes of unstructured data and scale it horizontally [33]. Although entity resolution from social media does involve processing of unstructured data, yet it has certain nuances. Most of the work in this regard so far has been in the area of entity extraction. Our effort of developing a generic EIIM system from social media would be a pioneering effort in the field of entity resolution and would create new avenues of research. As already discussed in the applications, it would open new doors in the field of social media data integration, cyber security, online payment and social media analytics.

One of the main emphases in the realm of unstructured textual content for last two decades has been in the task of extracting named entities and categorizing them into types. Competitions like MUC (Message Understanding Conference), CoNLL (Conference on Computational Natural Language Learning) and ACE (Automatic Content Extraction) spearheaded the development of new techniques in this domain. This led to the development of sophisticated tools like Stanford NER [34], OpenNLP [35], GATE [36], LingPipe [37] and NLTK [38]. Variety of techniques ranging from hand-coded rules, automatic rules, to statistical machine learning techniques like hidden Markov models, maximum entropy and conditional random fields have been proposed. A comprehensive survey of the techniques could be found in [2, 39]. A study of various efforts in extracting information from micro-blogs could be found in [40] and a survey of named entity recognition and classification could be found in [41]. Efforts have been made by the industry in building crowd sourced knowledge bases like freebase and dbpedia for the purpose of entity extraction. A recent effort from the industry for extracting entities from social media and building scalable knowledge bases for doing so has been documented in [42, 43]. In spite of the recent efforts in the field of entity extraction and unstructured text, there is no generic framework that solves the problem of persistently collecting and managing entity identity information from social media.

Traditionally, identity resolution has been a subject of system administration and management of user identities in large organizations. For the first time [15] showed the intersection of identity management, master data management and entity resolution could be used for managing identities of real-life entities in information systems, that could further play an important role in data integration and information quality. Entity identity management in social media mainly comprises of resolving and integrating profiles of the same person in social networking websites. The FOAF project has been playing an important role in all such efforts [44, 45, 46]. A very nice endeavor has been made by the OKKAM project for integrating and managing the multiple entity identifiers in various knowledge bases across the Internet [47]. To our knowledge, we are the first to propose a framework for collecting and extracting identity information of different types of entities from social media and use the concepts of entity identity management and entity resolution for persistently managing their identities with respect to time.

Conclusion and Future Work

In this paper we introduced the idea of Entity Identity Information Management (EIIM) from social media and discussed how it could be used for tracking unstructured references of different types of entities. We explained the current EIIM model and discussed how it could be extended to the domain of social media. The

transition of EIIM from the closed world of structured information systems to the unstructured open world of social media introduced new challenges. We gave an explanation of the challenges in different sections of the paper, wherever considered necessary. The new domain of social media led us to redesign the EIIM processes and components. New EIIM components and EIIM life cycle suitable for social media references were introduced in this paper. We discussed how information extraction could be performed from the unstructured social media references to entities with the help of ontologies representing entity identities. The ontologies could be used for building knowledge bases of entity identity information, which could be further used for resolving entity references from social media. Overall, the EIIM cycle aims at ensuring persistent storage of entity identities from social media and consistent tracking of references to the entities being monitored by the EIIM framework, satisfying the constraint of entity identity integrity. We also discussed about the necessity and widespread applications of an EIIM system in social media. To our knowledge, ours is the first effort in organizing the different tenets in the fields of entity resolution, information quality, data integration, entity extraction, and semantic web into a generic EIIM framework capable of functioning in the domain of social media.

As a research in progress we plan to implement a scalable system based on the framework presented in this paper. We are interested to study the implicit problems of social media data for EIIM and devise techniques to overcome them. One of the major aims of the research would be to come up with new or extended ontologies pertinent to the domain of social media in order to represent identity information of different types of entities. We also plan to study the problem of entity resolution from social media references. The challenging scenarios and new domain would require us to design a new evaluation framework for measuring the degree of identity integrity of the system. We see this as an opportunity to develop new evaluation metrics. Work is already in progress for revamping the already existing open source software OYSTER that implements current EIIM. New modules are being developed for tuning the software to the needs of the unstructured world of big data. We wish to include the components of the presented framework into OYSTER and show its efficacy in resolving and managing unstructured entity references from social media.

References

[1] PewResearchCenter. Social networking fact sheet.

[2] Piskorski, Jakub, et al. "Information extraction: Past, present and future." Multi-source, Multilingual Information Extraction and Summarization. Springer Berlin Heidelberg, 2013. 23-49.

[3] Lim, E. P., et al. (1993, April). Entity identification in database integration. In Data Engineering, 1993. Proceedings. Ninth International Conference on (pp. 294-301). IEEE.

[4] Pang, Bo, et al. Opinion mining and sentiment analysis. Foundations and trends in information retrieval, 2(1-2):1{135, 2008.

[5] Howard N, Philip, et al. Opening closed regimes: what was the role of social media during the arab spring? 2011.

[6] Singh K, Vivek, et al. Mining the blogosphere from a socio-political perspective. In Computer Information Systems and Industrial Management Applications (CISIM), 2010 International Conference on, pages 365{370. IEEE, 2010.

[7] Banerjee, Nilanjan, et al. User interests in social media sites: an exploration with micro-blogs. In Proceedings of the 18th ACM conference on Information and knowledge management, pages 1823{1826. ACM,

2009.

[8] Mahata, Debanjan, et al. Learning from the crowd: an evolutionary mutual reinforcement model for analyzing events. In Proceedings of the 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, pages 474{478. ACM, 2013.

[9] Sheth, Amit, et al. Understanding events through analysis of social media. Proc. WWW 2011, 2010.

[10] Ajmera, Jitendra, et al. A crm system for social media: challenges and experiences. In Proceedings of the 22nd international conference on World Wide Web, pages 49{58. International World Wide Web Conferences Steering Committee, 2013.

[11] Oh, Onook, et al. Information control and terrorism: Tracking the mumbai terrorist attack through twitter. Information Systems Frontiers, 13(1):33{43, 2011.

[12] Tumasjan, Andranik, et al. Predicting elections with twitter: What 140 characters reveal about political sentiment. ICWSM, 10:178{185, 2010.

[13] Kumar, Shamanth, et al. Tweettracker: An analysis tool for humanitarian and disaster relief. In ICWSM, 2011.

[14] Zhang, Xue, et al. Predicting stock market indicators through twitter i hope it is not as bad as i fear. Procedia-Social and Behavioral Sciences, 26:55{62, 2011.

[15] Zhou, Yinle, et al. Entity identity information management (eiim). In International Conference on Information Quality (ICIQ-11), Adelaide, Australia, pages 327{341, 2011.

[16] John R Talburt. Entity resolution and information quality. Elsevier, 2011. Arkady Maydanchik. Data quality assessment. Technics publications, 2007.

[17] Paul Hemp. Death by information overload. Harvard business review, 87(9): 82{89, 2009.

[18] Benevenuto, Fabricio, et al. Detecting spammers on twitter. In Collaboration, electronic messaging, anti-abuse and spam conference (CEAS), volume 6, page 12, 2010.

[21] Chris Grier, Kurt Thomas, Vern Paxson, and Michael Zhang. @ spam: the underground on 140 characters or less. In Proceedings of the 17th ACM conference on Computer and communications security, pages 27{37. ACM,

2010.

[22] Bontcheva, Kalina, et al. Twitie: An open-source information extraction pipeline for microblog text. In Proceedings of the International Conference on Recent Advances in Natural Language Processing. Association for Computational Linguistics, 2013.

[23] Scott Prevost. An information structural approach to spoken language generation. In Proceedings of the 34th annual meeting on Association for Computational Linguistics, pages 294{301. Association for Computational Linguistics, 1996.

[24] Derczynski, Leon. Microblog-genre noise and impact on semantic annotation accuracy. In Proceedings of the 24th ACM Conference on Hypertext and Social Media, pages 21{30. ACM, 2013.

[25] Ritter, Alan, et al. Named entity recognition in tweets: an experimental study. In Proceedings of the Conference on Empirical Methods in Natural Language Processing, pages 1524{1534. Association for

Computational Linguistics, 2011.

[26] Talburt, J. & Zhou, Y. (2013). A practical guide to entity resolution with OYSTER. In Shazia Sadiq (Ed.), Handbook on Research and Practice in Data Quality, Springer, pp. 235-270.

[27] Ding, Xiaowen, et al. Entity discovery and assignment for opinion mining applications. In Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining, pages 1125{1134. ACM, 2009.

[28] Broder, Andrei, et al. A semantic approach to contextual advertising. In Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval, pages 559{566. ACM, 2007.

[29] Jeffrey Ghannam. Social media in the arab world: Leading up to the uprisings of 2011. Center for International Media Assistance, 3, 2011.

[30] W Lance Bennett, et al. Taken by storm: The media, public opinion, and US foreign policy in the Gulf War. University of Chicago Press, 1994.

[31] Winkler, William E. "The state of record linkage and current research problems." Statistical Research Division, US Census Bureau. 1999.

[32] Benjelloun, Omar, et al. "Generic entity resolution in the serf project." IEEE Data Engineering Bulletin, June 2006 Issue (2006).

[33]Lars, Kolb, et al. "Dedoop: efficient deduplication with Hadoop." Proceedings of the VLDB Endowment 5.12 (2012): 1878-1881.

[34] Finkel, Jenny Rose. "Named Entity Recognition and the Stanford NER Software." 2007-03-01]. http://nlp. stanford. edu/software/jenny-ner-2007. pdf(2007).

[35] Baldridge, Jason. "The opennlp project." (2005).

[36] Cunningham, Hamish. "GATE, a general architecture for text engineering."Computers and the Humanities 36.2 (2002): 223-254.

[37] Baldwin, Breck, et al. "LingPipe." Available from World Wide Web: http://alias-i.com/lingpipe (2003).

[38] Bird, Steven. "NLTK: the natural language toolkit." Proceedings of the COLING/ACL on Interactive presentation sessions. Association for Computational Linguistics, 2006.

[39] Sarawagi, Sunita. "Information extraction." Foundations and trends in databases 1.3 (2008): 261-377.

[40] Hua, Wen, et al. "Information extraction from microblogs: A survey." Int. J. Soft. and Informatics 6.4 (2012): 495-522.

[41] Nadeau, David, and Satoshi Sekine. "A survey of named entity recognition and classification." Lingvisticae Investigationes 30.1 (2007): 3-26.

[42] Deshpande, Omkar, et al. "Building, maintaining, and using knowledge bases: a report from the trenches." Proceedings of the 2013 international conference on Management of data. ACM, 2013.

[43] Gattani, Abhishek, et al. "Entity extraction, linking, classification, and tagging for social media: a wikipedia-based approach." Proceedings of the VLDB Endowment 6.11 (2013): 1126-1137.

[44] Bouquet, Paolo, et al. "Entity-centric Social Profile Integration." Proceedings of the International Workshop on Linking of User Profiles and Applications in the Social Semantic Web (LUPAS 2010). 2010.

[45] Bortoli, Stefano, et al. "Foaf-O-Matic-Solving the Identity Problem in the FOAF Network." SWAP. 2007.

[46] Raad, Elie, et al. "User profile matching in social networks." Network-Based Information Systems (NBiS), 2010 13th International Conference on. IEEE, 2010.

[47] Bouquet, Paolo, et al. "OkkaM: Towards a Solution to the``Identity Crisis"on the Semantic Web." SWAP. Vol. 201. 2006.