# Harvesting Entity-relation Social Networks from the Web: Potential and Challenges

Saeed Amal
University of Haifa
samal@campus.haifa.ac.il

Tsvi Kuflik
University of Haifa
tsvikak@is.haifa.ac.il

Einat Minkov
University of Haifa
einatm@is.haifa.ac.il

## ABSTRACT

We describe a graph-based entity profiling system (GBEP) that extracts information about persons of interest from the Web and uses this information to construct a joint social graph. GBEP then employs graph-based measures to assess inter-personal relatedness, performing *social recommendation*. Importantly, GBEP provides detailed explanations for its suggestions in the form of relational connecting paths. Initial positive results were obtained for recommending related conference participants to each other using a joint social graph constructed for this purpose.

## CCS CONCEPTS

• **Information systems → Collaborative and social computing systems and tools**; • **Computing methodologies →** *Information extraction*;

## KEYWORDS

Graph-based Recommendation; Information Extraction

## 1 INTRODUCTION

Relational information about an entity of interest can be represented as an entity-relation graph. For example, a graph representing *Albert Einstein* would include links to nodes denoting entities or concepts such as 'University of Zurich', or 'Quantum Theory'. While relevant information about personas like Einstein is available in structured form from public resources like Wikipedia, for most persons, their profiles must be constructed from raw Web data. We apply information extraction techniques to automatically construct an *entity profile* in response to a query that specifies a *person* name. Importantly, multiple personal profiles can be readily unified into a joint graph, comprising a heterogeneous entity-relation social network. It is then possible to address complex queries such as: "Who are the *persons* most related to *person* $p$?", or, "*how* are the persons represented by nodes $p$ and $q$ related?"

Our prototype of GBEP automatically extracts relational information about persons of interest from on their homepages. We then unify the graphs of multiple personal profiles to form a social network, in which similarity assessments and recommendation can take place. We report preliminary results of *social recommendation* using GBEP: ranking the participants of the IUI'15 conference by their relatedness to each other. Such application may promote the generation of new social and professional ties.

Previously, Adamic and Adar [1] extracted personal profiles from the Web, however their focus was on social community exploration, while we are interested in social recommendation. Accordingly, we place emphasis on presenting detailed supporting evidence to the user in the form of labeled and weighted relational connecting paths. Another recent related research pursues social recommendation in academic conferences [2], but they only consider direct co-authorship as indicator of social affinity. Our targeted social similarity suggestions are more extensive in that they involve diverse entities as well as indirect relations. Initial feedbacks suggest using GBEP is engaging and surprising.

## 2 GRAPH-BASED ENTITY PROFILING

*Personal profile construction.* Given a person name $t$ and her homepage, we build a graph profile $G_t$ which displays her connections with related typed entities $E_t$, which we identify from the semi-structured homepage. Entity mentions are often available in structured form being tagged with hyperlinks. In order to increase coverage, we also apply the Stanford named entity tagger [1] to identify *person*, *location* and *organization* entity name mentions that appear within the unstructured text. The target person $t$ and the related named entities $E_t$ are represented as typed nodes in $G_t$,[2] having a direct edge link from $t$ to each related entity $e \in E_t$ (i.e., the personal graph is star-shaped.) Ideally, the graph edges should be assigned a semantic *relation type*, $r(t, e)$, that succinctly describes the identified inter-entity association. In our case study, we consider high-level relation types that characterise scholars, including *education* (i.e., *studied-at(t,e)*), *employment*, and *publications.* We assign the edge types automatically based on the local context that surrounds the entity mention (five tokens before and after the mention) and its content string. The results of 10-fold cross validation of a Naive Bayes classifier trained using a set of labeled examples and bag-of-word features measured 0.87 and 0.82 in precision and recall, respectively.

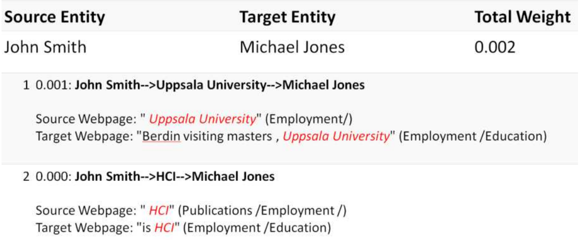
*Connecting people.* Personal profiles were constructed in this fashion for the participants of the Intelligent User Interfaces conference in 2015. Overall, 594 personal profiles were generated for participants for whom a homepage was identified. The individual

*UMAP'17, July 09-12, 2017, Bratislava, Slovakia*

ACM ISBN 978-1-4503-4635-1/17/07.
https://doi.org/http://dx.doi.org/10.1145/3079628.3079656

[1] http://nlp.stanford.edu/software/CRF-NER.shtml

[2] We leave disambiguation and unification to future work.



| Source Entity | Target Entity | Total Weight |
|---|---|---|
| John Smith | Michael Jones | 0.002 |

1 0.001: **John Smith-->Uppsala University-->Michael Jones**

Source Webpage: " *Uppsala University*" (Employment/)
Target Webpage: "Berdin visiting masters , *Uppsala University*" (Employment /Education)

2 0.000: **John Smith-->HCI-->Michael Jones**

Source Webpage: " *HCI*" (Publications /Employment /)
Target Webpage: "is *HCI*" (Employment /Education)

**Figure 1: Presentation of the supporting paths for a computed inter-person relatedness score.**

graphs were unified into a compact yet sparse joint social graph consisting of 70K edges and 23K nodes.

In this study, we wish to highlight to each participant a list of related conference attendants. It is likely that some of the predicted connections correspond to existing acquaintances, yet it is desired to bring to one's attention potential new acquaintances, and the respective social contexts. This task corresponds to the query: "who are the person nodes most related to $p$, and *why*?". We believe that the graph-based suggestions must be explained, so as to engage users and obtain their trust. Formally, this task involves ranking *person* nodes in the graph by their graph-based similarity to the node representing the focus person $t$. We apply the Personalized PageRank measure, conducting a two-step random walk process, to address this query [4]. The relatedness score of node $p$ with respect to $t$ equals the summation of the weights of the individual paths that connect them. Essentially, nodes that connect over a larger number of paths, as well as shorter paths, are assigned higher relatedness scores.

Recommendation systems typically only provide with numeric scores. In contrast, we provide the user with explanations for the suggestions made, namely the set of paths over which the entities connect in the graph, having a *path* denote a sequence of labeled *entities* and *relations*. Figure 1 displays the connecting paths between two (weakly linked) IUI participants whom we name here 'John Smith', and 'Michael Jones'. The figure shows the computed similarity score, along with two paths that account for their interpersonal relatedness. The first path connects the two persons via the entity 'Uppsala University' over a relation labeled as *employment* (for 'John Smith'), and *education* as well as *employment* for 'Michael Jones'. In addition to the predicted relation types, available lexical context is presented ('..visiting masters..'). This path seems interesting and non-trivial, as only a minority of researchers attended or visited Uppsala University. This fact may therefore ignite interest and motivate a conversation. The second path connects the two persons over the concept 'HCI'. This path is less interesting, as most of the conference participants are involved in HCI research. Indeed, the weight of this path is low (nearly zero); since a large number of *person* nodes link to the node denoting 'HCI', its contribution to the similarity score is low [4].

## 3 EVALUATION: SOCIAL RECOMMENDATION

We requested two dozens of IUI'15 participants to experience with the system over the Web, and provide us with feedback in free form. Following is a summary of their feedbacks.

*Ranking quality.* Some of the feedbacks pertained to the perceived correctness of the rankings. While one of the respondents commented that "the selected persons are adequate and the weight of the link as well", others noted cases in which persons who they knew well and collaborated with were ranked below people with whom they were less familiar with. Detailed feedbacks also pointed out some disambiguation issues, e.g., multiple mentions of the name "Huang" refer to different people, as well as errors in our third-part named entity tagger. These errors can be alleviated by improving named entity recognition. Learning from user feedback may also help promote informative paths. Nevertheless, we find the feedbacks to be encouraging; for example, one user defined 70% of the rankings as relevant and interesting. Although our main focus in not on optimizing the rankings by familiarity, we find that GBEP should be tuned and measured with respect to this requirement in the future to meet users expectations.

*Surprise.* Our goal is to rather point out new or unknown interesting ties, so as to encourage the user to make new contacts, or to 'break the ice' when being introduced to or meeting yet-unknown persons in a social setup such as a scientific conference. Connecting entities may suggest topics for conversation and encourage the exploration of mutual background. Several feedbacks indeed used the word 'surprised', e.g., "I was surprised to see A. at the top but when I checked out their research profile it makes some sense."

*Clarity.* Supposedly, detailing the relational paths that connect the user to a related person is more intuitive and convincing compared with mere numerical scores. Indeed, the comments cited so far indicate that the users took advantage of the system as intended, exploring the connecting paths that associate them to the suggested persons. The feedbacks indicate that this presentations increases users' engagement, e.g., consider the following positive comment collected in our survey: "I certainly found it an interesting activity to go through the list for 10 minutes and check out the home pages of some of these people for which I was unfamiliar (or in some cases had forgotten)".

In summary, we described a prototype for generating personal profiles that uses information extraction techniques to automatically process Web data into structured entities and relations information. We performed social recommendation using a graph that included the profiles of a scientific conference attendants. Initial feedbacks indicate that the suggested rankings, as well as the graph-based relational explanations generated, are sensible, surprising at times, and engaging. In the future, we would like to improve and personalize the random walk scheme using path-based learning techniques based on users' feedback[3, 4], and explore additional applications of this framework.

## REFERENCES

[1] Lada A Adamic and Eytan Adar. 2003. Friends and neighbors on the Web. *Social Networks* 25, 3 (2003).

[2] Peter Brusilovsky, , Jung Sun Oh, Claudia López, Denis Parra, and Wei Jeng. 2017. Linking information and people in a social system for academic conferences. *New Review Of Hypermedia And Multimedia* (2017).

[3] Ni Lao, Einat Minkov, and William W. Cohen. 2016. Learning Relational Features with Backward Random Walks. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*.

[4] Einat Minkov and William W Cohen. 2010. Improving graph-walk-based similarity with reranking: Case studies for personal information management. *ACM Transactions on Information Systems (TOIS)* 29, 1 (2010), 4.