IJMTARC – VOLUME – VI – ISSUE – 24, OCT-DEC, 2018 ISSN: 2320-1363

A Review OnSocial Identity Relation via Heterogeneous Behavior Modeling

[1] Madala Ashok
PG Scholar, Dept of CSE,
Prakasam Engineering College, Prakasam(Dt), AP, India.
[2]k.v srinivasarao
Asst Professor , Dept of CSE,
Prakasam Engineering College, Prakasam(Dt), AP, India

Abstract:

The study the matter of large-scale social identity linkage across wholly completely different social media platforms, that's of important importance to business intelligence by gaining from social data a deeper understanding and lots of correct identification of users. This paper proposes HYDRA, a solution framework that consists of three key steps: (I) modeling heterogeneous behavior by semi permanent behavior distribution analysis and multi-resolution temporal knowledge matching; (II) constructing structural consistency graph to measure the high-order structure consistency on users' core social structures across wholly completely different platforms; and (III) learning the mapping perform by multi-objective improvement composed of every the supervised learning on pair-wise ID linkage knowledge and additionally the cross platform structure consistency maximization. exhaustive experiments on 10 million users across seven widespread social network platforms demonstrate that HYDRA properly identifies real user linkage across wholly completely different platforms, and outperforms existing state of-the-art algorithms by a minimum of 2 hundredth below wholly completely different settings, and 4 times higher in most settings.

Index Terms—Social identity linkage, structured Learning, heterogeneous behavior, multi-resolution temporal information matching

1.INTRODUCTION

The recent blossom of social network services of all kinds has revolutionized our social life by providing everyone with the ease and fun of sharing various information like never before (e.g., micro blogs, images,

videos, reviews, location check-ins). Meanwhile, probably the biggest and most intriguing question concerning all businesses is how to leverage this big social data for better business intelligence. In particular, people wonder how to gain a deeper and better understanding of each individual user from the vast amount of social data out there. Unfortunately, information of a user from the current social scene is fragmented, inconsistent and disruptive. The key to unleashing the true power of social media analysis is to link up all the data of the same user across different social platforms, offering the following benefits for user profiling.

Completeness. Constrained by the features and design of each, any single social network service offers only a partial view of a user from a particular perspective. Cross-platform user linkage would enrich an otherwise-fragmented user profile to enable an all-around understanding of a user's interests and behavior patterns.

Consistency. For various reasons, information provided by users on a social platform could be false, conflicting, missing and deceptive. Cross-checking among multiple platforms helps improve the consistency of user information.

Continuity. While social platforms come and go, the underlying real-world users remain, who simply migrate to newer ones. User identity linkage makes it possible to integrate useful user information from those platforms that have over time become less popular or even abandoned. In this paper, we study the problem of automatically linking user accounts belonging to the same natural person across different social media platforms. It is beneficial to first explore the research challenges for a better understanding of this problem.

1.1 Our Contribution To address these challenges

Here propose HYDRA, a framework for cross-platform user identity linkage via heterogeneous behavior modeling. Compared with the record linkage problem long studied in database community [8, 11, 24], our technical breakthrough comes from taking advantage of two important features unique to social data: (I) user behavior trajectory along temporal dimension: both empirical and social behavior studies (e.g., [23]) demonstrate that,

over a sufficiently long period of time, a user's social behavior exhibits a surprisingly high level of consistency across different platforms. (II) user's core social network structure: the part formed by those closet to the user, and is called "core structure" for short. A user's core structures across different platforms share great similarity and offer a highly discriminative characterization of the user. Based on (I), we model the behavior similarity among online users with multi-dimensional similarity vectors with the following information: a) the relative importance of the user attributes, which measures how likely two user accounts belong to the same person when any one of their attributes is identical; b) the statistical divergence of topic distributions, describing the potential inclination of users over a long period of time; c) the overall degree of matching between users' behavior trajectories, capturing the highly correlated actions between user accounts over a certain period of time. Based on (II), we develop a linkage function learning methodology by taking full advantage of the agreement of the social structure level behavior consistency. The key intuition is to propagate the linkage information based on the linked users and the strong interaction along their social structures. Consequently, the linkage function can be effectively learned even with partial ground truth linkage information. In summary, our key contributions are: 1. Heterogeneous Behavior Modeling. We design a new heterogeneous behavior model to measure the user behavior similarity from all aspects of a user's social data. The proposed framework is able to robustly deal with missing information and misaligned behavior by long-term behavior distribution construction and a multiresolution temporal behavior matching paradigm. 2. Structure Consistency Modeling. We propose a novel social structure modeling method to leverage users' core social network structure to identify user linkage. We measure the high order pairwise behavior similarity and structure consistency by a graph representation. The model is learned to maximize the structure consistency,

which is equivalent to a convex objective function minimization. By incorporating structure consistency, our model is capable of identifying user linkage even when ground-truth labeled linkage information is insufficient. 3. Multi-objective Model Learning. We put forward a multiobjective optimization (MOO) framework [19] to solve the overall social identity linkage problem, which jointly optimizes the supervised learning on labeled user linkage pairs and the cross-platform structure consistency maximization. To deal with missing information, we further enhance the model into an iterative learning one. We also provide theoretical proof that our model is a generalization of the traditional semi-supervised learning, and can be efficiently solved by convex optimization. 4. Experiments on Large-scale Real Data Sets. We evaluate HYDRA against existing state-of-the-art approaches on two big real data sets — I) five popular Chinese social network platforms and (II) two popular English social network platforms — a total of 10 million users on seven social media platforms amounting to more than 10 TB data. Experimental results demonstrate that HYDRA significantly outperforms existing algorithms in identifying true user linkage across different platforms. Road Map. Section 2 discusses related work. Section 3 formally defines the social identity linkage problem. Section 4 presents an overview of our approach. Section 5 presents our heterogeneous behavior model and Section 6 proposes the multi-objective model learning. Section 7 presents the experimental results on different data sets. Finally, Section 8 concludes the paper.

2. RELATED WORK

2.1 User Linkage across Social Media

User linkage was firstly formalized as connecting corresponding identities across communities in [31] and a web-search-based approach was proposed to address it. Previous research can be categorized into three types: user-profile-based, user-generated-contentbased and user-behavior-model-based. User-profile-based methods collect tagging information provided by

users [13] or user profiles from several social networks and then represent user profiles in vectors, of which each dimension corresponds to a profile field (e.g., username,profile picture,description,location,occupation, etc.) [18, 22, 27]. Methods in this category suffer from huge effort of user tagging, different identifiable personal information types from site to site, and privacy of user profile. User-generated-content-based methods [16], on the other hand, collect personal identifiable information from public pages of user-generated content. Yet these methods still make the assumption of consistent usernames across social platforms, which is not the case in large-scale social network platforms. User-behavior-model-based methods [32] analyze behavior patterns and build feature models from usernames, language and writing styles. Unfortunately, previous methods [1] have not handled the missing information prevalent among usernames, usergenerated content, user behavior and social structures[2]

2) have not explored the underlying reasons for the missing information and its impact on user identity linkage[3]; 3) have not well formalized the user linkage problem with a solution of a sound theoretical foundation. To the best of our knowledge, our work is the first to link users across different social media platforms by integrating all the social data associated with a user in a unified model.

2.2 Authorship Identification across Documents

Authorship identification is a task that identifies the authors by analyzing their writing and language styles from their corresponding documents. Previous studies on authorship identification can be categorized into two kinds: content-based and behavior model based. Content-based-methods identify content features across a large number of documents [7,5,6]. Behavior-model-based methods capture writing-style features [33], or build language models [21] to identify content authorship. However, different from the document setting, social media platforms are characterized by data of

much greater heterogeneity, complicated network structures and a high degree of missing information, which could easily compromise most authorship identification methods. 2.3 Entity Resolution across Records User linkage is also in one way or another related to problems from other research communities including co-reference resolution in natural language processing [4], inter-media data retrieval [26], entity matching [28], record linkage in database [8, 11, 24], and name disambiguation in information retrieval [20, 14], which can be generalized as entity resolution across different records. In contrast to previous studies, we consider the user linkage problem in a much more challenging setting where we examine multiple features along time-line with missing and misaligned information across multiple media platforms. Also related are previous studies on user identification on a single site and deanonymization in social networks, which have been well surveyed in [16, 32].

3. FRAMEWORK OVERVIEW

In this paper, we propose HYDRA, which integrates both usersheterogeneous behavior and their core social network structure intoa unified multi-objective user linkage framework. HYDRA is composedof the following three main steps as illustrated .

Step 1.Behavior Similarity Modeling.calculate the similaritybetween two users of a pair for all user pairs via heterogeneousbehavior modeling. Details are discussed in Section 5.

Step 2.Structure Consistency Modeling.construct the structureconsistency graph on user pairs by considering both the corenetwork structure of the users and their behavior similarities. Detailsare discussed in Section 5.

Step 3.Multi-objective Optimization. Based on the previoustwo steps, we convert the SIL problem into a two-class classificationproblem and construct multi-objective optimization whichjointly optimizes the prediction accuracy on the labeled user pairsand multiple structure consistency measurements across different platforms. Details are discussed in Section 6.

4. HETEROGENEOUS BEHAVIOR MODEL

The key challenges in modeling user behavior across different social media platforms are (I) the heterogeneity of user social data and (II) the temporal misalignment of user behavior across plat-forms. The high heterogeneity of user social data can be appreci-ated by the following categorization of the data about a user avail-able on a typical social platform.

1. User Attributes. Included here are all the traditional struc-tured data about a user, e.g., demographic information, con-tact information, etc. (Subsection 5.1)

2. User Generated Content (UGC). Included here are the un-structured data generated by users such as text (reviews, micro-blogs, etc.), images, videos and so on. Modeling is primarily targeted at topic (Subsection 5.2) and style (Subsection 5.3).

3. User Behavior Trajectory. User behavior trajectory refers to all the social behavior of a user exhibited on the platforms along the time-line, e.g., befriend, follow/unfollow, retweet, thumb-up/thumb-down, etc. (Subsection 5.4)

To address these two challenges, we propose a behavior model-ing framework which computes the similarity between two users by capturing the heterogeneity in their behavior as well as the char-acteristics of their temporal evolution.

4.1 User Attribute Modeling

Textual Attributes. Common textual attributes in a user pro-file include name, gender, age, nationality, profession, education, email account, etc. While user profile information is effective in distinguishing different users, the relative importance of these at-tributes could be different, since attributes such as gender and pop-ular names like “John" are not as discriminative in identifying user linkage as some others such as email address. Yet, the weights of the attributes used in the matching can be learned from large train-ing data by probabilistic modeling.

Specifically, given a set of N

labeled training user pairs from different platforms, the relative importance of the attributes can be estimated by data counting. For a specific attribute a_k, k = 1; :::; M_A, we estimate the relative importance score of a_k by the following equation:

where $P_D(k)$ represents the number of user pairs matched on a_k in the positive labeled set P_D, and $N_D(k)$ represents the number of pairs matched on a_k in the negative labeled set N_D. " denotes a small real number used to avoid over-fitting. If a_k is absent for user i or i^0, it is denoted as a missing feature.

Visual Attributes. Besides textual attributes, visual attributes such as face images used in the profile can also help link users. However, such information could be very noisy as the face images might not be real, or come with poor illumination and severe oc-clusion. We design a matching scheme as shown in Figure 4 to compare two user profile images. In particular, if faces are detected from both images, the pre-trained classifier is used to determine if the two faces correspond to the same person. We use the face de-tector, facial feature extraction and face classifier provided by [12].

4.2 User Topic Modeling

An important feature of social media platform is that in general, over a sufficiently long period of time, the UGC of a user collectively gives a faithful reflection of the user's topical interests. Faking one's interests all the time defeats the purpose of using a social network service. Therefore, we propose to model a user's topical interests by a long-term user topic model. We first construct a latent topic model using Latent DirichletAllocation on every textual message, the output of which is a probability distribution over the topic space. We then calculate the multi-scale temporal topic distribution within a given temporal range for a user using the multi-scale tem-poral division similar to [17]. First, the temporal axis is divided into a series of time buckets with predefined scales (e.g., 16 days or 8 days). Then all the distribution vectors within a time bucket are aggregated into one topic distribution. After that, the corresponding similarity between the topic distributions in

each time bucket can be constructed. Finally, the overall similarity between user i and i^0 is calculated by averaging over the similarities of all the time buckets.axis is divided into multiple time buckets with different scales (we use 1, 2, 4, 8, 16 and 32 days in this paper to guarantee the optimal performance), then all the topic distribution vectors within each bucket are aggregated into a single distribution, which represents the topic distribution pattern within this time bucket. In C_t denotes the number of time buckets when the scale is selected to be 16. Correspondingly, the number of time buckets will be $2C_t$ and $4C_t$ respectively for 8 days and 4 days. Based on this, the similarity of topic evolution of a specific scale between two users can be simply calculated by averaging over the similarities of all temporal intervals, where each similarity can be measured by the chi-square kernel or histogram intersection kernel [17]. Finally, all the similarities calculated using different time scales are concatenated into a similarity vector.

The proposed long-term user topic model captures the behavior similarity from pair-wise topic correlation at a series of coarse-to-fine resolutions. In this paper, we analyze the following distribution types using this proposed strategy:

Content Genre Distribution. The content genre measures the relevance between the textual messages and popular topics on so-cial media sites, e.g., sports/ music/ entertainment/ society/ history/ science/ art/ high-tech/ commercial/ politics/ geography/ traveling/ fashions/ digital game/ industry/ luxury/ violence.

Sentiment Pattern Distribution. According to studies in sentiment mining [10], we can model sentiment patterns using a two-dimensional space (arousal-valence) [10] or roughly group all emotions into several categories, e.g., happy/ fear /sad /neutral. It can be done by extracting representative emotional key words in the textual content and learning a sentiment vocabulary. After that, each textual message can be represented by a probabilistic distribution on the sentiment vocabulary. We use the scheme in this section to compute a multi-scale similarity measure on sentiment patterns between two users.

4.3 User Style Modeling

The language style of a user including personalized wording and emoticon adoption is usually well reflected in comments, tweets and re-tweets (e.g. function words extraction [16]), which is beneficial to distinguishing between different users. To model a user's characteristic style, we extract the most unique words of each user by a simple term frequency analysis on the whole database. Note that since these unique words may also be inaccurate, we select the k (k = 1; 3; 5) most unique ones after removing stop words from the least-used terms of the whole user data repository. Such choices of k have been adopted widely in related studies [29].For user pairs, we can measure S_{lea} (the similarity on the unique word pattern) by word matching (the words should be converted into a uniform format, such as lower-case and singular form):

$$S_{lea} = \frac{\#matched\_words}{k} \quad (4)$$

4.4 Multi-resolution Behavior Modeling

User behavior trajectory is a unique feature of social media data laying out a user's social behavior along the timeline. In this paper, we are mainly concerned with the following patterns:Mobile Trajectory and Location Information. Social media sites with location-based-service provide strong support and incentive for users to record and share their locations. Generally, users with similar trajectory patterns and no conflicting instances over an extended period of time are likely to be the same person in real life.Multimedia Content Generation and Sharing. Users may post similar multimedia content on the web. For example, they may upload or share exactly the same image/ video/music. However, If a high level of synchrony is observed over an extended period of time between two user accounts from different platforms, it is reasonable to hypothesize that these two users correspond to the same person.

A natural solution is to construct a set of pattern-matching sensors, one for each modality (location, visual, textual and audio), and use them to collectively evaluate user similarity. However, as people are not always using multiple social platforms simultaneously, a significant amount of information could be missing in such a task. We therefore propose a multi-resolution

temporal behavior model to perform pattern matching with the ubiquitous presence of missing information.

As shown in Figure 1, given two users i and i^0, we first construct a set of pattern-matching sensors with different temporal searching ranges. If matched patterns (denoted by pentagons) are identified within the selected range of a pattern-matching sensor, a positive stimuli signal would be generated.

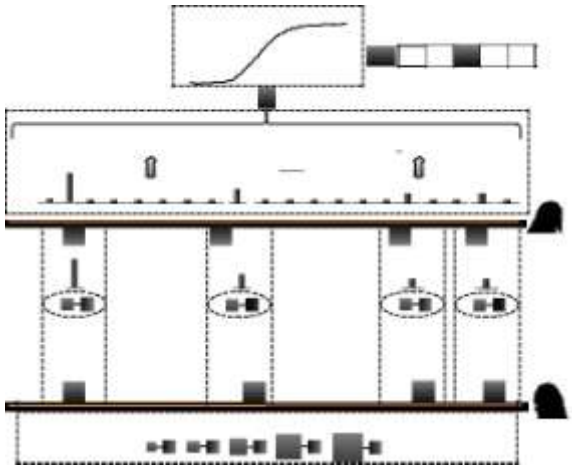


Figure 1: Multi-resolution temporal behavior modeling. A set of pattern-matching sensors are designed. For two users, sen-sors are used to detect the corresponding type of matched be-havior within certain temporal scales. When all the matched behaviors have been detected by sensors, an l_q norm pooling and nonlinear sigmoid mapping then aggregates all matched behavior signals into a multi-resolution similarity vector.

The choice of l_q-norm is inspired by bio-stimulation. It has been found that the maximum stimulation from a pooled signal set plays a significant role for perception. When q approaches infinity, the signal selection tends to better approximate the max-imum stimulation (i.e., max-pooling). Since the pattern-matching would be performed under different temporal scales, we can extract a multi-resolution temporal matching pattern between two users on the sparsely and asynchronously occurring patterns. The sigmoidfunction S^b_{mr} = is a typical nonlinear transformationfunction, where the parameter can be tuned on the specific validation dataset. The pattern-matching sensors we construct in this paper are the following:

Location Matching Sensor. A location matching sensor calculates location adjacency by a Gaussian kernel on geo-coordinates of user i and user i^0 within the predefined spatial range [17].Near Duplicate Multimedia Sensor. A near duplicated image sensor or down-sampling method [9] is constructed for near dupli-cate multimedia sensor.

6. MULTI-OBJECTIVEMODEL EARNING

IJMTARC – VOLUME – VI – ISSUE – 24, OCT-DEC, 2018 ISSN: 2320-1363

Based on the heterogeneous behavior modeling from user at-tributes, UGC and behavior trajectories as explained in Section 5, we propose to learn the linkage function via a multi-objective optimization framework.Supervised Learning. Some social media platforms allow users to log in to different platforms with one account. For example, we can use a Facebook account to log in to Twitter. We collect such user-provided linkage information as the ground-truth label infor-mation. We notice that the labeled training pairs collected by our paradigm is much cleaner (precision over 95%) than the approach in [16] (precision around 75%) where the labeled training pairs are automatically generated based on the uniqueness (n-gram proba-bility) of user names. We also collect label information by user attribute matching as the pre-linked label information. By utilizing the collected label information, we minimize the structured loss (SVM objective function) on the labeled training data.

Figure 2: Structure consistency maximization. Given two plat-forms, we measure both behavior similarity and structure con-sistency among their most frequently communicating friends (elliptical rings), especially those with ground truth linkage in-formation (linked by black arrows). The arrows within each platform indicate how the linkage information can be propa-gated along the social structure of each user. Consequently, the true user linkage (red dashed arrows) are correctly identified while the falsely linked user pairs (green dashed arrows) are filtered out.

IJMTARC – VOLUME – VI – ISSUE – 24, OCT-DEC, 2018 ISSN: 2320-1363

7. EXPERIMENTAL EVALUATION

7.1 Experiment Setup

Real Data. We use two publicly available large-scale real data sets for our experiments. The first one, referred to as "Chinese", in-cludes five popular social networks services which were originated from China and have since gained global popularity.

1. SinaWeibo: (www.weibo.com) A hybrid of Twitter and Face-book with a user base of 500 million users and 47 million daily active users by December 2012.
2. TecentWeibo: (t.qq.com) Another twitter-like micro-blogging service with 500 million users and over 100 million daily ac-tive users.
3. Renren: (www.renren.com) A social network service dubbed as the Facebook of China with 162 million registered users.
4. Douban: (www.douban.com) A social network service for people to share content on topics of movies, books, music, and other off-line events in Chinese cities, with over 100 mil-lion monthly unique visitors.
5. Kaixin: (www.kaixin001.com) A social network service with 160 million registered users.

We use 5 million Chinese users in this data set, each with ac-counts on every one of the five platforms. The time span of this data set is from June 2012 to June 2013.

The second one, referred to as "English", includes two globally popular social networks: (1) Twitter (twitter.com); and (2) Face-book (www.facebook.com). We use 5 million Chinese users in this data set each with accounts on both Twitter and Facebook. The time span of this data set is from June 2012 to June 2013.

For the social networks above, we collect user profiles (e.g. gen-der, city, and favorites), social content (e.g. tweets, posts, and sta-tus), social connections (e.g., friendship, comments, and repost or retweet contents), and timeline information (e.g., time index for each behavior).

Our ground truth of the linkage of each user across all the plat-forms are provided by a third-party data provider who has access to each Chinese user's national ID number, IP address and home ad-dress used by the user to register all accounts on different websites, all of which collectively serve as the most reliable data to uniquely identify a natural person and link all the different accounts. Note that users in the English data set are all Chinese users of our choice.

In the following experiment result, x-axis is the decreasing ranked result (user is by degree, and community is by size). The ratio be-tween the labeled data to unlabeled data is set to 1=5, but we have also tested other ratio settings in our experiment.

Experiment Environment. Our experiments and latency observations are conducted on 5 standard servers (Linux), with Intel (R) Xeon (R) Processor E7-4870 (30M Cache, 2.40 GHz, 6.40 GT/s Intel (R) QPI, 10 cores), 64 GB main memory and 10,000RPM server-level hard disks.

Compared Methods. We compare both our

methods with the fol-lowing state-of-the-art approaches and our own baselines.

(I) MOBIUS: a behavior-modeling approach to link users across social media platforms [32].

(II) Alias-Disamb: an unsupervised data-driven approach based on username analysis to link users across platforms [16].

(III) SMaSh: a record linkage approach finding linkage points over Web data [11].

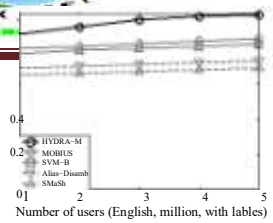
(IV) SVM-B: binary prediction on user pairs using support vec-tor machines on the proposed similarity calculation schemes.

(V) HYDRA-Z: a degenerate version of our model HYDRA where all the missing features are filled with zeros.

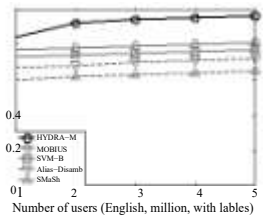
(VI) HYDRA-M: our model HYDRA with missing features filled with the core social network friend structure described in Section . Without specification, we call HYDRA-M as HYDRA[7].Parameter Settings. To achieve better performance of all the approaches, a validation set with 5 million user pairs and their ground truth labels have been used.

For the pair-wise similarity calculation in this paper, the parameters (e.g., " for user profiling, q and for multi-resolution temporal similarity modeling) are tuned by a grid search procedure to maxi-mize the performance of a linear SVM on the validation set. Then the optimized multi-dimensional similarity $x_{ii}0$ are used for model construction of (IV), (V) and (VI).

For both HYDRA-Z and HYDRA-M, we need to tune the model



(c) Precision in English



(d) Recall in English

Figure 4: Performance w.r.t. #labeled pairs.

parameters$_L$, $_M$, p, $_S$ and $_D$. We construct the models on the training data and conduct parameter tuning on the validation set. In the following sections, we will illustrate the functional properties with respect to different model parameter settings.

Evaluation Metrics. In our experiments, we use precision and recall to evaluate the effectiveness, and the total execution time (at different scales) to evaluate the efficiency. Precision is defined as the fraction of the user pairs in the returned result that are correctly linked. Recall is defined as the fraction of the actual linked user pairs that are contained in the returned result.

The parameters of all the kernels for HYDRA are tuned strictly according to the methods described in the previous sections.

8. CONCLUSION

In this paper, we link up user accounts of the same natural per-son across

different social network platforms. We propose a frame-work, HYDRA, a multi-objective learning framework incorporating heterogeneous behavior modeling and core social network structure. We evaluate HYDRA against the state-of-the-art solutions on two real data sets — five popular Chinese social networks and two popular English social networks, a total of 10 million users and more than 10 tera-bytes of data. Experimental results demonstrate that HYDRA outperforms existing algorithms in identifying true user linkage across different platforms.

10. REFERENCES

[1]T. W. Athan and P. Y. Papalambros. A note on weighted criteria methods for compromise solutions in multi-objective optimization. Engineering Optimization, 27:155–176, 1996.

[2]M. Belkin, P. Niyogi, and V. Sindhwani. Manifold regularization: A geometric framework for learning from labeled and unlabeled examples. Journal of Machine Learning Research, 7:2399–2434, 2006.

[3]S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein. Distributed optimization and statistical learning via the alternating direction method of multipliers. Foundations and Trends R in Machine Learning, 3(1):1–122, 2011.

[4]J. Cai and M. Strube. End-to-end coreference resolution via hypergraph partitioning. In COLING'10.

[5]R. Cilibrasi and P. M. B. Vitanyi. Clustering by compression. IEEE Transactions on Information Theory, pages 1523–1545, 2005.

[6]W. Cui, Y. Xiao, H. Wang, Y. Lu, and W. Wang. Online search of overlapping communities. In SIGMOD'13.

[7]O. de Vel, A. Anderson, M. Corney, and G. Mohay. Mining e-mail content for author identification forensics. SIGMOD Record, 30(4):55–64, 2001.

[8]A. K. Elmagarmid, P. G. Ipeirotis, and V. S. Verykios. Duplicate record detection: A survey. IEEE Transactions on Knowledge and Data Engineering, pages 1–16, 2007.

[9]R. C. Gonzalez and R. E. Woods. Digital image processing. 1992.

[10] A. Hanjalic and L.-Q. Xu. Affective video content representation and modeling. IEEE Transactions on Multimedia, pages 143–154, 2005.

[12] O. Hassanzadeh, K. Q. Pu, S. H. Yeganeh, R. J. Miller, M. Hernandez, L. Popa, and H. Ho. Discovering linkage points over web data. PVLDB, 6(6):444–456, 2013. http://www.briancbecker.com/bcbcms/site/proj/facerec/fbextract.html.

[13] T. Iofciu, P. Fankhauser, F. Abel, and K. Bischoff. Identifying users across social tagging systems. In ICWSM'11.

[14] D. V. Kalashnikov, Z. Chen, S.

Mehrotra, and
R. Nuray-Turan. Web people search via connection analysis.
IEEE Transactions on Knowledge and Data Engineering, pages 1550–1565, 2008.

[15] S. Kumar, R. Zafarani, and H. Liu. Understanding user migration patterns in social media. In AAAI'11.

[16] J. Liu, F. Zhang, X. Song, Y.-I. Song, C.-Y. Lin, and H.-W. Hon. What's in a name?: an unsupervised approach to link users across communities. In WSDM'13.

[17] S. Liu, S. Wang, K. Jeyarajah, A. Misra, and R. Krishnan. TODMIS: Mining communities from trajectories. In ACM CIKM'13.

[18] A. Malhotra, L. C. Totti, W. M. Jr., P. Kumaraguru, and
V. Almeida. Studying user footprints in different online social networks. In ASONAM'12.

[19] R. T. Marler and J. S. Arora. Survey of multi-objective optimization methods for engineering. Structural and multidisciplinary optimization, 26(6):369–395, 2004.

[20] Y. nanQian, Y. Hu, J. Cui, Q. Zheng, and Z. Nie. Combining machine learning and human judgment in author disambiguation. In CIKM'11.

[21] J. Novak, P. Raghavan, and A. Tomkins. Anti-aliasing on the web. In WWW'04.

[22] A. Nunes, P. Calado, and B. Martins. Resolving user identities over social networks through supervised learning and rich similarity features. In SAC'12.

[23] G. Pickard, W. Pan, I. Rahwan, M. Cebrian, R. Crane,
A. Madan, and A. Pentland. Time-critical social mobilization. Science, 334(6055):509–512, 2011.

[24] M. Sadinle and S. E. Fienberg. A generalized fellegi-sunter framework for multiple record linkage with application to homicide record systems. Journal of th American Statistical Association, 105(502):385–397, 2013.

[25] B. Schlkopf and A. J. Smola. Learning with kernels: support vector machines, regularization, optimization, and beyond.
Cambridge: TheMITPress, 2002.

[26] J. Song, Y. Yang, Y. Yang, Z. Huang, and H. T. Shen. Inter-media hashing for large-scale retrieval from heterogeneous data sources. In SIGMOD '13.

[27] J. Vosecky, D. Hong, and V. Shen. User identification across multiple social networks. In NDT'09.

[28] J. Wang, G. Li, J. X. Yu, and J. Feng. Entity matching: How similar is similar. PVLDB, pages 622–633, 2011.

[29] J. Weston, C. Leslie, E. Ie, D. Zhou, A. Elisseeff, and
W. Noble. Semi-supervised protein classification using cluster kernels.
Bioinformatics, pages 55–64, 2005.

[30] P.-L. Yu, Y.-R. Lee, and A. Stam. Multiple-criteria decision making: concepts, techniques, and extensions. Plenum Press New York, 1985.

[31] R. Zafarani and H. Liu. Connecting

corresponding identities across communities. In ICWSM'09.

[32] R. Zafarani and H. Liu. Connecting users across social media sites: A behavioral-modeling approach. In KDD'13.

[33] R. Zheng, J. Li, H. Chen, and Z. Huang. A framework for authorship identification of online messages: Writing-style features and classification techniques. Journal of the Association for Information Science and Technology, 57(3), 2006.