Harnessing the Social Web: The Science of Identity Disambiguation

Matthew Rowe and Fabio Ciravegna
OAK Group
Department of Computer Science, University of Sheffield,
Regent Court, 211 Portobello Street,
S1 4DP Sheffield, United Kingdom
{m.rowe, f.ciravegna}@dcs.shef.ac.uk

ABSTRACT

Personal information is one of the most widely available and accessible forms of information on the World Wide Web. The sensitive nature of such information has lead to a rise in malevolent web practices such as identity theft and lateral surveillance. To avoid falling victim to such practices, web users must manually identify all web resources which cite them and if necessary, remove the sensitive information. In this paper we discuss how automated techniques can be deployed to replace such manual processing and how the Social Web can be harnessed in order to collect seed data for use by such techniques. We demonstrate, through an evaluation, the performance of such techniques with respect to web citation sparsity and how automated techniques are able to outperform humans performing the same task.

Keywords

Digital Identity, Disambiguation, Social Web, Machine Learning, Inference Rules

1. INTRODUCTION

The World Wide Web has evolved into an interactive information network, allowing web users to collaborate and share information on a massive scale. A large amount of information now available on the World Wide Web is personal information, which has either been disseminated voluntarily (i.e. a personal web page, profile page) or involuntarily (i.e. telephone directory, electoral register). The sensitive nature of personal information and its widespread visibility has lead to a rise in malevolent web practices such as lateral surveillance and identity theft. In order to avoid falling victim to these practices, web users are forced to find web resources (web pages, data feeds) which *may* contain their personal information and then decide which of those web resources *do*. This decision process disambiguates web resources which are *identity web references* for a given person, however performing this process manually is time consuming and costly. Furthermore, as more information is published on a regular basis, the process must be repeated constantly and handle an ever increasing information load.

In order to overcome the need for the manual discovery of *identity web references* automated techniques are required. To function effectively such techniques require seed data - background knowledge describing the identity of a given person (i.e. their biographical information, their social network) - however producing this seed data manually is expensive and restricted by access to resources: one of the hard problems within the machine learning community.

In this paper we explain how the Social Web platforms can be utilised as a source for seed data. Profile pages are a common feature of such platforms, together with social networks compiled and maintained by users. Theoretical discussions in sociological studies [4, 6] have explored the relationship between real-world identities and digital identity representations which are constructed on Social Web platforms. We present a user study which provides empirical evidence to support such discussions, where the findings from the study show a strong correlation between real identities and digital identity representations. We present two identity disambiguation techniques, each of which employ seed data collected from the Social Web in a unique manner. We explain the science behind the methods and present an evaluation of the techniques against a baseline measure of human processing. The results from the evaluation indicate the suitability of deploying automated techniques, supported by seed data collected from the Social Web, to replace human detection of *identity web references*.

This paper is structured as follows: section 2 discusses digital identity on the Social Web by explaining current studies of the relationship between real and digital identities. Section 3 explains how seed data is collected from the Social Web and presents disambiguation techniques which use this seed data in a bespoke manner. Section 4 evaluates these techniques against a baseline measure of human processing. Section 5 finishes the paper with conclusions drawn from this work.

2. HARNESSING THE SOCIAL WEB

The World Wide Web is now composed of platforms offering an interaction paradigm with the user as the focal point. This evolution into a *Social Web* has produced platforms such as Facebook and MySpace, both of which provide users with the functionality to create profiles which are visible within the public domain. In doing so, Web users who lack the technical expertise and know-how to build and deploy web pages are now provided with an online presence. Work within the field of sociology has explored the use of such platforms in terms of the digital identity representations which users construct. Work by [6] states that Social

Web Science Conf. 2010, April 26-27, 2010, Raleigh, NC, USA.

Web platforms are primarily used to maintain offline relationships, rather than establishing relationships in such a space. Similar work by [4] also claims that Web users use such platforms to reinforce established offline relationships by continuing offline interactions in an online environment. In both [6] and [4] the social network of an individual is considered as an important, and uniquely identifying, facet of a person's identity.

2.1 Comparing Real World and Digital Identities

Although sociological studies provide an insight into the dynamics of digital identity on Social Web platforms, such findings are largely hypothetical and lack empirical evidence to justify their claims. Furthermore if Social Web platforms are to be used as a source for seed data, from which automated disambiguation techniques can be supported, then the relationship between digital identity representations and real-world identities must be assessed more closely. To this end we conducted a user study using 50 participants from the University of Sheffield (staff, students and researchers) with an even split of 25 male and 25 female. The user study was designed to explore the similarity between the digital identities which users construct on a Social Web platform and their real world identities. We chose Facebook as the Social Web platform on which to assess digital identity information - given that this is the most popular and widely used in the UK with 22 million users[1]. To provide a quantifiable measure of the similarity between real-world and digital identities we compared the offline social networks which the participants maintained in the real world with the online social networks that they constructed. The experiment used for the study was broken down into three stages as follows:

The *first stage* involved each participant describing his/her social network in the real world. Each participant was given a web page form with 20 rows, one for each person in the participant's social network. More rows could be added should the participant wish to add more than 20 people. This stage produces the participant's social network containing the strongest and most important (strong-tied) relationships [3], given that these will be the first that come to mind. The *second* stage extracted the social network of each participant from Facebook using the Social Circular Facebook application[2]. Image and conversation data was analysed for behavioural trends: images were extracted that each participant featured in. For each participant the images were analysed to find which other people appeared within the images. In order to analyse the conversation data, all messages each participant shared were extracted. In a similar manner to the image analysis, the messages were analysed to find which people had sent and received them. The *third* stage involved the comparison of the two social networks. Each participant compared his/her real world social network from stage 1 with the digital social network from stage 2. If a person appeared in both the real-world and digital networks then the participant flagged that person. In comparing the networks, the overlap in people present in both networks is derived and therefore gives a quantifiable measure of the replication of offline social networks in an online environment.

[1]http://www.clickymedia.co.uk/2009/10/uk-facebook-user-statistics-october-2009/

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

2.1.1 Measures

To assess the relationship between real world and digital social networks, adapted versions of the information retrieval measures precision and recall were used. The definition of precision from [18] describes a measure which returns *"...the proportion of retrieved information that is relevant..."*. Given that in this study relevant information is the real world social network and retrieved information is the digital social network, precision is redefined as a measure of the *relevance* of a participant's digital social network as follows:

$$relevance = \frac{|real \cap digital|}{|digital|} \quad (1)$$

This metric of relevance therefore measures the proportion of the digital social network that is denoted by strong-tied relationships and is therefore relevant. The definition of recall from [18] defines a measure which returns *"...the proportion of relevant material that is actually retrieved..."*. Given that relevant material is the real world social network and retrieved material is the digital social network, recall is redefined as a measure of *coverage* which gauges the extent to which a participant's real-world social network is replicated within a virtual environment. This is defined as follows:

$$coverage = \frac{|real \cap digital|}{|real|} \quad (2)$$

2.1.2 Findings

Figure 1 shows the results from the study. We found that the majority of each participant's real world social network was replicated in their digital social network. This was characterised by a range of coverage from 0.5 to 1 - where only 1 person out of the 50 participants achieved 1 - with an average coverage measure of 0.77. This therefore indicates that, on average, 77% of a given Social Web user's real-world relationships are repeated in an online environment, thereby suggesting that the digital identity which users build on a Social Web platform mimics their real-world equivalent.

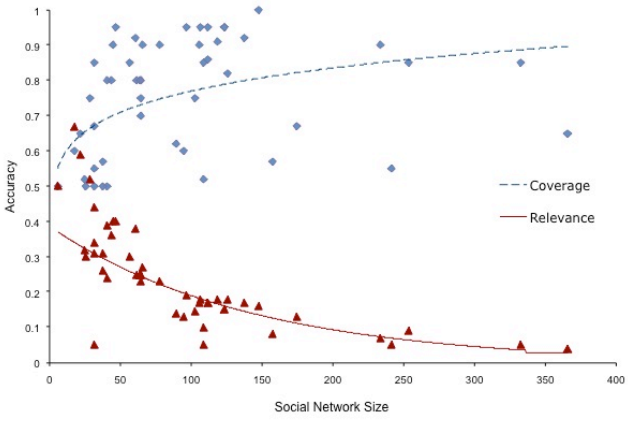


Figure 1: Relevance and Coverage measures from social network comparisons

The results from the study yielded an average relevance measure of 0.23, indicating that, overall, 23% of a given So-

cial Web user's online social network consists of important, strong-tied relationships and the remaining 77% are weak-tied relationships. [5] argues that relationships found on Social Web sites form part of much vaster social networks, and are mainly classed as weak-ties; where the two people are acquainted and nothing more. The findings from the study support such a hypothesis. The results from Figure 1 also demonstrate that coverage tends towards 0.9 as the online social network increases in size. Although the adoption of Social Web platforms by Web users has become a mainstream phenomena, there are still users who resist following their friends and peers in signing up to such services.

The behaviour of each participant within their social network was analysed by collecting image and message data. Images featuring each participant were analysed to see which other people appear in those images, therefore deriving a cumulative count for each member of the participant's social network. A similar approach was used for analysing the messages sent and received by each participant to count how many times each social network member had shared a message with the participant. We observed how the data forms a power law [12] curve where the head of each graph contains a high frequency of people with whom very interactions occur, whereas the tail contains a low frequency of people with whom many images and messages are shared. Within the tail of the graph we found, on average, 92% of those people to be members of the real-world social network. This indicates the large degree to which web users continue their offline lives within a digital space. Our findings are consistent with work comparing online and offline social networks [2, 16, 17] and provides empirical evidence to support the theoretical discussions from [4, 6].

3. IDENTITY DISAMBIGUATION

As we described earlier, the current practice by which web users identify web resources (web pages, data feeds) which are *identity web references* is time consuming and laborious. The identification process is user centric (it concentrates on a single person) and is broken down into two stages: first possible web citations are gathered, and second web resources which cite a person are found by disambiguating the set of possible citations. We are concerned with the latter of these stages, *identity disambiguation*. We now explain how we collect seed data from the Social Web and gather possible web citations for a given person, before moving on to present two complimentary disambiguation techniques, one using a *supervised* approach and the other using a *semi-supervised* approach, and the science behind their methodologies.

3.1 Collecting Seed Data from the Social Web

To support automated disambiguation techniques seed data is collection from Social Web platforms. Our user study indicates how Social Web users construct digital identity representations which mirror their real-world equivalent, therefore by leveraging profile information from such platforms we are provided with identity information. To collect such identity information the APIs for Facebook and Twitter are queried and the returned XML responses are converted into RDF, using the FOAF [1] and GeoNames[3] ontologies to describe biographical and social network information. The use of such Semantic Web technologies provides a useful means by which a consistent interpretation of information can be achieved from disparate sources. Social graphs are produced, describing individual fragments of a person's digital identity, by combining these fragments together we produce a single social graph containing a more complete digital identity representation from which automated disambiguation techniques can be supported. The fragments are combined using inferences over the social graphs to fuse information about the same people together - i.e. identifying a social network member in two different social graphs as referring to the same person. For a more detailed explanation of this technique we refer the reader to [14]. Once combined together we are provided with a single social graph, describing the digital identity of a person distributed across multiple Social Web platforms, for use as seed data. An example social graph looks as follows using notation 3 (n3)[4] syntax for RDF:

```
<http://www.dcs.shef.ac.uk/~mrowe/foaf.rdf#me>
        rdf:type foaf:Person ;
        foaf:name "Matthew Rowe" ;
        foaf:homepage <www.dcs.shef.ac.uk/~mrowe> ;
        foaf:mbox <m.rowe@dcs.shef.ac.uk> ;
        foaf:knows <http://www.dcs.shef.ac.uk/~mrowe/foaf.rdf#fabio> ;
        foaf:knows <http://www.dcs.shef.ac.uk/~mrowe/foaf.rdf#sam> .
<http://www.dcs.shef.ac.uk/~mrowe/foaf.rdf#fabio>
        rdf:type foaf:Person ;
        foaf:name "Fabio Ciravegna";
        foaf:mbox <fabio@dcs.shef.ac.uk>;
        foaf:homepage <http://www.dcs.shef.ac.uk/~fabio> .
<http://www.dcs.shef.ac.uk/~mrowe/foaf.rdf#sam>
        rdf:type foaf:Person ;
        foaf:name "Sam Chapman" ;
        foaf:mbox <sam@dcs.shef.ac.uk> ;
        foaf:homepage <http://www.dcs.shef.ac.uk/~sam> .
```

Our intuition is that a person will appear in a web page with individuals that he/she knows - e.g. work colleagues. Therefore seed data collected from Social Web platforms provides a suitable source for such data, as it is intrinsically social and describes the relationships which a given person maintains offline - as we demonstrated in the previous section. It is worth noting also, that this line of thinking is not a solitary one. Within the literature describing identity disambiguation, social network information is widely used [10, 9].

3.2 Gathering Possible Web Citations

Automated disambiguation techniques require both seed data and a collection possible web citations - we refer to possible web citations as *unlabelled* data. Assessments are then made of this *unlabelled* data, thus disambiguating the web resources which cite a given person. Possible web citations must therefore be gathered for our automated techniques. Search engines remain the most useful and popular web sites on the Web[5], based on such information gateways, we query the World Wide Web (WWW) using Google and Yahoo, and Semantic Web using Watson[6] and Sindice[7] using the person's name, and collect the returned results. The provided seed data is represented using RDF and ontologies to describe the data in a machine-processable form. Web resources returned from queries are converted into the same format using different techniques depending on the format of

[3]http://www.geonames.org/ontology/

[4]http://www.w3.org/DesignIssues/Notation3

[5]http://www.alexa.com/topsites

[6]http://watson.kmi.open.ac.uk/WatsonWUI/

[7]http://sindice.com/

the content within the web resources: if the web resource is a XHTML document containing 'lowercase' semantics such as RDFa or Microformats, then GRDDL [19] is used to glean an RDF model from the page. This produces an RDF model from the OAK people page[8] as follows:

```
<http://oak.dcs.shef.ac.uk/people>
        foaf:topic <http://oak.dcs.shef.ac.uk/people#fabio> ;
        foaf:topic <http://oak.dcs.shef.ac.uk/people#matt> ;
        foaf:topic <http://oak.dcs.shef.ac.uk/people#sam> .
<http://oak.dcs.shef.ac.uk/people#fabio>
        rdf:type foaf:Person ;
        foaf:name "Fabio Ciravegna";
        foaf:mbox <fabio@dcs.shef.ac.uk>;
        foaf:homepage <http://www.dcs.shef.ac.uk/~fabio> .
<http://oak.dcs.shef.ac.uk/people#matt>
        rdf:type foaf:Person ;
        foaf:name "Matthew Rowe" ;
        foaf:homepage <www.dcs.shef.ac.uk/~mrowe> ;
        foaf:mbox <m.rowe@dcs.shef.ac.uk> .
<http://oak.dcs.shef.ac.uk/people#sam>
        rdf:type foaf:Person ;
        foaf:name "Sam Chapman" ;
        foaf:mbox <sam@dcs.shef.ac.uk> ;
        foaf:homepage <http://www.dcs.shef.ac.uk/~sam>
```

If the page contains no such 'lowercase' semantics - i.e. a HTML document - then we construct an RDF graph from information which we find within the page. We use a combination of HTML Document Object Model manipulation - to identify regions of a given web page to extract information from - and then apply Hidden Markov Models to extract person information (i.e. name, email, homepage). An RDF model is built from the extracted information, using the same form as the above RDF snippet in n3 syntax. A full description of this technique, however, falls outside the scope of this paper, instead we refer the reader to [15] for a description of this method.

Following the generation of RDF models for all of the returned web resources, we are provided with a collection of possible web citations in a common format with the seed data. By using Semantic Web technologies in such a way, we achieve a consistent interpretation of information from heterogeneous sources, thus providing our disambiguation techniques with the necessary input data - seed data and possible web citations - from which the techniques can function.

3.3 Identity Disambiguation

3.3.1 Inference Rules

Using the seed data we are able to build rules which *infer* a web resource as citing a given person. Rules provide a means of logically inferring conclusions based on the presence of information. Richard Jeffrey [8] states that: *"Logic is the science of deduction. It aims to provide systematic means for telling whether given conclusions do or not follow from given premises."* The grounding of rules in logic reflects the cognitive process by which we are able to decide whether a web page refers to the topic, or person, we are interested in or not. Using *a prior* knowledge of the topic, a person assesses the available information and derives a conclusion. In terms of *identity disambiguation* this cognitive process is mimicked by taking the known seed data - describing the digital identity of a given person - and forming rules which provide a logical basis from which a conclusion can be derived - i.e. as to whether the web resource cites the individual or not. We now explain how we build our collection of rules to infer *identity web references*.

Rules are built by first extracting RDF instances, for use as features, from the RDF model of the provided seed data. This seed data contains resources which are indicative of the person whose *identity web references* we wish to detect; biographical information, social network members and geographical locations. Therefore we use RDF instances found within the given RDF model as features by extracting subgraphs - where each subgraph represents a single resource. To extract features we extract the *Resource Leaves* from a graph (G) surrounding a given resource (r) as an RDF instance using the following formal construct - this produces a tripleset of the instance including only triples where the object of the triple is not the subject of another:

$$RLS_G(r) = \{< r, p, o > \mid < r, p, o >\in G \wedge \nexists p', o' < o, p', o' >\in G\} \quad (3)$$

Once we have extracted all the features, and therefore the RDF instances, from the seed data, a rule is built for each instance as follows: a skeleton rule is created which detects the presence of the person's name within a web resource - this is to filter out any web resources which may have been returned by accident. Then for each triple in the RDF instance description, we add this triple to the rule's antecedent, removing the literals and resources and replacing them with variables. This allows the knowledge pattern to be matched, which associates information identifying the person with information in a given web resource. If the predicate is defined as being inverse functional, meaning that it denotes a unique property of the instance, then we create a separate skeleton rule, add the triple whilst replacing the literals and resources with variables, and add this rule to our collection of rules. We repeat this process for all RDF instances within the seed data, and return a collection of rules. An example rule produced using this process looks for the name of a person whose web citations are to be found, together with their email address in a given web resource. Using SPARQL rule syntax[13] this rule would look as follows:

```
CONSTRUCT {
  <http://www.dcs.shef.ac.uk/~mrowe/foaf.rdf#me> foaf:page ?url
}
WHERE {
  <http://www.dcs.shef.ac.uk/~mrowe/foaf.rdf#me> foaf:name ?n .
  <http://www.dcs.shef.ac.uk/~mrowe/foaf.rdf#me> foaf:homepage ?h .
  ?url foaf:topic ?p .
  ?p foaf:name ?n .
  ?p foaf:hompeage ?h
}
```

The knowledge patterns included within the *antecedent* (the body) of the rule depends on the features of the seed data. We create rules based on first-order logic, thus allowing the inclusion of variables within a rule, this in turn allows patterns to be built without the need for the inclusion of literals (i.e. explicitly defining the inclusion of a person's name). The intuition behind this strategy is that information describing social network members in the seed data will vary such that one person may be described using only his/her name, whereas another will contain the name and email address. By using only the structure of this information both of these patterns are covered, providing more generalised rules for inferring web citations.

[8] http://oak.dcs.shef.ac.uk/people

3.3.2 Self-training

Inference Rules use a *supervised* strategy by constructing rules based on the provided *labelled* seed data and then applying them to the *unlabelled* data - the collection of possible web citations. This places a requirement on the seed data: it must contain sufficient feature coverage to enable the detection of all web citations. Instead it may be the case that the person, whose web references are to be found, features in many web resources with his/her colleagues, yet the user profiles on the Social Web platforms do not contain these relationships. One solution to this problem is the use of a *semi-supervised* disambiguation technique which use both *labelled* - the provided seed data - and *unlabelled* data - to learn from disambiguation decisions.

To investigate such an approach we have explored the use of Self-training, a *semi-supervised* machine learning strategy. This technique treats the disambiguation task as a binary classification problem (i.e. Does the web resource mention person X or not?), thus restricting web resource labelling to two class labels; positive - cites the person - and negative - does not cite the person. Self-training begins by training a machine learning classifier using the provided seed data as positive training data and web resources from the *unlabelled* data, which are identified as negative examples, as the negative training data. The classifier is then applied to the *unlabelled* data and each web resource is labelled as either citing the person (positive) or not (negative), coupled with a confidence score. The training data is then enlarged with the web resources from the *unlabelled* data which have the strongest confidence scores for both the positive and negative labels. The classifier is then retrained using the training data and reapplied to the *unlabelled* data, thereby *Self-training* the classifier.

The use of such an approach mirrors the process by which humans learn from their observations in order to make future decisions. Although the provided seed data may only contain a facet of a person's identity, by learning from classifications, the technique is able to detect web citations which would previously have been missed. We now explain the technical aspects of the approach including feature extraction and the Self-training strategy.

3.3.2.1 Feature Extraction.

The seed data and the collection of possible web citations are provided as a collection of RDF graphs - where a single web resource is represented by an individual RDF graph. To enable such graphs to be used as machine learning instances, features must be extracted for analysis by classifiers. To extract features, the Resource Leaves construct from Equation (3) is used. As we iteratively classify web resources, our approach *learns* additional instances (events, publications) within the classified RDF graphs - given that these instances may not be present within the collected seed data.

3.3.2.2 Feature Similarity Measures.

By using RDF graphs as machine learning instances, and RDF instances as features, we able to explore the effects of several different RDF graph comparison techniques on the disambiguation task. We investigate the effects of three techniques to detect equivalence between RDF instances as follows:

1. *Jaccard Similarity*: Matches RDF instances when their graphs are completely identical: including all literals, resources and predicates.

2. *Inverse Functional Property Matching*: Matches RDF instances as equivalent when a given property value matches in both instances and the property is defined as inverse functional (i.e. two person instances have the same email address).

3. *RDF Entailment*: According to Pat Hayes' Interpolation Lemma[9] and work in [11], we match two RDF instances: g and h, when the graph of h entails g ($h \models g$) (every model of h is also a model of g).

3.3.2.3 Mapping RDF Graphs into a Machine Learning Dataset.

In order to use RDF graphs as learning instances from which machine learning classifiers can be trained, we must represent these models as binary feature vectors ($\vec{x}$) where each feature represents a distinct RDF instance. Therefore we map each of the RDF graphs into a machine learning dataset for learning using the following three steps:

1. The graph of each RDF instance is extracted from the RDF model.

2. Each extracted subgraph is representing as a tripleset.

3. Binary feature vectors are built using a feature indexer (i.e. $\overrightarrow{x_i}$ where the index i denotes a given feature) by deriving indexes for the features of the given RDF graph using one of the described similarity measures: if a feature is matched then the relevant index is returned, otherwise a new index is created.

Self-training functions as a wrapper algorithm in that machine learning classifiers are used as "blackboxes", inputting the *labelled* and *unlabelled* data, and outputting class labels and classification confidence levels. By using such a strategy we are able to investigate the performance of different machine learning classifiers on the disambiguation process. We compared three different classifiers, two discriminative: Perceptron and Support Vector Machines, and one generative: Naive-Bayes. Perceptron and SVM construct their discriminative hyperplanes within the feature space based on the labelled instances within the dataset, and Naive Bayes derives the conditional probabilities from labelled instances and their features to build the initial generative model.

3.3.2.4 Generating Negative Training Data.

Our classification task is binary, we must train a classifier to differentiate between a web resource which cites a given person and one that does not. Seed data collected from the Social Web is positive, it contains features which describe a given person such as his/her biographical information and social network. We must therefore provide negative data to couple this positive data from which a classifier can learn. To generate negative training data we use a Rocchio classification model [20]. This method uses a vector space model by constructing two prototype vectors: a positive vector and a negative vector. The vectors act as centroids within the

[9] http://www.w3.org/TR/2004/REC-rdf-mt-20040210/#entail

vector space, such that an unlabelled instance which has more features in common with a given centroid will appear closer to it within the vector space. In such a setting - without pure negative data - the negative prototype vector is built from the *unlabelled* data. Instances from the *unlabelled* data are then compared with each prototype vector, if the cosine similarity with the negative prototype vector is greater than the positive vector, then the instance is added to the negative set. All instances within the negative set are then ranked, based on the strength of their cosine similarity with the negative prototype vector, and the top n vectors are kept as negative training set, where n is the number of instances in the positive set - given that these vectors appear away from the positive prototype vector in the vector space.

3.3.2.5 Self-training.

The Self-training strategy, as shown in Algorithm 1, functions by first using the seed data to train a classifier. Unlablled instances are classified using the initially trained classifier and the instances are ranked based on their classification confidence. Training data - the positive and negative sets - are enlarged with the single strongest positive and negative instances from the unlabelled data based on these rankings. The strongest instances are chosen to ensure that no misclassifications are made. Such is the nature of Self-training that any misclassifications have the tendency to *snowball* by reinforcing themselves and leading to poor disambiguation results. The Self-training process is repeated until all the *unlabelled* data has been assigned labels and belongs in either the positive or negative training sets. These sets are then used as the final classifications.

Algorithm 1 ItEnl(P,N,U) : Iterative Self-training based on enlargement of training data according to classification rankings. P is the positive set, N is the negative set and U is the universal set

Input: P, N and U
Output: P and N
1: **while** $U \neq \emptyset$ **do**
2: Train Ψ using P and N
3: Classify U using Ψ
4: Rank U based on classification confidences
5: Enlarge training sets by strongest k positive (U_k^P) and negative (U_k^N) instances
6: $P = P \cup U_k^P$
7: $U = U - U_k^P$
8: $N = N \cup U_k^N$
9: $U = U - U_k^N$
10: **end while**
11: **return** P and N

4. EVALUATION

Evaluation of our disambiguation techniques compares the accuracy of a *supervised* method against a *semi-supervised* method. Our intuition prior to performing the evaluation was that a *supervised* approach would fail to detect a large portion of web citations, yet ensure that the detected citations where correct. Conversely the *semi-supervised* method would return web citations which may contain errors, yet ensure that a large number of web citations were found.

4.1 Experimental Setup

4.1.1 Evaluation Measures

Inference Rules and Self-training divide a set of possible web citations into two distinct sets; those that refer to a given person (*positive*) and those that do not (*negative*). In order to evaluate the performance of our techniques, accuracy is assessed using information retrieval metrics [18] featuring two sets of documents: A denotes the set of relevant web resources and B denotes the set of retrieved web resources. Precision measures the proportion of web resources which are labelled as citing a given person and are true references ($precision = |A \cap B|/|B|$). Recall measures the proportion of true references that were retrieved ($recall = |A \cap B|/|A|$). F-measure provides the harmonic mean between the two measures as follows:

$$f - measure = \frac{2 \times precision \times recall}{precision + recall} \quad (4)$$

4.1.2 Dataset

A dataset was compiled using 50 members of the Semantic Web and Web 2.0 communities as participants. For each person seed data was gathered from Facebook and Twitter. Possible web citations were gathered by querying the WWW and the Semantic Web and the first 100 results from each query were downloaded. The dataset, following removal of duplicates, contained 17300 web resources with $\sim$ 346 web resources to be analysed for each participant. RDF models were generated from each web resource using the described techniques. We built a gold standard by manually labeling the dataset. We also assessed the performance of the described disambiguation techniques based on web presence levels of the participants, where we defined *web presence* as the proportion of web resources that refer to the participant given as a percentage. (e.g. if participant A appears in 50 of the 350 collected web resources, then the web presence level of that participant is 14%).

4.1.3 Baseline Measure: Human Processing

Human disambiguation of *identity web references* was used as a baseline measure. This measure was derived using a group of 12 raters who manually processed a portion of the dataset. Three different raters identified web citations for each evaluation participant, thereby providing three distinct sets of results. Interrater agreement was then used to find the agreement of results between the raters according to techniques in [7], thus providing measures for precision, recall and f-measure. The average of the three separate rater agreement measures for each participant was then computed.

4.2 Results

4.2.1 Inference Rules

Inference Rules achieve high levels of precision with respect to Self-training and human processing (as shown in Table 1). This is due to the strict nature of the rules, requiring a web resource's knowledge structure to exactly match the tripleset of the graph pattern in the body of the rule. A consequence of this strict matching is poor levels of recall. Inference Rules utilises a *supervised* technique such that only *labeled* data (seed data) is used to construct the rules, this inhibits the technique's ability to learn features

Table 1: Accuracy of Inference Rules and permutations of Self-trained classifiers and feature similarity measures with respect to Human Processing

		Precision	Recall	F-Measure
	Rules	0.955	0.436	0.553
Perceptron	Entailment	0.629	0.905	0.728
	IFP	0.630	0.878	0.715
	Jaccard	0.651	0.820	0.700
SVM	Entailment	0.613	0.910	0.731
	IFP	0.628	0.864	0.711
	Jaccard	0.755	0.695	0.691
Naive Bayes	Entailment	0.629	0.629	0.628
	IFP	0.649	0.652	0.649
	Jaccard	0.713	0.619	0.633
	Humans	0.765	0.725	0.719

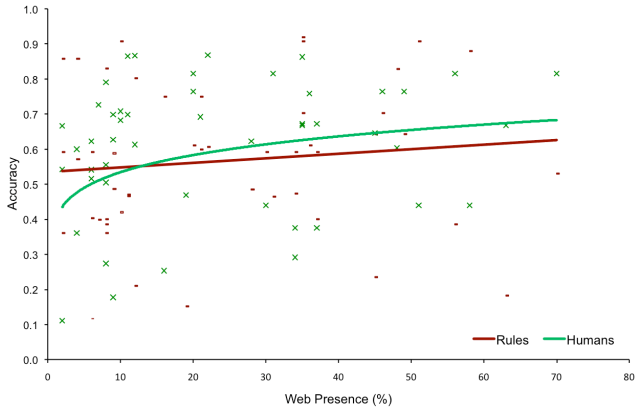


Figure 2: F-Meaure levels of Inference Rules and Humans with respect to Web Presence levels

from identified web citations. In comparison to Self-training, recall levels are significantly lower as a result of this inability to learn. Human processing of the dataset ensures that precise results are achieved, whilst, in certain cases, missing web citations. We believe that this is akin to a "*needle in a haystack*" problem if the web presence level of an individual is very low. As the results indicate the cognitive process of discovering these low-levelled individuals is difficult, however using an automated technique, such as Rules, offers a suitable approach to discovering sparse data with high precision levels.

With respect to web presence levels, as shown in Figure 2, the performance of inference rules improves as the web presence level increases. Similarly, human processing also improves as the web presence level increases. Where evaluation participants have a low web presence level, humans perform poorly. As we mentioned previously, humans are unable to detect a low number of web citations due to the sparsity of citations providing humans with a large portion of web content to look through until a web citation is found, and fewer features identifying a given person to learn from. Conversely inference rules are not effected by web citation sparsity, and instead outperform humans at low web presence levels.

4.2.2 *Self-training*

Self-training, unlike inference rules, achieves greater recall levels for the sake of precision. In the case of SVM combined with Entailment, a recall level of 0.91 is achieved, thereby ensuring that 91% of web citations are returned. Overall the use of Entailment, and its ability to permit variance in features when matching, yields higher recall levels than the two alternative measure. Conversely, the strict graph matching required by Jaccard similarity ensures that precision levels are higher than both Entailment and IFP, whilst leading to a reduction in recall as web citations are missed where Entailment and IFP detect such references.

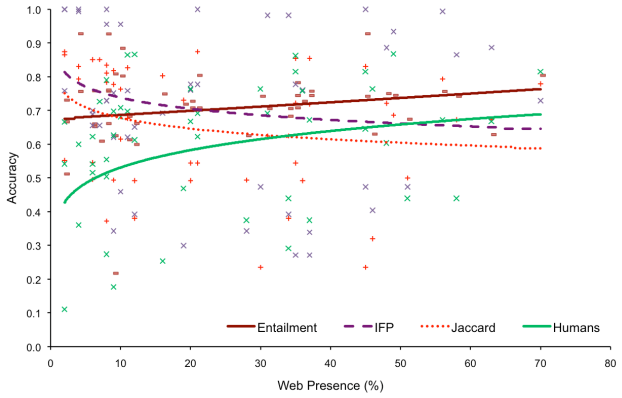
In terms of the classifiers, there is little variance between the two discriminative classifiers: Perceptron and SVM. Each build hyperplanes in the vector space to enable classification of the *unlabelled* data into either the positive or negative sets. The use of a generative model however, denoted by the performance of Naive Bayes, yields poorer f-measure levels than the discriminative classifiers. Such results indicate the suitability of classifiers which attempt to learn a mapping function for *unlabelled* data rather than an overall density model.

Similar to Inference Rules, Figure 3 demonstrates the ability of Self-training to yield high f-measure levels for low web presence levels in comparison to humans. The use of entailment as a feature similarity measure ensures that, as the web presence level increases, performance remains the same. In the case of Perceptron and SVM combined with Entailment we observed how humans are constantly outperformed regardless of web presence levels. Such results demonstrate the effectiveness of self-trained discriminative classifiers for disambiguating *identity web references*. F-measure levels of all classifiers when combined with either IFP or Jaccard degrade as web presence levels increase.

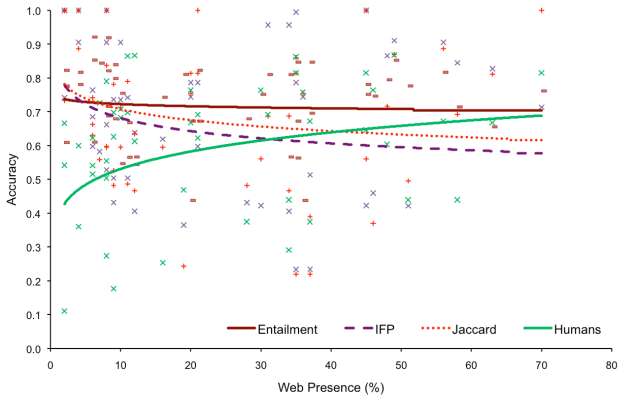
5. CONCLUSIONS

In this paper we have described two distinct disambiguation techniques, both of which provide a novel means by which human disambiguation of *identity web references* can be replaced. Inference Rules employs a *supervised* strategy thereby yielding high precision levels, guaranteeing that web resources inferred as web citations are correct, whilst achieving poor recall levels due to the technique's inability to learn from its inferences. Self-training, using a *semi-supervised* strategy, achieves high recall levels, ensuring that a large portion of web citations are identified, whilst yielding lower precision levels. Overall, and with respect to the human processing of the dataset, we found Self-training to be the most suitable means by which the manual process of *identity web references* disambiguation could be replaced.

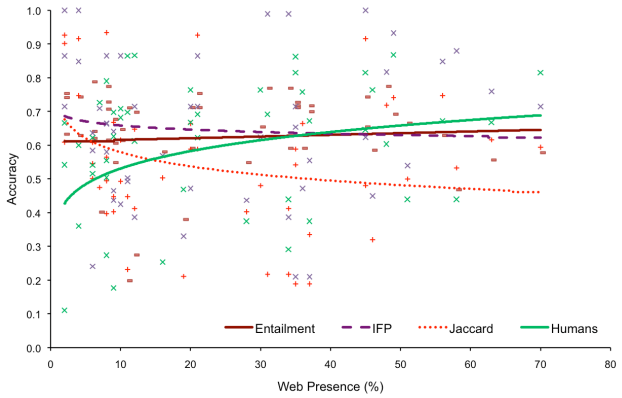
A natural progression of the presented work is to combine the two techniques. We plan to conduct experiments which primarily use Inference Rules to increase the available positive data, given the high precision levels of this method. Using this increased positive data, Rocchio classification will then be applied to generate negative examples. The positive and negative training data will then be used to train a machine learning classifier, and apply Self-training to improve the learnt hypotheis and increase recall levels. We believe that such a process is akin to the approach by which humans increase their primary understanding of a given topic: by reading up about the subject, before moving on to find more information. Without this initial phase, information detection will be poor, instead a more substantial overview of the topic ensures that additional, related information can be identified.



(a) Perceptron



(b) SVM



(c) Naive Bayes

Figure 3: F-Measure levels of Self-trained classifiers and Humans with respect to Web Presence levels

6. REFERENCES

[1] D. Brickley and L. Miller. FOAF vocabulary specification. Technical report, FOAF project, May 2007. Published online on May 24th, 2007 at.

[2] B. Clarke. Friends forever: How young adolescents use social-networking sites. *IEEE Intelligent Systems*, 24:22–26, 2009.

[3] J. Donath and D. Boyd. Public displays of connection. *BT Technology Journal*, 22(4):71–82, October 2004.

[4] N. B. Ellison, C. Steinfield, and C. Lampe. The benefits of facebook friends: Social capital and college students' use of online social network sites. *Journal of Computer-Mediated Communication*, 12:1143–1168, 2007.

[5] R. Gross, A. Acquisti, and H. J. Heinz, III. Information revelation and privacy in online social networks. In *WPES '05: Proceedings of the 2005 ACM workshop on Privacy in the electronic society*, pages 71–80, New York, NY, USA, 2005. ACM.

[6] J. Hart, C. Ridley, F. Taher, C. Sas, and A. Dix. Exploring the facebook experience: a new approach to usability. In *NordiCHI '08: Proceedings of the 5th Nordic conference on Human-computer interaction*, pages 471–474, New York, NY, USA, 2008. ACM.

[7] G. Hripcsak and A. S. Rothschild. Agreement, the f-measure, and reliability in information retrieval. *Journal of American Medical Informatics Association*, 12(3):296–298, 2005.

[8] R. Jeffrey. *Formal Logic*. McGraw-Hill, 1989.

[9] L. Jiang, J. Wang, N. An, S. Wang, J. Zhan, and L. Li. Two birds with one stone: A graph-based framework for disambiguating and tagging people names in web search. In *18th International World Wide Web Conference (WWW2009)*, April 2009.

[10] D. V. Kalashnikov, Z. Chen, S. Mehrotra, and R. Nuray-Turan. Web people search via connection analysis. *IEEE Transactions on Knowledge and Data Engineering*, 20(11):1550–1565, 2008.

[11] S. Liu and J. Zhang. Retrieving and matching rdf graphs by solving the satisfiability problem. In *WI '06: Proceedings of the 2006 IEEE/WIC/ACM International Conference on Web Intelligence*, pages 510–513, Washington, DC, USA, 2006. IEEE Computer Society.

[12] M. E. J. Newman. Power laws, pareto distributions and zipf's law. *Contemporary Physics*, 46:323, 2005.

[13] E. Prud'hommeaux and A. Seaborne. SPARQL Query Language for RDF. Technical report, W3C, 2006.

[14] M. Rowe. Interlinking distributed social graphs. In *Proceedings of Linked Data on the Web Workshop, WWW09*, April 2009.

[15] M. Rowe. Data.dcs: Converting legacy data into linked data. In *Proceedings of Linked Data on the Web Workshop, WWW2010*, 2010.

[16] K. Subrahmanyam, S. Reich, N. Waechter, and G. Espinoza. Online and offline social networks: Use of social networking sites by emerging adults. *Journal of Applied Developmental Psychology*, 29(6):420–433, November 2008.

[17] A. L. Traud, E. D. Kelsic, P. J. Mucha, and M. A. Porter. Community structure in online collegiate social networks. Technical Report arXiv:0809.0690, American Physical Society, Sep 2008.

[18] C. J. van Rijsbergen. *Information Retrieval*. Butterworths, London, 2nd edition, 1979.

[19] W3C. Gleaning resource descriptions from dialects of languages (GRDDL). W3c recommendation, W3C, September 2007.

[20] G. Wang, Y. Yu, and H. Zhu. Pore: Positive-only relation extraction from wikipedia text. In *ISWC/ASWC*, pages 580–594, 2007.