Identifying Individuals using Identity Features and Social Information

Matthew Rowe
Web Intelligence Technologies Lab
Department of Computer Science
University of Sheffield, UK
m.rowe@dcs.shef.ac.uk

Abstract: This paper presents an approach for the disambiguation of individuals using the semantics of identity and social circles. Identity information is extracted and integrated to provide a presentation of existing identity information currently on the web relating to a given individual. Communities are used to discover identity resources, share them socially, and critique the resources based on the accuracy and volume of their content. As motivation for this research issues concerning identity theft, online fraud and cyber stalking are considered, where the growth of the social web has contributed to the rise in such practices. Monitoring identity on the web would go some way to addressing these issues.

Keywords: community, disambiguation, identity, integration, semantic web, social web

1. Research Problem Overview

The motivations behind my research have been the growth of the social web over the past 2 years, and the rise in online identity theft and cyber stalking [1]. The work presented in this extended abstract provides an approach that identifies, extracts and integrates occurrences of identity information from the web. The approach is split into three parts: Extracting identity information and social network mining, integrating identity information, and resource discovery.

The semantics of identity are used to perform the extraction process by recognising identity features within a web resource and extracting the content relating to these features. Social networks are mined and pruned to derive the social circle that an individual belongs to using social content such as socially tagged images and conversation data. The social circle is then used to recognise identity information on the web, by parsing text content from web pages to derive the names of individuals, and comparing them against the social circle.

Disambiguating between information describing different individuals using features of identity and the pruned social circles is used to aid with the integration of identity information. Social circles are used to provide a useful technique to disambiguate individuals by their acquaintances. Resource discovery is supported using a community of users by sharing resources containing identity information; the

community is responsible for rating, and critiquing the resources, and discovering more identity resources upon which they are shared with the community.

Semantic technologies provide useful techniques and methods to identify individuals. An individual's identity will be formalised to encapsulate uniquely identifiable properties, reasoning is performed to derive additional information relating to the individual. The extraction of information will be carried out using a populated ontology containing personal details belonging to an individual. Identity information from various resources will be integrated together, and disambiguated according to their semantics. Involving a community to aid with extraction, by sharing vital resources, will also use social technologies. Social feedback methods will also be used for feature selection by allowing an individual to select the properties of their identity they believe to be the most prevalent.

The work presented within this abstract employs a both a combination of existing methodologies such as social networking mining, and original techniques for disambiguation of individuals. The motivation of this work places emphasis on monitoring online information, and providing risk assessments to concerned users, those who wish to discover what information exists relating to them. This extended abstract is structured as follows: Section 2 discusses the state of the art divided into the three previously mentioned areas. Section 3 is similarly divided into three areas outlining the work plan by explaining the various investigations being conducted. Section 4 explains the evaluation methods to be used, and section 5 presents concluding remarks.

2. Related Work and Contributions

Extracting Identity Information and Social Network Mining

The state of art on information extraction distributed throughout various textual sources includes standard information extraction mechanisms and approaches that can be applied to identity data such as classic wrapper induction [6] for information extraction from structure sources, and more up to date approaches such as support vector machines [4] for the extraction of information from free text. My work has focused on the semantics of identity, what properties are more prevalent than others, and how the community can influence the prevalence of identity features when extracting identity information.

The state of the art within the area of social network mining commonly uses techniques such as entity co-occurrence [8], [3] for extracting the strengths and ties among individuals. A seed set of entities is produced that models the names of individuals that commonly co-occur together in the same context. State of the art work presented in [9] also demonstrates how social networks can be mined from Semantic description files via FOAFnet. Advancement on previous work is demonstrated in [10] and [5] where relations between individuals within the same social network are not only identified but are also given labels denoting the tie between them. My research will investigate the inference of relationship strengths that bind relationships. Using such methods I am investigating how social cliques and circles play an important role in identifying individuals, similar to real life identification through acquaintances.

Integrating Identity Information
Social networks from two separate sources are integrated together in [11], enabling the integration of identity information. Disambiguation is performed using a context sensitive algorithm, considering the properties and relations surrounding the entities in question. Utilising both community selected prevalent identity features and social circles, my work will contribute to the state of the art by offering a social approach to the feature selection problem and disambiguating objects using social bonds. By incorporating a user within the disambiguation process bootstrapping is performed by allowing the user to select the features of their identity that they believe provide their most unique features.

Resource Discovery
Work in [7] describes a framework to allow users to share information within a community portal by adding and removing metadata from already existing information. My work contributes to the state of the art by sharing resources containing identity information, and supervising the process of information extraction by allowing individuals to select their prevalent identity features. User based feedback is used enabling individuals to rate resources based on their usability and accuracy of retrieved content.

3. Work Plan

Extracting Identity Information and Social Network Mining
Regarding the discovery of identity information I have focussed on the semantics of identity. I am currently defining a manually created ontology encapsulating the properties of identity, and able to capture an individual's identity properties. Future work includes the designing of a methodology to efficiently discover identity information from the wider web. In order to extract identity information I have investigated the use of support vector models for community supported identity extraction from semi-structured web resources. Following work will investigate focussed crawling using specialised web queries, indexing a subset of the web, and community supported blocking mechanisms.

To mine social networks I have created several mechanisms to extract social network data from social networking sites that will be used to seed sets for a wider mining process. A working prototype of this approach is available for use[1]. The next stage of work will investigate the pruning of social network data to derive social cliques and circles, and the investigation of the effects and application of identity discovery through the use of social cliques and clusters.

Integrating Identity information
The work I have done to date has investigated the integration of object data from heterogeneous web resources, and the disambiguation of objects. The disambiguation of objects can then be applied to identity information. Further work will investigate

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

the use of pairwise decision models, and community supervision of integration where I anticipate that the use of feature selection using decision models will be an important aspect of disambiguating identity information.

Resource Discovery
To date I have researched approaches to share resources through social bookmarking tools and similar web applications. Future work will investigate the adoption of collective intelligence approaches when sharing identity resources, and the use of feedback mechanisms to rate and prioritise identity resources.

4. Evaluation

Extracting Identity Information and Social Network Mining
Identity extraction will be evaluated for precision, recall and error rate. Evaluators of the approach will be required to find all occurrences of their identity manually to create a gold standard, detailing what identity details are present. Extracted identity information will then be evaluated against the gold standard. Precision and error rate will be used to evaluate extracted social networks through comparison against real life social networks for each individual performing the evaluation. Evaluators will also be required to validate relationship strengths within their social circle, and the members of their social circle.

Integrating Identity Information
Evaluation will be performed using exhaustive user testing to derive the precision, recall and error rate. Each individual will verify the all information items relating or not relating to them, and the integrated information to identify incorrectly integrated information and incorrectly excluded data.

Resource Discovery
The evaluation of the sharing mechanism will be performed using social studies of users when using the approach, testing for user satisfaction through questionnaires. The approach should provide a useful means for sharing identity resources through an easy to use, yet effective methodology.

5. Conclusions

This paper presents an overview of the research that I am currently conducting. The research when broken down into the three areas can be summarised further to include information extraction, information integration and sharing mechanisms. The first areas being largely concerned with existing semantic web technologies and their adaptation to these areas of work. The third area is largely centred around the social web, and current sharing mechanisms being employed by social web sites and services. A combination of both semantic web and social web technologies would incorporate the user at a more intrinsic level by supervising the information extraction and integration stages.

The state of the art will be contributed to mainly in the area of social network mining, and the use of the derived social circles to disambiguate individuals. The work that I have carried out so far has been largely concerning research within each separate area of work, and I have reached a position to begin implementation of an approach to disambiguate individuals using social circles. Information retrieval metrics have been chosen to evaluate the extraction and integration of information because of their widespread usage in similar applications. Evaluating for user satisfaction was selected when evaluating the discovery of resources in order analyse the effectiveness of the sharing mechanism.

In relation to the addressing of issues such as identity theft, online fraud, and cyber stalking, the presented approach provides a methodology to monitor the occurrence of identity information, and using semantic technologies reasoning can be performed to asses the risk of an individual being a victim of such practices. The approach must have sufficient flexibility to allow assessments to be made based on alternative requirements, such as different identity features. Disambiguating identity information performs a crucial role when assessing the risk of identity theft, any information that is wrongly classified could contribute to providing a false analysis.

References

1. Atkinson. S., Jagodzinski. P., Johnson. C., Phippen. A. D.: Personal Privacy: Exploitation or Control through Technology. Proceedings of the Sixth International Network Conference (INC2006), Plymouth, UK. 11-14 July, pp. 269-276 (2006).
3. Hamasaki. M., Matsuo. Y., Ishida. K., Nakamura. Y., Nishimura. T., Takeda. H.: Community Focused Social Network Extraction. Proceedings of 2006 Asian Semantic Web Conference (2006).
4. Huang. T-M., Kecman. V., Kopriva. I.: Kernel Based Algorithms for Mining Huge Data Sets, Supervised, Semi-supervised, and Unsupervised Learning, pp. 260. Springer-Verlag, Berlin, Heidelberg (2006).
5. Jin. Y., Matsuo. Y., Ishizuka. M.: Extracting Social Networks among Various Entities on Web. The Semantic Web. pp. 487-500. International Semantic Web Conference (2006).
6. Kushmerick. N., Weld. D., Doorenbos. R.: Wrapper induction for information extraction, IJCAI-97 (1997).
7. Maneewatthana. T., Wills. G., Hall. W.: Adaptive Personal Information Environment based on the Semantic Web. In: HT 2005 - ACM Workshop on Hypertext and Hypermedia, 6-9 September. Salzburg, Austria (2005).
8. Matsuo. Y., Hamasaki. M., Nakamura. Y.: Spinning Multiple Social Networks for the Semantic Web. Proceedings of the 2006 Asian Artificial Intelligence Conference. (2006).
9. 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).
10. Mori. J., Tsujishita. T., Matsuo. Y., Ishizuka. M.: Extracting Relations in Social Networks from the Web Using Similarity Between Collective Contexts. Proceedings of ISWC 2006 (2006).
11. Aleman-Meza. B., Nagarajan. M., Ding. L., Sheth. A., Arpinar. B., Joshi. A., Finin. T.: Scalable semantic analytics on social networks for addressing the problem of conflict of interest detection. ACM Transactions on the Web Journal. Vol. 2. (2008).