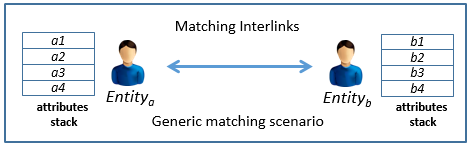
**Linking user profiles in social networks: A comparative review**

**Abstract.** Social Networking Sites (SNS) have become a treasured resource for social network analytics. Information disseminated across these platforms can be analyzed using different facets, due to the heterogeneity of the contents shared by the subscribers. The core entity that forms the SNS is the user, who typically creates a profile that represents his/her identity and that contains public and private information filled by himself/herself. Users can register themselves on one or more networks. The existence of user profiles belonging to a single user across different SNS poses several challenges to the research community. The main technical issue in that context is to detect the same user profiles across several different social networks by leveraging a set of mechanisms that identify the similarity among the user profiles. This problem is commonly referred to as *entity matching* or *identity linkage* on social networks. In this review, we describe and compare the 27 most important (to the best of our knowledge) research papers in this area. The main contributions of this article are to provide a systematic and integrated review of papers in this area, to provide comparative points that simplify the understanding of such systems, and finally to discuss future research avenues.

**Keywords:**Identity Linkage, Profile Matching, Survey, Social Networks

**1   Introduction**

The process of registering on most social networking websites is  accomplished in an easy manner, launched by the creation of a new user account. Typically, it is necessary to provide an email and a password to obtain a valid user profile. A user can later access his own profile and modify various settings, such as a profile image, location, relationship status, interests and other information. People can register themselves on several sites; moreover, the same user can create different accounts on a single platform by providing the same information for each profile while using different email accounts. So far, social networking sites did not setup any mechanisms to detect or discover if two accounts belong to a single user and to potentially merge them. This problem relates to a well-known issue in the database literature known as Record Linkage (RL). RL is the task of accurately identifying records corresponding to the same entity from one or more data sources [22, 27]. The RL process normally starts by resolving entities in a database (Entity resolution), matching them using convenient data matching techniques and finally merging similar records.



**Fig. 1.** Abstraction of Profile Matching Scenario

The scenario is similar on social networks; Figure 1 shows a generic model that gives a high-level overview of the process. Suppose that we have two social networks, social network *SNa* pcm:subscript too small and social network *SNb*, and we have a user entity *Entitya* on *SNa* and user entity *Entityb* on *SNb*. The objective is to discover whether these two profile accounts should be linked and/or merged. Approaches to user profile matching usually start by defining a set of features (matching interlinks) to link user profiles. Interlinks are categorized into attribute information or context and semantic information [38]. Attribute based matchers employ profile attributes, such as screennames, profile images, birthdates, etc. Context or semantic matchers compare the behavioral likeness of user profiles. For instance, in [12, 3] the authors use timestamps between posts across two social networks to decide if two profiles correspond to the same identity. According to [38] the matching criteria can be processed naively by treating all profiles as one set. The other method that is considered as more efficient is the blocking algorithm pcm:cite something, which suggests to partition the entity set into blocks of related entities. However, to the best of our knowledge none of the previous work on entity matching for social networks used blocking.

The research community investigated a number of matching mechanisms. Timestamp variations between user posts is a widely used behavioral feature. However, this feature can be weak if the user is mostly active on one social network only. Profile attributes like image, location, and others, might be non-updated for a long period. In the end, all matching criteria pose a set of challenges; for attribute based criteria, information can be private or distinct. User profile image can be distinct between two resources, as context information might overlap for different users holding the same names. In Table 1, we show a taxonomy of each profile matching paper including the year of publishing, the feature being used and the resources.

This paper represents the first literature survey on the topic, and covers the past efforts in profile matching for social networks. We analyzed and compared twenty seven research papers in this area. This survey is structured as follows: section 1.1 demonstrates a background definition on profile matching types. Section 1.2 illustrates the difficulties and challenges in this research field. Section 2 lists the approaches and algorithms on profile matching composed into two sub sections. Section 3 discuss the similarity metrics that have been used. Section 4 presents for each approach the results and datasets that have been used and finally section 7 proposes a novel algorithm for interlinking user profiles on social networks.

**Table 1.** Taxonomy of articles (articles above 2014 are bolded)

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| References | Year | Features | | | Resource(s)  (Flickr: FL,YouTube:YT) |
| attribute basic features | | context & activity features |
| Goga [7] | 2013 |  | ■ | | Yelp,Flickr,Twitter |
| Sha et al. [8] | **2016** |  | ■ | | TW,FB,Google+ |
| Zafarani [1] | 2009 | ■ |  | | 12 different SNS |
| Goga [9] | **2015** | ■ |  | | TW, FB,LinkedIn,FL |
| Bennacer [18] | **2014** | ■ |  | | FL, Live-Journal, TW &YT |
| Van Le [10] | **2014** | ■ | ■ | | Truong Xua |
| Raad [2] | 2010 | ■ |  | | FB, LinkedIn |
| Jain 2013 [11] | 2013 | ■ | ■ | | TW, FB |
| Liu [12] | **2014** |  | ■ | | TW, FB, Sina Weibo, Renren |
| Zafarani [3] | 2013 |  | ■ | | SNS, Blogs, forums (CMS) |
| Nunes [13] | 2012 | ■ |  | | TW, FL, LinkedIn |
| Motoyama [4] | 2009 | ■ |  | | FB,MySpace |
| Bartunov [5] | 2012 | ■ | ■ | | TW, FB |
| Vosecky [6] | 2009 | ■ |  | | FB, StudiVZ |
| Shen [14] | 2014 | ■ |  | | G+,TW,Foursquare |
| Liang [19] | **2015** | ■ |  | | TW, FB |
| Roedler [20] | **2015** |  | ■ | | TW, FL, Foursquare |
| Panchenko [17] | **2015** | ■ |  | | FB, VKontakte |
| Nguyen [16] | **2015** | ■ |  | | FB, TW |
| Peled [15] | **2016** | ■ |  | | FB, Xing |
| Jain 2015 [27] | 2007 | ■ |  | | TW, FB, Instagram, Tumblr |
| Perito [28] | 2008 | ■ |  | | Google, eBay |
| Szomszor [30] | 2010 |  | ■ | | del.icio.us, FL |
| Iofciu [31] | **2015** |  | ■ | | Delicious, FL & StumbleUpon. |
| Malhotra [34] | 2012 | ■ |  | | TW and LinkdIn |
| Zhang [37] | **2014** |  | ■ | | TW, LinkedIn |
| Vosoughi [35] | **2015** |  | ■ | | TW, TW |
|  |  |  |  |  |  |

**1.1   Profile Matching Background**

***Profile matching for identity linkage***: identity linkage is the problem of linking entities in data resources that represent the same identity, by leveraging a series of data matching mechanisms. An entity can for example be a person or a product, or any other uniquely identifiable instance. The problem is then as follows: given a social network user, the goal is to use matching functions that match several profiles on one or several social network as belonging to the same user. More specifically, the goal is to find exact profile matches in the same or different SNS by leveraging essential or auxiliary matching attributes. The significance of having more effective identity resolution approaches is growing due to the fast increase in the number of users and profiles on SNS. This can for example by used by organizations to unify or find richer contents on their users for various applications.

***Profile matching for recommendation:*** profile matching can be used to identify profiles that are similar to a given user’s profile, for instance on online dating applications [30]. The problem is to recommend similar profiles based on one profile, by finding the intersections or overlaps between various profiles. Key attributes in this context can be age, interests, location and others. This problem is usually denoted as *people-to-people* *recommendation* [31, 32]. It uses a variety of techniques such as collaborative filtering [33], content-based filtering or hybrid systems to reach a solution.

**1.2   Profile Matching: Difficulties and Challenges**

Matching two profiles across two different SNS is a difficult task. The first challenge to tackle is the varying structure of the social networks. In addition, privacy and security issues between the networks make things even more complex. For example, LinkedIn often provides publicly available user affiliations, which might be private on other platforms. This obviously renders the problem of linking two profiles by their affiliations tricky. Recently, Facebook provided a new publicly available profile attribute: profile creation date. However, many other social networks do not provide this feature. This can severely complicate the linking process. For this reason, researchers tend to use auxiliary features beyond user profile attributes: the user behavior on the platforms, or content analysis on the information they share are widely used. Profile attributes are often more accurate because they are directly contributed by the user. However, these are not always available making auxiliary features like content, behavior, and timestamp especially attractive even though they are not always present on the platform and are potentially more difficult to analyze.

**2   Profile Matching Algorithms**

We compose the profile matching algorithms into two segments. The algorithms that rely on user profile information and the others that rely on content and activity features. We imply by user profile information those that can be filled upon creating an account on a specific social network by the user himself, such as: screen names, location, age, gender, work and education. However, these attributes are not limited to the mentioned ones; it can be extended depending on the SNS being used. For instance, LinkedIn provided a profile attribute called “BIO”, that allows subscribers to write a short paragraph that summarizes themselves. The second segment is the content information, which have been used widely by researches arguing that basic profile information leads to privacy obstacles; hence, it impedes the matching process. We named this information by “content and activity”, which includes a variety of features such as behavioral information [12, 3], geo-tags [20], user messages [8] and any other information that are not extracted from limited profile attributes. Some approaches intersect the two segments, using the profile attributes as generic information that can be used to query another social network, to search for a user holding the same name. In the second phase, they leverage content information to confirm the matching of two profiles [10]. In Table 2, we show for each approach what are the used matching interlinks.

**Table 2.** Matching Features used by Each Approach

|  |  |
| --- | --- |
| Reference | Matching features (Interlinks) |
| Goga et al. [7] | Geo-location, timestamp of posts, writing style |
| Sha et al. [8] | User message (posts, tweets, retweets) |
| Zafarani et al. [1] | Usernames |
| Goga at al. [9] | Profile public-attributes |
| Bennacer et al. [18] | Network topology, public information |
| Van Le et al. [10] | Topics exists in user posts |
| Raad et al. [2] | Profile public-attributes |
| Jain et al. [11] | Public-attributes, social network, self-mentions (URIs) |
| Liu et al. [12] | Long-term behavioral analysis |
| Zafarani et al. [3] | Information redundancies in behavioral patterns |
| Nunes et al. [13] | Profile public-attributes |
| Motoyama et al. [4] | Profile public-attributes, email |
| Bartunov et al. [5] | Profile public-attributes, friendship links |
| Vosecky et al. [6] | Profile public-attributes |
| Shen et al. [14] | Public attributes, neighborhood features, quasi (inferred) features |
| Liang et al. [19] | Profile attributes, friendship links |
| Roedler et al. [20] | Timestamp of posts, device generated geo-tags |
| Panchenko et al. [17] | Usernames, friend lists |
| Nguyen et al. [16] | User public information |
| Peled et al. [15] | Profile public-attributes, network features |
| Jain et al. 2015 [27] | Historical values of attributes |
| Perito et al. [28] | Usernames |
| Szomszor et al. [30] | Tag-clouds |
| Iofciu et al. [31] | Usernames, tags |
| Malhotra et al. [34] | username, display name, location, profile image, and number of connections |
| Zhang et al. [37] | Local features: Usernames, language, URL, popularity. External features: location, avatar |
| Vosoughi et al. [35] | Language models, temporal activity |

**2.1   Approaches based on profile attributes and public information**

Goga et al. 2015 [9] propose a matching schema that takes into consideration the reliability. Hence, to model a reliable matching schema, they define four characteristics: availability, consistency, non-impersonability, and discriminability. By achieving each one of the following attributes, they can reach a reliable matching. Availability goal is to ensure that the attribute value is available on both social networks where profiles are being matched. Consistency goal is to ensure that each entity provide same or similar profile information on the different profiles he manage. Non- impersonability goal is to ensure that information provided by users is not fake, otherwise, fake information can lead to matching troubles. The second contribution of this work is, to select the training and testing data set carefully.

Raad et al. [2] is one of the earlier works who solve this problem. The goal of this research was to create a framework that finds the similarity among user profiles across different social networks. They consider solving this problem using all the user’s profile attribute. The matching mechanism starts by assigning weights to each one of the profile attributes based on its importance. To finally decide the matching occurrence, they create a decision-making algorithm and assign it this task. The authors also discriminate their work from others, by including the semantic similarity between profiles. Such semantic attribute can be topic interest and description. To perform a successful matching procedure, the authors also assign for each profile attribute a specific similarity function. Hence, they do benefit from a variety of similarity metrics that are compatible with each profile attribute. In order to perform easier matching process, they transform all the profiles into a FOAF syntax before conducting the matching. Ultimately, the interest of this matching framework lies in covering all the user profile attributes in the matching process, and stratifying for each attribute its correspondence similarity function, and assigning a suitable weight for each attribute.

Motoyama et al. [4] start by searching similar profiles using email addresses. However, if this search fails, they conduct the matching by relying on a set of public profile attributes; age, gender, location and others as a first step. The second step is to decide whether the profiles does really match, for this task they employ the boosting machine learning algorithm. In addition to the profile attributes, this research extracts attributes from profile content. For instance, they try to search for the email addresses inside the image of the user profile.

Vosecky et al. [6] goal was to identify users across multiple social networks based on profile matching. The authors use profiles from two social networks: Facebook and StudiVZ. This research published in 2009 was also one of the earlier approaches that address the profile matching problem. To achieve their target, the authors represent each profile by a vector of information. The elements of each vector are the profile attributes themselves. As [2] suggests, in this paper, the authors assign weights for each profile attribute. To decide the similarity, authors distinguish between three classes of matching. Partial, exact and fuzzy matching. Exact matching occurs with attributes that holds predefined values; such as gender. As same as for partial and fuzzy matching, each depends on the attributes. The result of matching was based on a predefined threshold. Profiles must be matched if the final similarity score is more than the threshold.

Shen et al. [14] attack the privacy concerns behind user accounts matching. The target of this research was to raise user awareness against profile matching systems. They first infer whether two accounts are linkable or not by proposing User Account Linkage Inference (UALI) method. After that, if two profiles are likely being matched, they notify users to control their profile information through proposing a method called Information Control Mechanism (ICM). To infer if a pair of user accounts linkable or not, authors have leveraged three types of features: basic profile attributes, neighborhood (like friends network) and quasi features.  Quasi features does not exist in user profile, however, they have inferred by a special algorithm.

Nunes et al. [13] propose an approach for user profile matching between social networks. They use features extracted from user profile attributes. To achieve the desired target, they employ a set of supervised machine learning algorithms: SVM, random forests and decision trees. All research experiments have been conducted on three social networks: Twitter, Flickr and LinkedIn.

Peled et al. [15] introduced an entity matching approach to match user accounts across Facebook and Xing social networks. They rely on the set of publicity available attributes crawled using iMacros software, in addition to the user’s social network. The distinction in this work from others is the use of six different machine-learning algorithms for this problem. Concluding, the LogitBoost performs the best for this task.

Jain et al. 2015 [25] Link user profiles’ accounts between Twitter, Facebook, Tumblr and Instagram. They based this matching on a novel feature by detecting the historical modifications of user’s profile attribute information. Historical values of attributes can definitely link two users with same identity. However, the authors did not discuss a universal solution for the privacy issues behind accessing historical information, which is considered as highly sensitive.

Perito et al. [28] is the only research that solve user profile linking relying only on usernames comparison. The authors claimed that usernames alone could solve this problem. The authors leverage Markov-Chains model to estimate missing characters in usernames. Two similarity methods used to perform the comparison between usernames: Edit distance and term-frequency inverse-document-frequency (tf-idf). The dataset used in this research is one of the largest in this area with a total of 10 million usernames from eBay and Google.

Nguyen et al. [16] tackled user privacy into consideration. They propose a public-information-only approach to match users between social networks. Relying on user public information just important to respect user privacy, and dealing just with his publicity available data. However, does this lead to effective matching solutions? In this context, approaches that combine public information with content are the ones that can lead to better matching results.

Bennacer et al. [18] matched social accounts across Flickr, Live-Journal, Twitter and YouTube using names, emails and links to other webpages. In addition, they define a set of rules on the aforementioned attributes to ensure the matching accuracy. This algorithm is the first to match profiles iteratively by using newfound matches in each new matching step.

Panchenko et al. [17] matched profiles between FB and VKontakte social networks. This algorithm is the first to perform big data techniques (Hadoop and MapReduce) for the matching problem. The procedure starts by comparing usernames between two user profiles. If the first two letters of their last names match, they move to the next step. In the second phase, they search the maximum number of friend sharing, and to decide the matching occurrence, they define a set of threshold. The core contribution in this research is the use of millions of profiles and big data techniques to solve the matching problem. However, there are no real contributions proposed on the level of defining new matching interlinks, which might overcome the previous ones.

Zafarani et al. [1] provide an approach for mapping user across communities. They used data collected from twelve different communities. The main goal of this research is to connect these platforms using community mappings. To achieve this task the authors rely on usernames and URL with an accuracy reached 66%.

Malhotra et al. [34] contribution behind is not novel. The authors claimed they have achieved 99% of precision through automating the matching procedure which uses a set of machine learning algorithms. The authors decide to rely on public attributes: location, usernames, connections, profile image and description attribute from LinkedIn. They employ convenient similarity function for each pair of attributes. The SN used in this research are Facebook, and LinkedIn. However, the authors did not clarify how they matched attribute description between Facebook and LinkedIn, because this attribute is only available on LinkedIn.

**2.2   Approaches based on content and activity features**

Goga et al. 2013 [7] the goal of this research is to study how potential attackers can detect if  user accounts on different social networks belong to one entity. They utilize features exist  in posted contents, mainly: geo-location of post, timestamp of posts and the writing style of the users. To predict whether two accounts matches or not, the authors employ a binary logistic regression technique that takes as an input  the score of  the similarity functions. They used features extracted from three social networks: Twitter, Flickr and Yelp. The authors prove that their approach have dominated the matching algorithms that were based only on usernames.

Sha et al. [8] take care mainly about the privacy. Authors  argue that existing methods who rely on the pairwise similarity among user profile information are vulnerable, because users might change these profile contents. Hence, they propose profile-independent feature, which is user messages, where the goal here is to attack the portion of users who are  always aware of their profile privacy and hide their identities. Authors have treated all historical user messages as vector, and compute the similarities among these vectors. The experiments have conducted on three different social networks: Twitter, Facebook and Google+. To achieve their target, authors have employed four different classification algorithms: LogisticRegresion, KNeighbors, DecissionTree and SVM.

Van Le et al. [10] propose a system for user profile matching and modelling in social networks. Despite other approaches that focused on attribute based information, this research aim at discovering the hidden topics that lies inside user generated contents, through the use of Latent Dirichlet Allocation (LDA). The overall target of this research is to suggest finally, a list of friends to the user’s current social network. In addition to the content comparison between user profiles, the authors harnessed user profile information in the matching process, such as birthday, location, interests, etc. authors have composed attributes into types, and for each type they assign the convenient equation. For instance, to compare single-valued attributes that holds single value: (male, female), the equation used here is the complete equality. The same for other type of attributes. Research experiments have examined on a Vietnamese social network called “Truong Xua”.

Jain et al. [11] target is to find user’s identity on Facebook given his identity on Twitter. To achieve this task they use profile-independent features. Two main search axis have followed. The first is the use of content information; the second is the use of user’s social network. As a start point, they launch their matching process by searching using basic profile attributes, such as username and location. Secondly, they search the wanted profile using content (tweets) from Twitter as a search query to find the desired profile on Facebook. The authors propose also a novel matching interlink which is self-mention search. Two profiles are likely to be linked if one of them at a specific social network mentions his profile URL on the second social network inside his posts. The final dimension used in the matching process, is the social networks of users. Users are likely to be matched if they have similar network connection on both social networks. All search experiments in this paper have conducted on Facebook and Twitter social networks.

Liu et al. [12] introduced HYDRA for linking same user accounts. The motivation of this research issued from a set of challenges. The main challenge was the heterogeneity of behavior modelled across different social networks. Hence, matching user accounts using basic behavioral comparison, will not led to matching succeed. Therefore, authors propose a large-scale behavior modelling approach, which model the behavior of different social networks in a long period of time, which must raise the level of consistency.  Another crucial matching key proposed by the authors, is the core network social structure of a user. The aim of structural comparison is to find the most frequently communicating friends in the same social network and matching these findings in the other social network. The idea behind this research was to recommend new social networks to the current social media users based on the aforementioned matching criteria. All research experiments have conducted in both Chinese and English social networks, seven platforms used in total for experiments.

Zafarani et al. [3] proposed attribute-independent user profile mapping approach, by exploiting redundant information that exist from user’s behavioral patterns on social media sites. They argue that  behavioral information are unique due to variety of factors, such as user personality. Hence, user personality cannot be changed which will lead to effective mapping approach. The second contribution of this work is the use of machine learning techniques to raise the efficiency of identifying user. However, a set of approaches mentioned in the previous works have also used machine learning for matching tasks. The main process of this approach starts by building a set of features extracted from user behavior, this features serve as an input to a machine-learning framework, which will resolve the matching occurrence. Authors defined taxonomy of behavioral patterns that will strongly assist in the mapping procedure.

Szomszor et al. [28] correlate user profile information between del.icio.us and Flickr, by leveraging tag-clouds used by users across these two networks. The decision of correlating users across the two networks Is done by the comparison between these tag-clouds using cosine similarity function.

Iofciu et al. [29] match users between Delicious, Flickr and Stumble Upon social networks, leveraging tags used by users and usernames. They propose threefold approach. They test their matching using tag-based only comparison, username only comparison and a combination of both. For each one they employ a set of similarity measurements. For tag comparison, they used TF-IDF and BM25. For username comparison, they used a set of text similarity functions: such as LCS.

Roedler et al. [20] leveraged timestamps between user posts and geo-tages. They hypothesize that users use their social networks simultaneously; hence, if a user for example update his status on Facebook, he will do the same on Twitter. However, many users have accounts on different platforms, but active on one platform. Geo-tags used to inspire user’s geographical area by calculating distances between these tags.

Zhang et al. [37] matched users between LinkedIn and Twitter. The main contribution of this work is to match users even if they share different username attribute. To address the matching problem they used probabilistic classifier that estimates the matching accuracy based on the pre-defined features. However, the authors did not clarify if this approach is interoperable with other social networks, for instance: Facebook do not have attribute named description or Biography, which have been used by the authors. Moreover, the authors used location comparison, while did not mention how did they extract this feature.

Vosoughi et al. [35] used language models and temporal patterns to match users from Facebook and Twitter. Language models shown very high advantage to explore the writing patterns of users, which definitely will be efficient matching link. However, temporal attribute might leads to very low precision, especially for users who are active on one social network only, despite of the other, for instance a user X has a rich-content profile on Facebook, and empty-content profile on Twitter, which will fails the temporal comparison to detect the similarity between users.

**2.3   Profile matching algorithms based on social linkage**

Few algorithms have focused on social linkage to solve the matching problem.  Liang et al. [19] defined a Greedy-based matching algorithm that discover matchings, even if two users do not share the same friend network. They define a user closeness measure to detect the matchings. The idea is to link users based on mutual friend occurrences and not based on the friend list itself. In addition, they perform initially attribute similarity between profiles. Similarly, Bartunov et al. [5] did. They combine profile attributes with social linkage to address the linking issue.

**3   Similarity Metrics**

Conducting successful and effective matching task sometimes affected by the choice of the similarity metrics. In table 3, we list some syntactic similarity methods used by the state-of-the-art. This list cannot be limited to syntactic matching methods. However, user profile information can be various, syntactic information, semantic information, image and social networks. Depending on the type of information, we choose the convenient similarity method. Four basic methods for matching have been observed during our review. An example of syntactic and semantic matching methods that can applied on textual profile attributes, are Jaro-Winkler, Jaccard similarity, Cosine similarity. Each method is optimized on a specific information structure, username comparison leads to best matches using Jaro-Winkler, cosine similarity is optimized for comparing vector information [24], Jaccard [24] is used to compare series of tokens like educational background and professional experience. Geo-distance is for location comparison. Image matching methods is used to compare user profile images, such as RGB Histogram. Graph matching methods is used to compare user’s social networks (friends, followers, friends of friends). In addition to the previous traditional matching methods, researchers can use crowdsourcing platforms that leverage human intelligence alongside with computational intelligence. Crowdsourcing particularly used when the matching problem is with high complexity; hence, we assign the resolution of two profiles to the crowd.

**Table 3.** List of Similarity Metrics used by some Approaches

|  |  |
| --- | --- |
| Reference | Similarity metrics |
| Goga [7] | Cosine distance; trained series of similarity  functions |
| Goga [9] | Jaro, geodesic, phash, SIFT, |
| Van Le [10] | Cosine similarity |
| Raad [2] | Editdistance, jaro , softTFIDF, ESA |
| Jain [11] | RGB histogram, jaro distance |
| Zafarani [3] | Jaccards, cosine, Jensen-Shannon divergence |
| Panchenko [17] | Edit distance, own dictionaries, Levenshtein Automata |
| Nunes [13] | Levenshtein, Jaro-Winkler |
| Vosecky [6] | VMN, String Distacen Score(SDS) |
| Shen [14] | Jaro-Winkler, Geodist,OpenCV |
| Jain 2015 [25] | Variety of LCS, Jaro and Jaccard similarity functions |
| Perito [26] | Edit distance, TF-IDF |
| Szomszor [28] | Cosine similarity |
| Iofciu [29] | Exact match, Jaccard, Levenshtein, Smith-Waterman similarity and Longest Common Substring (LCS) for usernames. TF-IDF and BM25 for tags |
| Malhotra [34] | Jaro Winkler, Wordnet ontologies, Jaccard, geo-distance |
| Zhang [37] | Graph-scale X2 for image comparison  Jaccard, TFxIDF Jaro-Winkler for username comparison |

**4   List of Evaluations Metrics & Datasets by Different Approaches**

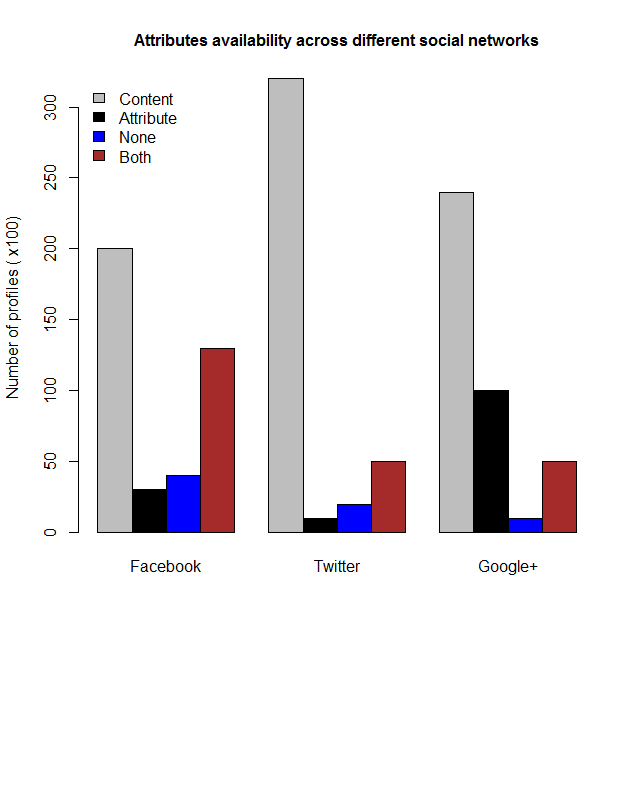
In table 4, we list for each work the evaluation metrics used and the size of dataset. The choice of the appropriate evaluation metrics commonly depends on the type of results. The area under the receiver operator curve (AUC) usually used when we use machine learning algorithms in our model, in order to evaluate the efficiency of our classifier. Precision (P) is used to find out the number of correctly matched profiles divided by the total number of profiles found. However, recall (R) is divided by the number of correct profile matched. F1-score, is used to measure the accuracy of results considering both precision and recall in the calculation. Accuracy (ACC), false positive rate (FPR) and True positive rate (TPR), are also widely used evaluation measures in the entity matching problem. ACC finds the portion of test results is classified correctly when using a machine-learning algorithm.

The largest datasets used, are by Liu [12] with 5 million profiles, Panchenko [17] used the largest dataset ever to match user from Facebook and the popular Russian social network VKontakte with 90 million profiles from the VKontakte and 3 million profiles from Facebook. Perito [26] is the largest dataset used to match profiles using only usernames: with a dataset of 10 million profiles from eBay and Google.

**Table 4.** Evaluation Metrics and Datasets

|  |  |  |
| --- | --- | --- |
| Reference | Evaluation Metrics | Dataset |
| Goga [7] | TPR, FPR | 13,629(Tw&Flc),1,889 (Tw,Yelp) |
| Sha [8] | P,R,F1-measure | TW,GP:1314,GP,FB:1576,TW,GP,FB:824 |
| Zafarani [1] | Accuracy | 36,214 usernames |
| Goga at al. [9] | P,R | Friend Finder(Dataset FF)G+(Dataset G+) |
| Bennacer [18] | P, R | 2M profiles |
| Van Le [10] | P | 507000 users |
| Raad  [2] | P, R | 50 profiles |
| Jain [11] | Accuracy, MAP | 543 users |
| Liu [12] | P, R | 5 million users |
| Zafarani [3] | Accuracy | 100,179 user pairs |
| Nunes [13] | Accuracy | 1756 pairs |
| Motoyama [4] | TPR, FPR | 1385 pairs |
| Bartunov [5] | P,R | 2K TW users,9K FB users |
| Vosecky [6] | Running time | 1980 unique users |
| Shen [14] | Recall, AUC | G+:258,TW:48,100,Foursquare:5,709 |
| Liang [19] | P, R | Bartunov [5] dataset |
| Roedler [20] |  | 2,588,981 geo-tagged tweets,  827,472 distinct TW users |
| Panchenko [17] | P,R | FB: 3M profiles,VKontakte: 90M |
| Nguyen [16] | Precision, recall |  |
| Peled [15] | AUC,TPR, FPR | 30,000 user profiles |
| Jain 2015 [25] | ACC,FNR, FPR | 8.7M profiles |
| Perito [26] | P,R | 3.5 Google,6.5 eBay profiles |
| Szomszor [28] | Improvement measurement | 502 user profiles |
| Iofciu [29] | Mean Reciprocal Rank (MRR) | 421,188 distinct users |
| Malhotra [34] | P, R,F1-measure | 883,668 users from FriendFeed  38,755 users from Profilactic |
| Zhang [37] | P,R,F1-measure  Identity based ACC | 154,249 LinkedIn profiles  9,750 matched identities |
| Vosoughi [35] | ACC ,Average rank | 11,224 FB and TW accounts |

**5 Performance Analysis: Content versus Attribute Based Approach**



**Fig. 2.** Attributes availability across different social networks

After conducting an exhaustive analysis and comparative study of different approaches, we have executed a small test on four thousand profiles to state some conclusions on which approaches performs better and why. Below are four derived baselines:

* As shown in Figure 2, we can conclude that users tend to keep their profile content public.
* Users who make their profile attributes public usually do that for content also (Figure 2 – Both)
* Attribute-only availability do not occur very often
* A number of users have shared neither content nor profile attributes about themselves, however, this constitutes the rare portion.

Hence, behavioral approaches for profile matching can still performs better than attribute ones. Indeed, the mix between the two approaches could be an optimal solution. In the upcoming section, we propose two novel attributes for each approach and show how using them can enhance each other’s.

**6   Beyond the Existing Matching Interlinks**

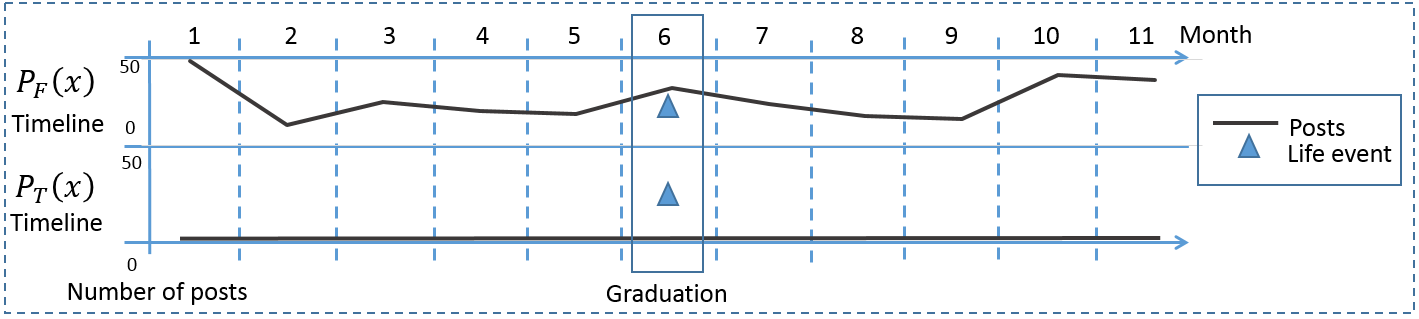
Although the aforementioned approaches have exhausted a wide variety of matching interlinks, research horizons are still promising in this research area.  This comprehensive study inspires and motivates us to propose new promising attributes that could enhance both categories (content and attribute based approaches). We tackle this problem and propose SocialMatching [39] a novel approach that leverages: (1) user *life events* such as graduation, marriage or new job, which used to enhance the behavioural approaches (2) *profile biographies*, which consist in small paragraphs that users write to comprise arbitrary information about themselves. These are used to enhance the attribute approaches. To evaluate our approach, we conducted experiments on 2,263 different profiles from Facebook matched with 5,694 Twitter users, and compared them with two baseline approaches. Our results show that SocialMatching achieves better results compared to the baselines approaches, showing that our system successfully bridges the gap between behavioural and attribute based approaches.

7 SocialMatching: Connecting identical user profiles on different social networks

We propose an approach called SocialMatching that aims at linking user profiles from Facebook to their exact profiles on Twitter leveraging two novel matching interlinks such as life events and biographies.

7.1 SocialMatching Conceptual Model

SocialMatching is composed of two phases: (1) LEBL (Life Event Based Linking) that connects profiles using life events, and (2) DEBL (DEscription Based Linking) that connects profiles based on profile biographies. Figure 3 represents a case study of SocialMatching (LEBL). It shows the user interaction on two different channels (Facebook and Twitter over time. We observe that user could have same life event (new job) mentioned on both channels, even if the content and behavior does not exist on Twitter.



**Fig. 3.** User timelines activity on Facebook and Twitter: upper layer shows high activity on Facebook, however, the under one shows no such activity on Twitter except life events. **This** **constitutes the key motivation of this research**

**Life events**

A life event post is not frequent. Which leads to the occurrence in specific circumstances, hence users who are not active on all their social networks and want to keep the audience updated, normally share these events, by publishing a post that describes them or by adding them to their timeline. In this research, life events were extracted from the users’ timelines on Facebook. Unlike Facebook, no formal representation of life events on Twitter exists. Consequently, we must perform alternative mechanisms to detect them; named entity recognition is employed to detect entities inside event posts.

**Profile descriptions (biographies)**

Profile description (biography) is an attribute that exists on both LinkedIn and Twitter. On LinkedIn, users can write a detailed biography about themselves. However, Twitter descriptions are much shorter. In 2017[[1]](#footnote-1), Facebook released a new profile attribute that allows users to define anything about themselves using characters. People can mention anything inside it (hobbies, life events, biographies, etc). Users can mention many things inside like new job, hobbies, favorite food, etc.

7.2 SocialMatching Problem Formulation

Let be a user profile on Facebook and a user profile on Twitter. is a known entity and it consists of a username () which is composed of a first name, last name, a list of life events where ( 🡨, ) and a description . For each we must match the exact user names on Twitter. (1) is the matching function and (2) the similarity function where *L* and *D* are the *L*ife events and *D*escription respectively.

(1)

is the similarity function, and is the total number of events on Facebook. We use the vector space model (cosine similarity) because we have a sequence of tokens to compare in both life events and biographies.

(2)

Two profiles are considered to be matched, if they have exact screenname matches and if the value of similarity score is higher than a predefined threshold *t*.

(3)

7.2.1 Threshold computation

We assign a threshold *t* for the similarity function. Each threshold is calculated according to the following mechanism:

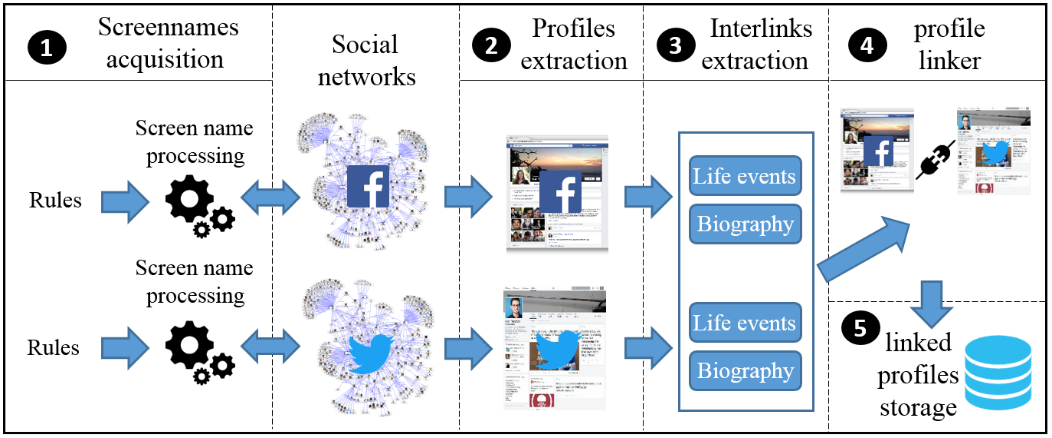
* 1000 matching candidates have been calculated on all of the corresponding similarity metrics, the threshold is obtained by calculating the average score of the truly matched pairs*.* Equation *t*s (4) Is the threshold function, where *s* is the similarity function (Cosine similarity in our case), *n* is the number of truly matched candidates and *N* is the total number of matchings (100).

(4)

The threshold function can be obtained also using further mechanisms such as machine learning algorithms, so it can learn from the average score and adopt the threshold value..

7.3 SocialMatching approach description

SocialMatching architecture is illustrated in Figure 4. (1) we start by retrieving screennames from Facebook, (2) then we extract the set of profile attributes corresponding to each user. After, we find the exact matching screennames from twitter. For each Twitter account, (3) we search for a similar life events on their timeline and compare them to those on Facebook, as well as we do the same for biographies. (4) Finally, we decide if two user profiles are linkable or not (5) and keep them stored.

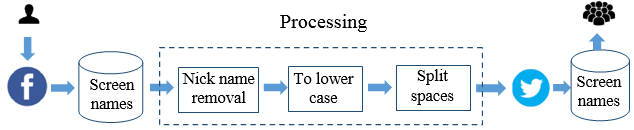


**Fig. 4.** SocialMatching Architecture

**Choosing social networking platforms**

Two social networks have been used in this approach: Facebook and Twitter. Facebook is the largest one across the world in the number of users, followed by Twitter. In our research, we detected that only 48.9% of the user profiles extracted from Facebook own a Twitter account. Facebook allows users to create structured life events that describe a certain circumstance in their life, these events are commonly published on the user’s wall (timeline) and can be public or private. Users also can publish a new update (status) that describes their life events. Contrariwise, Twitter does not provide any official feature through which users can update their life events. Twitter users can nevertheless post their own life events as Tweets. Tweets are consists of 140 characters, however, in September 2017 Twitter had doubled this limit.

**Processing Screennames and Profile Extraction**

All users want to create a profile on social network sites must define a valid first name and last name, the names that appear on user profiles are called screennames. People usually distinct their names to prevent the ambiguity with other user accounts, by modifying their usernames (a username is an id that can be accessed via the URL, e.g. facebook.com/userid). For instance, the exact screen name “Hussein Hazimeh” is available in fifty different profiles on Facebook, and each one has a different username definition. In plus, some users differentiate themselves by adding a nickname to the original screen names (e.g. Omar Abou Khaled (Professor)), or writing it in multiple languages. Screennames retrieval process starts by acquiring a set of screennames from the Facebook directory[[2]](#footnote-2), this directory contains people, pages and place names sorted alphabetically, with the URL for each entity. All screennames extracted were composed of Latin characters for both Facebook and Twitter datasets. In Figure 5, we illustrate the workflow of screenname processing. The maximum number of matches were 18 for a 

**Fig. 5.** Screennames processing workflow

Unique Facebook screenname. We remove all the nicknames and screennames that are different from Latin characters as well.

**Screen names similarity computation**

After filtering the screen names using the processing phases elaborated in figure 5, they are conducted to a matching phase using the convenient similarity metric. To reach this, we employed the Jaro-Winkler name similarity defined in (5) where *s1* is the processed screenname from Facebook and *s2*  is the processed screenname from Twitter, *m* is the number of matching characters and *t* the number of transposed ones. Jaro [6] is an algorithm commonly used for name matching in data linkage systems [40].

(5)

**Life Events Extraction**

After all screennames are manipulated carefully, and matched with the corresponding screennames on Twitter, all the life events with its dates are extracted for each username at the beginning. The maximum number of events extracted per entity was seven.

**Biography Extraction**

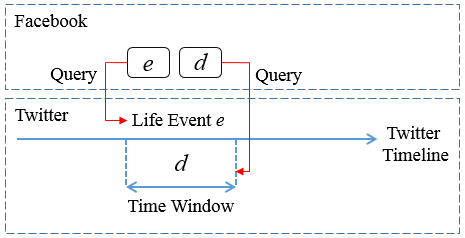
All biographies have been preprocessed prior storage. Due to the noisy content exists in them, several unwanted characters might be existing, and this could minify the matching performance. Hence, we clean all biographies by removing stop words, stemming the text, remove special characters and defining entities inside the text.

**Named Entity Extraction**

We have detected and extracted named entities from biographies and life events as well. A named entity example is a person, organization, location or a university. In life events, user can mention their work place or organization name for example, as well as in biographies, these entities can be share on biographies of two different social networks, in addition to persons. We used the state-of-the-art techniques to detect entities. Conditional Random Fields (CRFs) are widely used models. Given a specific biography or life event text, by applying CRF we can obtain and annotate all the entities inside this text.

**Life Events Searching Mechanism: A Time Window Approach**

Facebook life event is composed of a set of entities and a date. We model the timeline of a Twitter user as a series of time windows, as shown in Figure 6. A single time window is a fixed interval of time when the data stream is processed for querying. Suppose that we have a Facebook life event *e* that has a specific date *d*. Our objective is to query the time window with date *d* and try to find if we have existing similar life events to *e*. The similarity of two life events represented by the function *S(e1, e2)*. *S* is positive if one of the two following scenarios is occurred: (1) if the value of .



**Fig. 6.** Using event keywords and event date from Facebook as a query to search the twitter timeline for similar/exact events.

**Algorithm 1**: Life event Search

**Input:**

Life event : e

Date : d

Timeline : TL

Timewindow: TW

Post : p

Posts : P

Similarity\_function : S

**Output**:

Matched life event

1. **Begin**
2. TW=select P from TL
3. Where
4. TW=d
5. **Foreach** p in P
6. **Calculate**
7. S(p, e)
8. **If**(s>t)
9. Exit
10. **Endif**
11. **Endfor**

**Profile linker Workflow**

The first two inputs of the profile linker are: the Facebook profile and its corresponding Twitter profile ids. Matching problems are two-fold, the first one is (*1 to 1*) matching problem. This exists when there is only one matched screen name on Twitter. The second one is (*1 to n*) matching problem. This exists when there are more than two existing Twitter profiles. The matching process starts by comparing two biographies considering both named entities and matching score to take the matching decision. For e.g.

*PhD student at university of Leipzig* (1)

*Uni. Leipzig*(2)

In the biographies (1) and (2), the similarity score is very low. However, if we consider common entities between them, we can observe that the user writes about the university where he studies. In this situation, we decide to connect the two profiles from discarding them.

If the similarity score returns a null result, we query each user’s Twitter timeline with the named entities extracted from Facebook in a specific interval of time, Figure 6. Deeply speaking, the profile linker algorithm (Algorithm 2) gets a set of inputs as a first step as shown. The linking process is composed of two layers, the first layer starts by comparing two profiles by the simplicity attribute (biography). If the similarity score *S* is higher than the threshold value, we exit the process and connect these profiles, else we navigate into the second linking layer. In this layer, we compare profiles using the life events, matched profiles are collected using the search function in Algorithm 1, two life events are matched using the similarity function, and connected if they satisfy a threshold higher than t as well.

**Algorithm 2**: Profile linker

**Input:**

Flag=0

Facebook profile : FB\_profile

Twitter profiles list : TW\_profiles

Facebook profile biogrpahy : FB\_profile\_bio

Twitter profiles biography list : TW\_[profile\_bios]

Facebook life events named entities : FB\_life\_events\_NE

Twitter life events list : TW\_[life\_events]

Similarity\_function : S

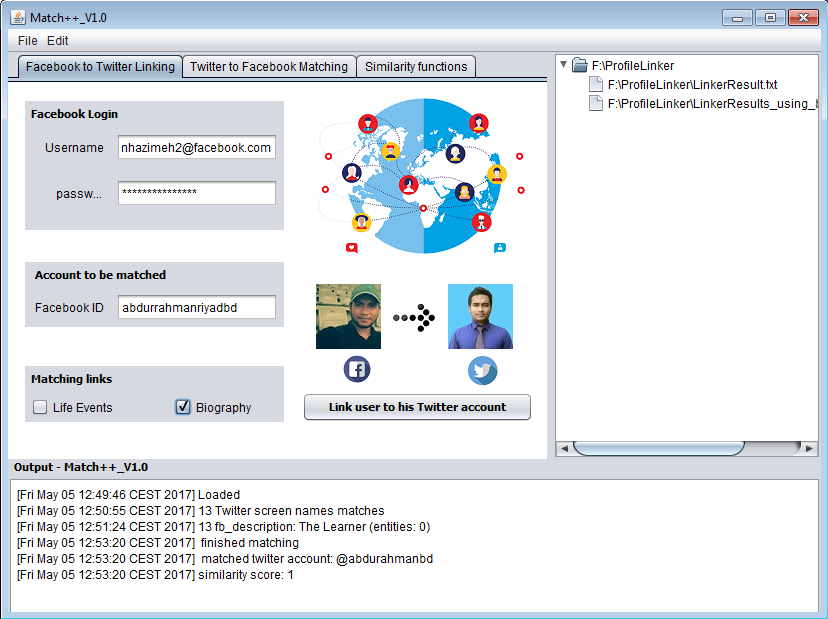
**Output**:

Two linked profiles

1. **Begin**
2. **Foreach** Twitter\_profile in Twitter\_profiles
3. **calculate**
4. S(TW\_[profile\_bio], FB\_profile\_bio)
5. **if** (S>t) then
6. Flag=1
7. **Link**(FB\_profile, TW\_profile)
8. **Exit()**
9. **End if**
10. **End for**
11. Merge (OSN-RSi. Feed)
12. **If**(Flag=0)
13. **Foreach** TW\_[life\_event] in TW\_[life\_events]
14. **Search**(TW\_timeline(TW\_[Life\_events],FB\_life\_events\_NE)
15. **If Search()** > null
16. **calculate**
17. S(search\_results, FB\_life\_events\_NE)
18. **If** S>t
19. Flag=1
20. **Link**(FB\_profile, TW\_profile)
21. **Exit()**
22. **Endif**
23. **Endif**
24. **Endfor**

8 Datasets and Experiment evaluation

In this paper, we used Facebook and Twitter platforms in our experiments. Additional social networks like Instagram can have more subscribers and much popular. However, Instagram is an image-only sharing social network, which is not enough to conduct wide matching experiments. Concerning LinkedIn, it contains much more information about users. But it is more oriented to professional’s environment, moreover, we do not have access to the user’s timeline. Therefore, even if the percentage of connected users between Facebook and Twitter is not huge, they remain the two better choices due to the richness of information tweeted and published on the user’s timelines.



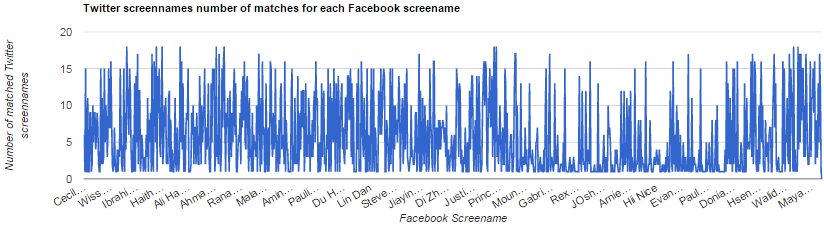
**Fig. 7.** Screenshot of Our System

Our dataset is composed of 2,263 different profiles and a total of 6,630 life events and 1948 profile descriptions from Facebook matching with 5,694 Twitter users. The built Facebook dataset available to be downloaded (upon request). All data from online social networks have been extracted using Selenium web driver, using a PC with 8GB of RAM and eight cores. The complete system code is downloadable on[[3]](#footnote-3).

**Dataset facts and profiles extraction**

The dataset collected from Facebook are collected based on the following mechanisms: (1) we select two random profiles from Facebook that have a public friend list; each one them has more than 1K friends. These lists have been extracted (2) and for each user in this list we have crawled from his profile the public life events published on his timeline and the profile description as well. For each event, we have extracted its content and the publishing date.

In Figure 8, as shown we display the total number of exact screen names matchings from Twitter with their corresponding Facebook profile screen name.

 **Fig. 8.** Screennames of Facebook variation with the number of profiles

#of profile

The maximum number of screen names matches is 18, and only 1,022 screen names were found on Twitter.

The maximum number of life events extracted from Facebook for each user is eight events. In addition, each class of event has a total of 2.2K at maximum.

**Table 5.** Life events dataset analysis

|  |  |  |
| --- | --- | --- |
| **Life event** | **Total number extracted** | **Named entities** |
| Travelled | 190 | City name |
| Started a new job | 1,481 | Company name |
| Left his job | 175 | Company name |
| Graduated from a school/university | 809 | School/university name |
| Started school | 2,261 | School name |
| Moved to a new city | 99 | City name |
| Engaged | 169 | Person name |
| Published a new paper | 1 | Paper/conference name |
| Get married | 278 | Person name |
| In a new relationship | 510 | Person name |
| Left studying at university | 363 | University name |
| Get a new award | 1 | Award name |

**Experiments evaluation**

To evaluate the performance of our approach, we used precision. Additionally, this approach is compared with other baselines.

Concerning life events, we select one state-of-the-art system, which compares the behavioural similarity between among user profiles. The system is called HYDRA [12] published in 2014. HYDRA compares also long-term behavioural activity in large-scale datasets. This system is one of the notable and novel contributions in the behavioral approaches area; based-on we decided to be chosen.

Concerning biographies, our system is compared with attribute-based approaches; we were able to prove that even if users do not share public attribute information, it is possible to connect these profiles using biographies. @I seek ‘fb.me’. [11] was the baseline compared with us due to the usage to a variety of profile attributes in this system, however, without biography.

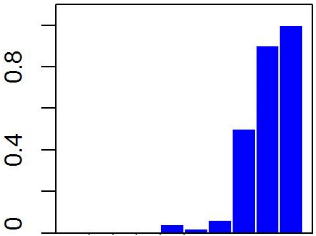
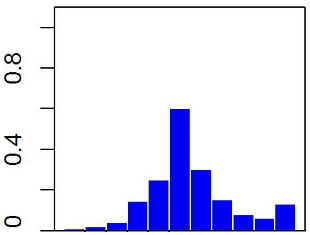
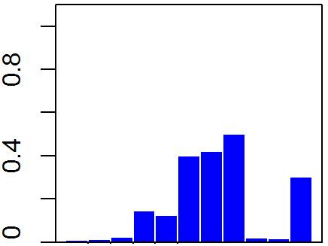
**Precision**

The system precision is evaluated prior and after defining named entities. Figures 9 and 10 shows how the precision is enhanced after defining entities on both the LEBL and DEBL approaches. The system is compared also with the baselines, in Figures 11 and 12 we show that both LEBL and DEBL are highly precise compared to the baselines. All the results shown in the figures take into account between 10 and 50 user profiles

|  |  |
| --- | --- |
| **Fig. 9.** LEBL precision with/without entities | **Fig. 10.** DEBL precision with/without entities |
| **Fig. 11.** LEBL precision compared to the baseline | **Fig. 12.** DEBL precision compared to the baseline |

**Similarity metrics**

Based on the predefined model, we mention that we used cosine similarity on the life events and biography attributes, and the Jaro-Winkler on screennames calculation.



Frequency (103)

0.2 0.4 0.6 0.8 1

0.2 0.4 0.6 0.8 1

0.2 0.4 0.6 0.8 1

Similarity score

Similarity score

Similarity score

1. Biography similarity (b) Life events similarity (c) screennames similarity distribution distribution distribution

Fig- 13. Similarity score distribution for each attribute

In figure 13 we show the distribution of these two similarity metrics on three different profile attributes. We can conclude from figure 13 (a) that similarity scores for biographies are congested around the interval [0.4,0.7], however, the same similarity metric show different conclusions in figure 13 (b), in this figure we observe that top matches are between 0.5 and 0.6. in figure 13 (c), the distribution is totally different from the previous ones. Screennames matching scores are distributed between [0.8,1] with null values for [0,0.4], this is due to the preprocessing phase that applied on them, moreover, screen names are composed of three words maximum.

**Baselines evaluation**

**DEBL Baselines**

Four random profiles are selected to show the gap between DEBL and its baseline. We prove that the baseline can fail to be linked using public attribute-based approaches, and show the high possibility of linking using profile descriptions. Table 2 shows a collection of four random profiles connected using biographies compared with @I seek ‘fb.me’ [11], providing the Twitter id and Facebook id of the user (we do not mention the screen names respecting the privacy of users).

**Table 6.** DEBL comparison with baselines

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Facebook id** | **Twitter id** | **@I seek ‘fb.me’** | **DEBL** | **#of Twitter profiles** |
| stevenjong | @StevenJong | No | Yes | 5 |
| fernanda.vasconcelo.9822 | @qbooomm | No | Yes | 14 |
| studioandrew | @andrewandraos | Yes (same image) | Yes | 5 |
| rodwell.mupungu | @minyango | Yes (same location) | Yes | 1 |

As shown in Table 6, DEBL can performs better than baselines. We see that baselines succeed in linking users via image attribute. However, image comparison can be more challenging than text. Others link users using location. However, Location can be same for different users. Hence, biographies can play a vital role when linking users.

**LEBL Baselines**

To compare LEBL to the behavioral baseline, we chose four random profiles, and show that information existing between users’ timelines can share different content of posts, timestamps between posts and other differences. Despite those, if we query the timeline with a specific type of information (such as Life events) we can reveal that the two profiles are identical. This can be enhanced in terms of accuracy (two same life events cannot be shared for two different users) and the lack of activity in one of the two profiles. In Table 7, we present in detail the analysis of the four different profiles.

**Table 7.** LEBL comparison with baselines

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Facebook id** | **Twitter id** | **HYDRA** | **LEBL** | **#of TW profiles** |
| itspeterwish | @peter\_wish | TW(textweets),FB(photos & videos) | Yes | 2 |
| hassan.mansour.528 | @mansourhassanhm | TW(latin letters),FB(Arabic latters) | Yes | 1 |
| 1406189719414584 | @IhabMortada | TW(Last public post 2015)  FB(last public post 1 hour ago) | Yes | 4 |
| 10154022294674058 | @venessabassil | Similar hashtags | Yes | 6 |

**9   Discussion and Conclusions**

This comprehensive review spots the light on a systematic review and comparison of profile matching algorithms in online social networks. Recent literature review in this area still missing to inclusive analysis. Sha et al. [8] state in their related works section that Zaafarani et al. [1] were the first who work on user linking. However, during our review, we detect earlier works [30]. Moreover, some approaches have never been referenced by the research community [35, 37].

Based on the type of matching feature used, we composed the approaches in this survey into two categories: attribute and context and activity features. The leading problem for all listed articles was to exploit effective matching interlinks in order to de-anonymize and link profiles belong to the same user entity across different SNS resources. The need of universal approaches is very important. Researchers judge that this is an important research field due to a set of reasons. Shen el al. [14] argue that building such model to infer the occurrence of matching between two profiles is important, to raise user awareness in protecting their profile public information. The other benefit for matching user profiles is to aggregate user information from various resources in order to create complement profiles that might serve business firms and research communities. Zafarani et al. [3] propose a behavioral matching approach to aggregate user information from different social networks. Noureddine et al. [23] introduce a research framework that serves the research community through correlating information from both digital libraries and social networks in order to create a complete and quality-aware researcher profiles. As discussed, the decision key that passes such algorithm is the concern of choosing matching features. Researchers early start by using profile basic information such as gender, age, screennames, location, interests, and others [9, 13, 2]. However, recently social network users raise their privacy attention of sharing personal information to publicity. This leads the research community to pry to more crucial and profile-independent features. Some researchers tend to explore the content of user profiles, this includes deep behavioral analysis [3, 12] and topic exploration of user messages [9, 10]. Alternative solutions can also be effective to solve the matching problem, by following a hierarchal methodology, which benefits from all user profile information as Jain et al. [11] did.

Research directions in this field are various. In case of matching attributes, those have not drained at all, the usage of new attributes are still possible. For instance, Facebook and Twitter, recently in 2016 provides two new public-only profile’s attribute. The *BIO*, which is a 160-character paragraph describes the user himself. We apply a small test on Facebook by collecting 3,271 distinct user accounts and found that 16.9% of them have written BIO about themselves. The second one is *join date,*the exact date when the user has registered himself to this social network (only by Facebook and Twitter). In case of matching algorithm, authors have followed only a naive process, none of them has used for example blocking approach, which considered more efficient [35]. In case of similarity metrics, many similarity functions used. However, exploiting more advanced ones also possible. N-gram models can be used for prediction purposes for words, characters or sentences; this can be used in context-based approaches.

References

1. R. Zafarani and H. Liu. Connecting corresponding identities across communities. ICWSM. 2009.

2. E. Raad, R. Chbeir, and A. Dipanda. User profile matching in social networks. In Network-Based Information Systems (NBiS). 2010.

3. R. Zafarani and H. Liu. Connecting users across social media sites: A behavioral-modeling approach. SIGKDD, ACM 2013.

4. M. Motoyama and G. Varghese. I seek you: searching and matching individuals in social networks. In WIDM, 2009.

5. S. Bartunov, A. Korshunov, S-Taek Park, W. Ryu, and H. Lee. Joint link-attribute user identity resolution in online social networks. SNAKDD Workshop, 2012.

6. J. Vosecky, D. Hong, and V.Y. Shen. User identification across multiple social networks. NDT ’09. 2009.

7. O. Goga, H. Lei, S. Hari Krishnan, G. Friedland, R. Sommer and R. Teixeir. Exploiting Innocuous Activity for Correlating Users Across Sites. WWW Conference 2013.

8. Y. Sha, Q. Liang and K. Zheng. Matching User Accounts across Social Networks based on User Message. ICCS 2016.

9. O. Goga, P. Loiseau, R. Sommer, R. Teixeira and K. P. Gummadi. On the Reliability of Profile Matching Across Large Online Social Networks. KDD 2015.

10. T. Van Le, T. Nghia Truong and T. Vu Pham. A Content-Based Approach for User Profile Modeling and Matching on Social Networks. MIWAI 2014.

11. P. Jain, P. Kumaraguru, Anupam Joshi. @i seek 'fb.me': identifying users across multiple online social networks. WWW (Companion Volume) 2013.

12. S. Liu, S. Wang, F. Zhu, J. Zhang, R. Krishnan. HYDRA: large-scale social identity linkage via heterogeneous behavior modeling. SIGMOD 2014.

13. A. Nunes, P. Calado, B. Martins. Resolving user identities over social networks through supervised learning and rich similarity features. SAC 2012.

14. Y. Shen, H. Jin. Controllable Information Sharing for User Accounts Linkage across Multiple Online Social Networks. CIKM 2014.

15. O. Peled, M. Fire, L. Rokach, Y. Elovici. Matching entities across online social networks. Neurocomputing 210. 2016

16. N. Hoang Long, J. J. Jung. Privacy-Aware Framework for Matching Online Social Identities in Multiple Social Networking Services. Cybernetics and Systems 46(1-2) 2015.

17. A. Panchenko, D. Babaev, S. Obiedkov. Large-Scale Parallel Matching of Social Network Profiles. AIST 2015.

18. N. Bennacer, C. Nana Jipmo, A. Penta, G. Quercini. Matching User Profiles Across Social Networks. CAiSE 2014.

19. W. Liang, B. Meng, X. He, Xianchao. GCM: A Greedy-Based Cross-Matching Algorithm for Identifying Users Across Multiple Online Social Networks. PAISI 2015.

20. R. Roedler, D. Kergl, G. Dreo Rodosek. Profile Matching Across Online Social Networks Based on Geo-Tags. NaBIC 2015.

22. L. Gu, R. Baxter, D. Vickers, and C. Rainsford. Record Linkage: Current Practice and Future Directions. CSIRO 2003

23. H. Noureddine, I. Jarkass, H. Hazimeh, A. Abou Khaled, E. Mugellini. CARP: Correlation Based Approach for Researcher Profiling. SEKE 2015.

24. W. W. Cohen, P. Ravikumar, S. E. Fienberg. A Comparison of String Distance Metrics for Name-Matching Tasks. IIWeb 2003.

25. P. Jain, P. Kumaraguru. Other Times, Other Values: Leveraging Attribute History to Link User Profiles across Online Social Networks. ACM (HT) 2015.

26. D. Perito, C. Castelluccia, M. Ali Kâafar, P. Manils. How Unique and Traceable Are Usernames? Privacy Enhancing Technologies . PETS 2011.

27. Elmagarmid, A. K., Ipeirotis, P. G., and Verykios, V. S. Duplicate record detection: A survey. IEEE Transactions on Knowledge and Data Engineering 19, 2007.

28. M. Szomszor, I. Cantador, H. Alani Correlating user profiles from multiple folksonomies. Hypertext 2008.

29. T. Iofciu, P. Fankhauser, F. Abel, K. Bischoff. Identifying Users Across Social Tagging Systems. AAAI 2011.

30. R. Nayak, M. Zhang, L. Chen. A Social Matching System for an Online Dating Network: A Preliminary Study. ICDM Workshops 2010

31. A. Krzywicki, W. Wobcke, Y. Sok Kim, Xiongcai Cai, M. Bain, A. Mahidadia, P. Compton. Collaborative Filtering for people-to-people recommendation in online dating. Int. J. Hum.-Comput. Stud. (2015)

32. X. Cai, M. Bain, A. Krzywicki, W. Wobcke, Y. Sok Kim, P. Compton, A. Mahidadia. Learning Collaborative Filtering and Its Application to People to People Recommendation in Social Networks. ICDM 2010.

33. J. S. Breese, D. Heckerman, C. Myers Kadie. Empirical Analysis of Predictive Algorithms for Collaborative Filtering. UAI 1998.

34. A. Malhotra, L. C. Totti, W. Meira Jr., Ponnurangam Kumaraguru, Virgílio A. F. Almeida. Studying User Footprints in Different Online Social Networks. ASONAM 2012.

35. S. Vosoughi, H. Zhou, D. Roy. Digital Stylometry: Linking Profiles Across Social Networks. Social Informatics - 8th International Conference. SocInfo 2015.

37. H. Zhang, M-Yen Kan, Y. Liu, S. Ma. Online Social Network Profile Linkage. AIRS 2014

38. H. Kopcke, E. Rahm. Frameworks for entity matching: A comparison. Data and Knowledge Engineering, 2009.

39. H. Hazimeh, E. Mugellini, O. Abou Khaled, P. Cudré-Mauroux. SocialMatching++: A Novel Approach for Inter-linking User Profiles on Social Networks. PROFILES’17@ISWC

40. W.E. Winkler, Overview of record linkage and current research directions, in: Bureau of the Census, Citeseer, 2006.

1. http://www.dailymail.co.uk/sciencetech/article-4992970/Facebook-tests-temporary-profile-status-feature.html [↑](#footnote-ref-1)
2. <https://www.facebook.com/directory> [↑](#footnote-ref-2)
3. <https://github.com/HusseinHESSO/ProfileLinking_v1.0> [↑](#footnote-ref-3)