Disambiguating Identity through Social Circles and Social Data

Matthew Rowe, Fabio Ciravegna

Web Intelligence Technologies Lab
Department of Computer Science
University of Sheffield, UK
{m.rowe, f.ciravegna}@dcs.shef.ac.uk

Abstract: This paper presents an approach to disambiguate extracted identity information relating to different individuals through the use of social circles. Social circles are generated through the extraction and pruning of social networks using the analysis of existing social data. Social data encompasses information such as images, videos and blogs shared within a social network. Identity information is extracted by involving the user in both selecting their key identity features for disambiguation, and validating the retrieved information. Our approach provides a methodology to monitor existing identity information, applicable to addressing such issues as identity theft, online fraud and lateral surveillance.

Keywords: communities, disambiguation, identity, semantic web, social networks, social web

1 Introduction

The social web has seen enormous growth over the past 2 years. For example, in the UK alone, there are now more than 9 million unique Facebook users, 5 million unique MySpace users and 4.1 million unique Bebo users [16]. Commonly, users of such services use these sites to create an on-line social environment very similar to the one they experience in their everyday life, but extended and complemented by the on-line features of these services. Common tasks include organising events, and interacting with friends through messaging, blogging and sharing photos.

In parallel with the growth of the Social Web, the number of cases of malicious use of personal information has also grown [18]. The problems of identity theft and online fraud are of great significance in many countries, where personal details of individuals are stolen daily and used for malicious purposes. Users of online social networking sites commonly share information intended for social interaction. However, such personal information (e.g. the date of birth provided to help people remind someone's birthday) can be misused with malicious intention (e.g. the date of birth is partly used to check a person's identity when accessing phone services in the UK). The number of reported cases of cyber stalking and online harassment has increased, and the practice of lateral surveillance has also risen, including reports of cases where potential employees are vetted based on their online presence.

Such issues present several challenges and opportunities for research. The development of technologies able to monitor personal information of a given individual provide a stepping stone to assessing the risk of an individual becoming a victim of identity fraud. This requires first and foremost the ability to identify and integrate personal identity information from heterogeneous web resources. This paper focuses on the first part of this task, i.e. the challenge of discovering personal identity information from semi-structured Web resources, and the required disambiguation of individuals contained within this information. Semantic technologies provide a useful means for carrying out disambiguation, by formalising identity and using semantic information to assert facts about an individual.

The approach presented in this paper utilises social circles derived from social data to disambiguate individuals. This is split into several stages:

Firstly a user's social network is extracted from social networking sites and integrated. The resulting network is then pruned into a social circle containing identifiable relationships with other individuals by analysing existing social information. A social circle is denoted as a group of people linked to a central individual by some identifiable common relation. Social information describes content related to a given person that contains social characteristics, and provides a useful source of socially annotated data due to the rise in sharing facilities in social Web sites. This can be of any type including blogs, images tagged with person names, and instant messaging conversations.

The user decides which identity features are best suited to minimally distinguish their identity from others. The process of discovering identity data begins by extracting information from the set of resources using these identity features: We use a social approach to allow users to rate and critique resources based on the accuracy and volume of information available. Each user has the ability to contribute their opinion about an identity resource based on the information present relating to them. Other identity features are used that have been repeatedly chosen by other users to extract data. Other resources are discovered and analysed automatically via the Web, e.g. FOAF-web.

Finally all resources found to contain the user's details are analysed to disambiguate between potential individuals sharing similar identity properties . The disambiguation procedure uses the user's generated social circle by analysing resources found to contain user details. Comparisons are made between the social circle and the resource to derive the probability of the information present belonging to the user. Upon completion of the disambiguation process, all known identity

information attributed to a person is displayed for validation. A user in the loop can then validate it, correct it and re-start the process of extraction.

In order to implement the approach we have created a Facebook application capable of extracting a users social network from the social networking site. Our application is able to access all the images and conversation data from Facebook relating to each member of a user's social network and prune the social network to produce a circle consisting of the user's closest friends. Images are chosen due to the social tagging application commonly found in social networking sites. Our implementation then uses this social circle for the disambiguation process. The Facebook application is available for downloading and testing[1].

This paper is structured as follows: Section 2 presents related work to disambiguating identity through social circles. Section 3 outlines the semantics of personal identity, and the selection of prevalent features. Section 4 details the presented approach, explaining the methods used and technical details. Section 5 explains the proposed methodologies of evaluation and explains the reasons for this. And finally, section 6 discusses the primary conclusions from the investigation so far, and outlines the proposed future work in this area.

2 Related Work

The related work to our approach covers several fields of research: Name disambiguation literature is included describing similar approaches to disambiguating persons using related contextual information. Social network analysis literature covers formal definitions of social circles and groupings. Social network mining literature discusses differing methods for extracting social network information from various sources. Object identification literature presents approaches that could be adapted to identifying individuals, and commercial systems are included detailing work towards identity disambiguation through the provision of identity theft risk assessments.

The problem of disambiguating individuals, also known as instance unification, is addressed in [1]. Citations are used to discover additional information about a given individual author; this information is then used to mine the web. Social networks are constructed surrounding the author based on the co-authorship of papers. Similarly work by [7] investigates the challenge of identifying misspelt and abbreviated names by using clustering together with Naïve Bayes to compute the probability of a given name belonging to a name cluster. Our methodology is similar to [1] by using existing information to derive the initial social network, however we utilise further social information such as image content and conversation data to prune the network.

Work to identify social circles is presented in social network analysis literature such as [14] where cliques are initially discovered from an individual's social network, categorised as a sub-graph where a relation connects all pairs of points within the graph. A social circle is the aggregation of overlapping cliques within the

[1] http://apps.facebook.com/socialcircular

[3] http://www.garlik.com

larger social network graph, and the key group of friends related to a given central individual. Social grouping enables socially linked individuals to be clustered based on a common relation where the relation can simply be a binary classifier used to prune an individuals social network to only contain those individuals positively classified. Our approach uses this definition to generate the social circle from the initial social network, we use classifiers over image and conversation data to derive social links.

A technique for mining social networks is presented in [9] and [6] utilising a three-step approach: Firstly, mining the web for social network information identifying links between two individuals, secondly monitoring real world interactions between individuals to confirm relationships between them, and thirdly, monitoring interactions between users on the web by capturing online communications between individuals. Further work will mine social network information from the wider web. At present our approach only generates social networks from social networking sites.

Work described in [10] uses a two-part methodology to gather social network information by mining information from the web and crawling for semantic documents containing information described using the FOAF [2] ontology. Social network information is mined from the web by querying a search engine, with pairs of names of individuals considered to be friends. The number of pages returned containing both names co-occurring is the count for that pair; this gives the strength of the relationship between the two individuals. Work by [4] presents an approach to social network extraction using FOAF files by crawling FOAF-web, extracting information from each FOAF file and aggregating with information from other FOAF files. Assertions are made about discovered individuals using the supplied semantic information. Our approach also uses FOAF-web to extract information, however we are only concerned with identity information during the mining phase. The later disambiguation phase checks the FOAF content for relationships corresponding to the social circle.

Work by [8] presents a methodology to identify labels for relations between two socially linked entities; the label word along with the two entities then form the query to be entered into a search engine to retrieve additional information. This methodology allows ontologies to be generated using the entities and relations that link them together. Threshold tuning is used to refine the importance between two entities using objective and subjective criteria. An approach described in [11] identifies relations between two socially linked entities and uses labels derived from the collective context of the entities to define the relation in a similar manner to work in [8].

Literature relating to object identification presents a similar approach to identity extraction by using features to recognise objects. An interesting methodology presented in [17] details a general method for learning rules to map data for object identification using domain independent attributes for identification. Work by [12] describes a heavily cited framework for object identification presenting a straightforward modulated approach using clustering for the pre-selection of similar object pairs. Literature such as [15] utilises the features of objects to predict the likelihood probability of a match occurring for object pairs. Work by [3] describes an automatic approach to identifying and disambiguating relevant information using a library of string metrics to deduce term record similarity. Our approach similarly uses

string metrics to compare the properties of identity at a low-level, mining the web for identity information, and disambiguating extracted identity information.

Several systems utilising social networks and social information offer users the ability to monitor their personal information. Garlik[3] offers services to enable the monitoring of personal information, but fails to correctly disambiguate between individuals in several cases. Garlik only uses the presence of social networking accounts when detecting personal information; our approach differs by using the information within the accounts to extract relationship information. Maltego[4] tracks social networks relating to a given individual through mining the web for information, and identifying real world links between people, and groups. Spock[5] is another similar application that crawls the social web and the wider web for occurrences of names that have been searched for, information is then aggregated together using the similarities in the derived content. Our work is similar to Maltego by mining the web for person names, however we place a greater emphasis on an individual's social circle rather than their wider social network. The presented approach also differs by allowing the user to select their most prevalent identity feature, and prioritising identity features with the greatest cumulative prevalence.

[4] http://www.maltego.com
[5] http://www.spock.com

3 The Semantics of Identity

The semantics of identity are extremely important when extracting identity information and disambiguating between individuals. The notion of personal identity described in [19] splits identity into 3 tiers: My identity containing persistent identity features; shared identity containing attributes assigned to an individual by others; and abstracted identity containing identity denoted by grouping. The first tier, 'My identity', contains identity features of the greatest significance when disambiguating one individual from the next; name, date of birth, etc. However, the prevalence of these identity features differs between individuals. Consider a scenario involving a man named John Smith, John must decide on what features of his identity are the most prevalent. He knows that his name is very common, but he knows that he is the only 'John Smith' on his street, so he selects his postcode. As certain identity features build a cumulative level of prevalence, these features become inherently used by the approach.

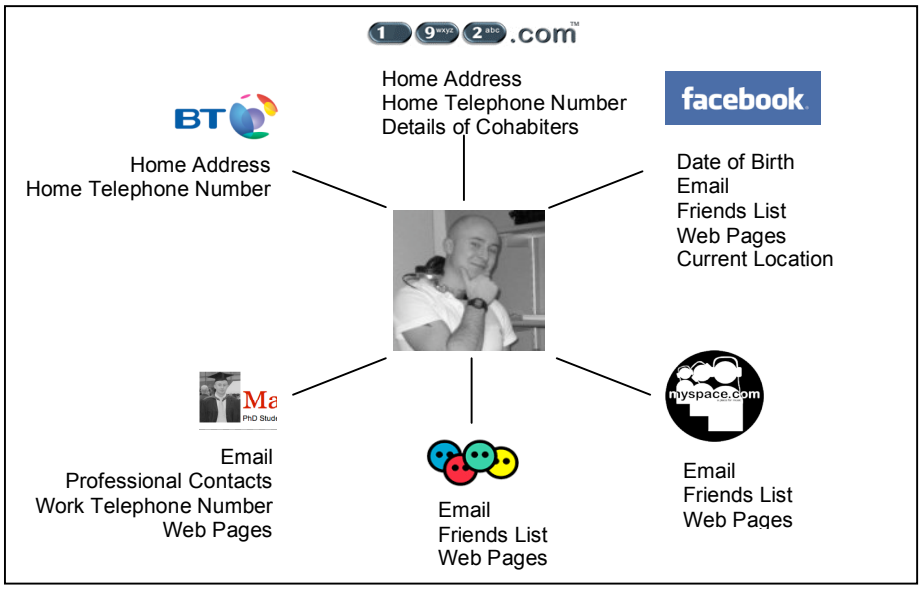

Figure 1. Distribution of features of my identity throughout the Web

As figure 1 shows the distribution of identity features contain enough information about a given individual to compile a fairly complete identity profile. The social network feature of identity is related to the 'Shared identity' tier presented in [19]; this tier contains temporary relationships prone to either becoming stronger or breaking down. The creation of an ontology to encapsulate the properties of an individuals identity and their social network based on the FOAF [2] specification is

required. FOAF provides sufficient properties to define certain identity information, however it fails to provide other features that would differ between domains e.g. Social security number, and national insurance number. Such a formalisation would aid with integration of identity information and reasoning.

4 An Approach to Disambiguating Identity using a Social Circle

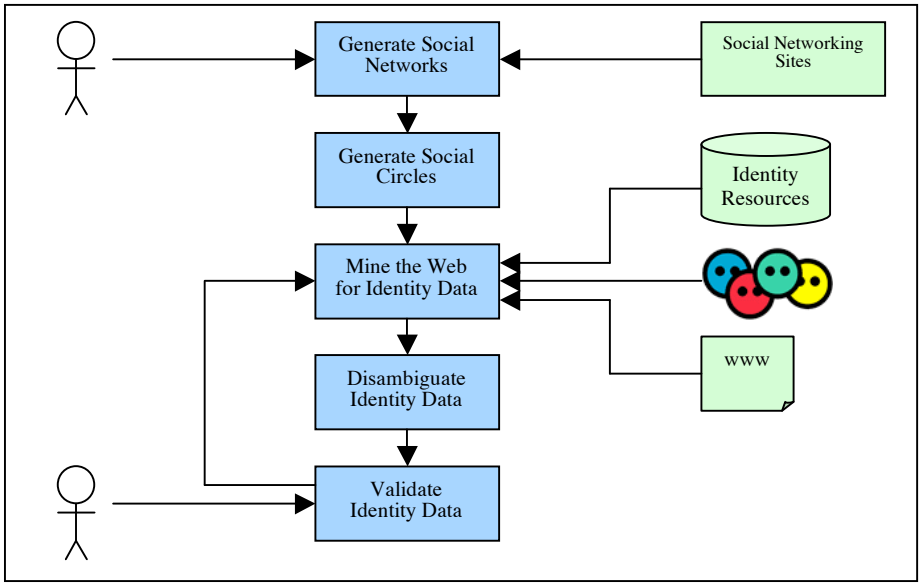

Figure 2. Approach Overview

This section explains the details of our approach to disambiguate individuals using social circles, and present disambiguated information for validation. This section is composed of sub-sections describing the processes performed in each stage of the approach as displayed in figure 2. To illustrate these processes we use the example of John Smith, a man who is concerned about his identity being stolen or misused, and wishes to know what information exists about him. We explain our approach in terms of a completed system.

4.1 Generating Social Networks

In order to generate social networks, existing accounts with social networking sites belonging to the user are needed to seed the disambiguation process. At present only Facebook data is being used to seed the approach. However, data from other accounts can be used. John logs into a specially created Facebook application. The application works by accessing Facebook's Developer API and extracting the social network

information into FOAF [2]. This compiles a comprehensive list of all of John's friends he is acquainted with. In the future, other social networking sites and alternative services can be used to seed the process providing John has an account with them. Existing approaches could also be used to generate the social network such as entity co-occurrence and mining FOAF-web for social networks [7,8].

4.2 Generating Social Circles from Multimedia Data

Using the algorithms described in figure 3 and figure 4, strengths are derived describing the relationships the user has with each member of their social circle. Several users of social networking sites are prone to adding friends who are more likely to be acquaintances and not necessarily people within their social circle or clique. Figure 3 details the derivation of relationship strengths from image data; images are extracted from a web resource annotated with unique identifiers of individuals. Should the unique identifier match that of the user, the social bond is strengthened. The same principle is used in the algorithm to derive the social bond from conversation data by analysing all received and submitted messages to identify the user.

```
-----------------------------------------------------------------------

   Extract all friends from friend list
   For each friend from friend list
       Extract all photos the friend appears in
             For each photo from photolist
               If you appear in the photo with your friend
               Increment friend strength by one
       Divide the friend strength by the total number of
   photos
        Store the social bond strength
-----------------------------------------------------------------------
```

Figure 3 – Algorithm deriving relationship strengths from images

```
-----------------------------------------------------------------------

   Extract all friends from friend list
   Extract all messages you have received
   For each friend from friend list
       Extract all the messages sent to their profile
       For each message they have received
           If you received the message
               Increment message count by 1
       Divide message count by the total number of
   messages
       For each message you have received
           If the message was sent by the friend
               Increment received message count by 1
```

```
      Divide received message count by the total number
  of received messages
      Add the message count to the received message count
      Divide combined message count by 2
      Store the combined message count
-------------------------------------------------------------------------
```

Figure 4 – Algorithm deriving relationship strengths from conversational information

Using these algorithms, John's social network is pruned to fewer components: For each of John's friends the Facebook application extracts all the photos they appear in. Each photo is then verified to see if John also appears in the photo. A score is kept to count how many times John appears in his friend's photos, this score is then used to derive the weighting of their relationship. The same process is carried out using conversational information. Both of the derived weightings are used to compute an average social bond, the <*social-bond*> property is added to the existing FOAF specification to contain the strength of the social bond for each friend.

4.3 Mining the Web for Identity Data

As discussed in section 3 the prevalence of identity features are purely based on the user. This approach incorporates both identity features specifically selected by the current user, and the social dynamic of cumulatively prevalent identity features following their frequent selection. The mining for identity data begins by using the features of identity that have been chosen by the user. John Smith knows that his name is very common, he must therefore select accompanying features of his identity. He knows that he is the only John Smith on his street so he chooses his postcode, and he also knows that his email address is unique so he selects that.

Matching identity information compares instances of identity to derive a similarity measure corresponding to the properties of each instance. At a meta-level this allows the comparison of objects regardless of the differences in the format of their properties. In our approach we only compare textual components, so one application of instance matching would use matching names and any other identity features through simple string matching using the SimMetrics package [3]. As resources are parsed, the information is analysed for any possible occurrences of the identity features being searched for. If any matches take place, then the URI of the resource is stored.

The mining process begins by firstly mining FOAF-web to extract semantic information about the user using the prevalent identity features selected by the user. Due to their semantically rich format, parsing FOAF files is a simple process allowing basic comparison of the identity features specified by the user and the found information. Using the <foaf:seeAlso> property, linked FOAF files are also mined to gather more Semantic information.

Secondly existing identity resources are accessed to extract identity information using the prevalent identity features selected by the user. Identity resources can be

shard by users of the implementation that contain identity information. The community of users are able to provide feedback regarding the accuracy and volume of the information present in each resource. Resources are marked by the community as useful sources for identity information, and become prioritised when the process of mining identity information begins. When mining both FOAF-web and identity resources for information, any found social content is flagged for later use. This includes any names observed from the individual's social circle.

Finally the wider web is mined for identity information by submitting structured queries to a search engine. Queries are structured to detail the most important and prevalent features of the user's identity selected at the start of the process, together with the cumulatively most prevalent identity features. In John's case three queries would be used: His name and his postcode. His name and his street name (derived from the postcode). His name and his email address.

4.4 Disambiguating Identity Data

Following the collection of all possible identity instances attributed to the user it is important to disambiguate between information relating to different individuals. Social circles are used to perform the disambiguation process. Each resource is parsed to derive information about any of the user's friends from their social circle. In our approach entity extraction is used to find any names within the identity resources and compile the names into a list using the rule based document annotator; Saxon [5]. Each name in the list is then compared with the names from the user's social circle by computing the Smith-Waterman-Gotoh distance using the SimMetrics [3] package, and comparing this distance to a predefined threshold. If a match occurs then the resource is marked for re-extraction, and the social bond between the matched friend name and the user is strengthened. The members of the user's social circle with the strongest bond are prioritised to force them to be compared to the name list first. Should no match be found, then the confidence level that the resource contains information relating to the individual is minimal, and as a consequence no further extraction is performed.

Identity information that was found relating to John Smith produces two pieces of information from different resources. After parsing the first resource three names were found and compiled into a list: John Smith, Bobby Moore, and Geoff Hurst. Bobby Moore is one of John's friends from his social circle. Therefore the information contained within the resource is valid and belongs to John. The second resource is parsed, and the names are added to a list: John Smith, and Boris Becker. Boris Becker is not in John's social circle; therefore we cannot tell if this resource contains information belonging to John.

Upon completion of the disambiguation process, the resources found to contain information correlating with the user's social circle are passed on for validation. Resources found to contain more people from the user's social circle are given a higher confidence rating and are therefore presented as being the strongest candidates for containing information relating to the user.

4.5 Validating Identity Data

Once the information has been aggregated and linked together, it is displayed to the user. The information is presented as a mapping showing the resource where the information occurs and within what context the recognition took place, describing the friends from the social circle that were matched around the user. The user is able to validate the results by confirming or rejecting the extracted information. If the returned information does not belong to the user then the extraction process should be performed again, but with less prevalence towards the friends who were responsible for the misidentification.

John is presented with an interactive diagram containing occurrences of his identity on the web. Each occurrence is labelled with the location, and his friends and identity properties that were also found there. Upon inspection of the information, he doesn't agree that one piece of information is about him. He clicks on the resource link to the web page containing the information and realises that he was correct; the information is about another John Smith. He returns to the diagram and informs the system of an incorrect find.

5 Evaluating Identity Disambiguation

The evaluation of the described approach uses a user based study consisting of 60 users each with a variable level of presence on the web, both within the wider web including personal web sites, and the social web including online accounts with social networking sites. This many users yield enough results to perform statistical evaluation. Each user tests the approach in two stages:

The first stage evaluates the pruning of each user's social network into a more compact and relational social circle. The evaluating user is required to analyse whether the derived social circle contains links that they deem to be appropriate with their peers, and mark any errors that exist.

The second stage of the evaluation process involves the evaluation of the disambiguated identity information. The evaluating user verifies all the extracted occurrences of their identity information prior to the disambiguation process, marking whether each resource does contain information describing their identity or not. Once the disambiguation process has completed, the user is then presented with disambiguated information for validation. This second evaluation step is already included in the previously described approach to provide a feedback mechanism for the mining of extracted information. Both steps of evaluation use information metrics to derive the precision, and error rate produced by the system, evaluating the efficiency and effectiveness. User satisfaction also is also evaluated using questionnaires completed by users of the implementation.

6 Conclusions and Future Work

The presented approach is currently being implemented and will be ready for evaluation by prospective users. Through implementation several interesting issues have arisen that will be investigated further. One of the main issues concerns the construction of search queries to the wider web. At present this last stage of the mining process can yield low levels of precision even when the user has declaratively specified their most prevalent identity features. This can be related to the lack of online presence in the search engine domain should a given individual not have any personal web sites, therefore a technique to tune the approach must be considered.

We believe that the described approach presents a novel technique both to the pruning of social networks to form social circles, and also for the disambiguation of individuals using social circles. The former challenge has been addressed in an abstract manner to allow the inclusion of alternative data sources such as blogs, and the sharing of bookmarks, or emails. The algorithm for deriving the social bonds can be applied to analyse any similar social interactions. Images were chosen due to the increase in the social tagging phenomena evident in several social networking sites. Such techniques could be easily adapted to provide risk assessments for users concerned with identity theft and online fraud. By providing a model that specifies what identity features must become accessible for identity theft to be possible, a user could be informed simply by submitting the identity information provided by our approach. This model must be an adaptive model to permit transfer between domains where the criteria for assessment may alter (i.e. Different countries require different identity features).

At present our approach uses two binary classifiers to derive the existence of a social bond between two individuals. We analyse conversations and images to classify individuals as being friends. The usage of such classifiers is assumed to generate a satisfactory social circle, however a further advancement would allow the comparison of classifiers. This would allow further evaluation of the proposed methodology for social circle generation, at present there is little indication of the quality of the classifiers being used.

The methodology that we present for performing the disambiguation process is fairly limited. It is composed of a straightforward comparison technique to derive the string distance between two words. An alternative suitable approach could utilise decision models compiled from the string distances of each name extracted from a single resource, a decision is then reached for a single resource from the analysis of all decision models [13]. The current approach is limited by only requiring a single name match from a resource to denote relation to the individual in question.

Upon completion of the implementation evaluation will take place to derive the effectiveness of generating social circles, and disambiguating between items of identity information. The methodology presented in chapter 6 sufficiently covers the two main objectives of this approach, although both evaluation steps do require an exhaustive process of auditing the generated information to produce a gold standard, particularly in the second stage.

Other future work is to develop a visualisation technique for user information similar to figure 1 detailing the distribution of personal information, and allowing the individual to analyse the information.

References

1. Aswani. N., Bontcheva. K., Cunningham. H.: Mining Information for Instance Unification. In the proceedings of 5th International Semantic Web Conference, Athens, GA, USA (2006).
2. Brickley. D., Miller. L.: FOAF Vocabulary Specification. (2004).
3. Chapman. S., Norton. B., Ciravegna. F.: Armadillo: Integrating Knowledge for the Semantic Web. Proceedings of the Dagstuhl Seminar in Machine Learning for the Semantic Web , 13-18 February (2005).
4. Finin. T., Ding. L., Zhou. L., Joshi. A.: Social Networking on the Semantic Web. The Learning Organisation, vol. 1 , no. 5, pp. 418-435 (2005).
5. Greenwood. M., Iria. J.: Saxon: An Extensible Multimedia Annotator. To Appear in the Proceedings of the International Conference on Language Resources and Evaluation, Marrakech, Morocco (2008).
6. Hamasaki. M., Matsuo. Y., Ishida. K., Nakamura. Y., Nishimura. Y., Takeda. H.: Community Focused Social Network Extraction. Proceedings of 2006 Asian Semantic Web Conference (2006).
7. Han. H., Zha. H., Giles. L. C.: A Model-based K-means Algorithm for Name Disambiguation. Semantic Web Technologies for Searching and Retrieving Scientific Data Workshop. International Semantic Web Conference (2003).
8. Jin. Y., Matsuo. Y., Ishizuka. M.: Extracting Social Networks among Various Entities on Web. The Semantic Web. International Semantic Web Conference 2006. pp. 487-500 (2006).
9. Matsuo. Y., Hamasaki. M., Nakamura. Y.: Spinning Multiple Social Networks for the Semantic Web. Proceedings of the 2006 Asian Artificial Intelligence Conference (2006).
10. Mika. P.: Bootstrapping the FOAF-Web: An Experiment in Social Network Mining. 1st Workshop on Friend of a Friend, Social Networking and the Semantic Web, Galway, Ireland (2004).
11. Mori. J., Tsujishita. T., Matsuo. Y., Ishizuka. M.: Extracting Relations in Social Networks from Web using Similarity between Collective Contexts. International Semantic Web Conference (2006).
12. Neiling. M., Jurk. S.: The Object Identification Framework. In KDD03 Workshop on Data Cleaning, Record Linkage, and Object Consolidation, Washington DC (2003).
13. Rendle. S., Schmidt-Thieme. L.: Object Identification with Constraints. In Proceedings of the Sixth international Conference on Data Mining (December 18 - 22, 2006). ICDM. IEEE Computer Society, Washington, DC. pp. 1026-1031 (2006).
14. Scott. J.: Social network analysis : a handbook. London, Sage (2000).
15. Singla. P., Domingos. P.: Object identification with Attribute-Mediated Dependences. In Conference on Principals and Practice of Knowledge Discovery in Databases (PKDD), pages 297--308, Porto, Portugal (2005).
16. Sweeney. M.: Facebook sees first dip in UK users. http://www.guardian.co.uk/media/2008/feb/21/facebook.digitalmedia (2008)
17. Tejada. S., Knoblock. C. A., Minton. S.: Learning Object Identification Rules for Information Integration. Special Issue on Data Extraction, Cleaning, and Reconciliation, Information Systems Journal. vol. 26. (2001).

18. Wallop. H.: Fear over Facebook identity fraud.
http://www.telegraph.co.uk/news/main.jhtml?xml=/news/2007/07/03/nface103.xml
19. Windley. P. J.: Digital Identity. O'Reilly Media (2005).