Ranking Entities on the Web using Social Network Mining and Ranking Learning

Yingzi Jin
University of Tokyo
eiko-kin@mi.ci.i.u-tokyo.ac.jp

Yutaka Matsuo
University of Tokyo
matsuo@biz-model.t.u-tokyo.ac.jp

Mitsuru Ishizuka
University of Tokyo
ishizuka@i.u-tokyo.ac.jp

ABSTRACT

Social networks have garnered much attention recently. Several studies have been undertaken to extract social networks among people, companies, and so on automatically from the web. For use in social sciences, social networks enable analyses of the performance and valuation of companies. This paper describes an attempt to learn ranking of entities from a social network that has been mined from the web. In our approach, we first extract different kinds of relational data from the web. We construct social networks using several relevance measures in addition to text analysis. Subsequently, the relations are integrated to maximize the ranking predictability. We also integrate several relations into a *combined-relational network* and use the latest ranking learning algorithm to obtain the ranking model. Additionally, we propose the use of centrality scores of companies on the network as features for ranking. We conducted two experiments on a social network among companies to learn the ranking of market capitalization, and on a social network among researchers for ranking of researchers' productivity. This study specifically examines a new approach to using web information for advanced analysis by integrating multiple relations among named entities.

Keywords

social network, ranking learning, relation extraction, search engine, web mining

1. INTRODUCTION

Social networks have attracted much attention recently. An increasing number of studies have investigated relation extraction and network extraction among named entities on the web. Several studies have been undertaken to extract social networks automatically from the web among people, companies, and so on [6, 8, 7, 5].

The extracted relations and social networks are useful for various applications [2, 6]. In the social sciences, in order to identify the prominence or importance of an individual actor embedded in a network, *centrality* measures have been used in social sciences. On the other hand, ranking network entities is an important topic in link mining. Given a network among entities, the goal is to find a good ranking function to calculate the ranking of each entity using the relational structure [1, 3].

WWW2008, April 21–25, 2008, Beijing, China.

Considering those two directions of recent studies–relation extraction from the web and ranking learning– the next feasible step is to learn ranking based on relations extracted from the web. This paper describes an attempt to learn the ranking of named entities from a social network mined from the web. It enables us to have a model to rank entities for various purposes: one might wish to rank entities for search and recommendation, or might want to have the ranking model for prediction. For example, if we want to rank companys by their market price, we can extract the social network of the company from the web and learn the ranking model based on the social network. Consequently, we can predict the market price ranking of a new company by considering its relations to other companys.

In our ranking learning model, given a list of entities, we first extract different types of relations from the web based on our previous work [7, 5]. Then we rank entities on these networks using different ranking indices. We designate these rankings as *internal rankings* because they are calculated directly from relational networks. Conversely, we designate the target ranking of given entities as the *external ranking*. We propose three approaches to learn and predict target ranking based on internal rankings: Simply choose the most predictive relation types; Combine multiple relations into one network, designated as the *combined-relational network*, to learn ranking using a probabilistic model; Integrate multiple ranking indices from social networks as company features. We conducted two experiments: using social networks of 312 companies in Japan to discern the ranking of market capitalization; using social netowrks of 253 researchers from University of Tokyo to learn the ranking of researchers' productivity. Several findings including web co-occurrence relations are important to produce good rankings for companies as well as researchers. Large companies are also famous in several relational networks.

This paper is organized as follows. The following section presents a description of an overview of a ranking learning model. Section 3 introduces our previous work for extracting social networks from the web. Section 4 describes ranking learning approaches based on extracted social networks. Section 5 describes experimental settings and the results. Section 6 presents some discussion before concluding the paper.

2. RANKING LEARNING MODEL

The motivation of our study is explained as follows: We can infer various relations among entities from the web. However, what we are often interested in is not the rela-

tion itself, but a combination of relations (e.g. finding a path), or the aggregated impact of the relations to each entity (e.g. centrality of the entity) [9]. If we can identify a type of relation or a typed network that is influential to some attributes of each entity, we can understand that the types of relation are important, and that it would be possible to execute an analysis using the extracted network. In short, our approach consists of two steps;

Step 1: Constructing Social Networks Given a list of entities with target ranking, we extract multiple social networks among these entities from the web based on preliminary studies.

Step 2: Ranking learning Rank entities on extracted social networks and determine a ranking model based on the feedback from correlation between *external ranking* (target ranking) and *internal ranking* (from the network itself).

Once we obtain a ranking model, we use it for prediction for unknown entities. Additionally, we can obtain the weights for each relation type, which can be considered as important for relations. If the important relations are identified, the social network can be visualized by specifically examining its relations. Alternatively, social network analysis can be executed based on the relations.

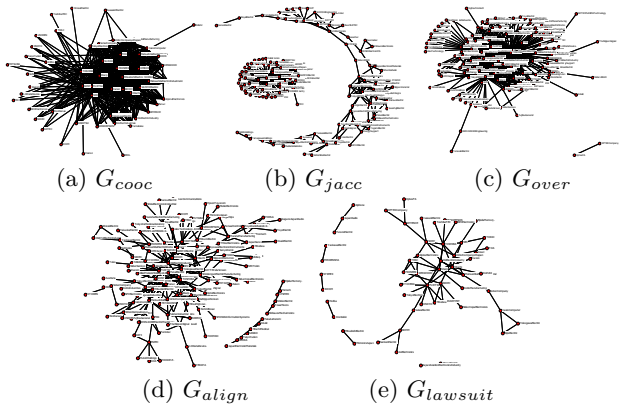
3. CONSTRUCTING SOCIAL NETWORKS FROM THE WEB

In this step, our task is, given a set of companies, we use a general search engine to construct a set of social networks $G_i(V, E_i)$, $i = 1, \ldots, m$, where m signifies the number of relations, V is the set of entities, and E_i is the set of edges with respect to the *i-th* relation. We are only interested in undirected networks.

The first kind of social network is extracted using a co-occurrence-based approach [6, 8, 7]. Given a person name list, the strength of relevance of two persons, x and y, is estimated by putting a query x AND y to a search engine. An edge will be invented when the relation strength by the co-occurrence measure is higher than a predefined threshold. Subsequently, we extract three kinds of co-occurrence-based networks: cooc network(G_{cooc}), jaccard network(G_{jacc}) and overlap network(G_{over}). The relational indices are calculated respectively using the Matching coefficient $n_{x\wedge y}$, the Jaccard coefficient $n_{x\wedge y}/n_{x\vee y}$ (also used by [8, 6]), and the overlap coefficient $n_{x\wedge y}/\min(n_x, n_y)$ (used by [7]).

Based on Web co-occurrence networks, Y. Matsuo et al. targeted the relations in a researcher community to classify relations using C4.5 as a classifier. In our experiments, we first extract jaccard network, then classify the edges into two kinds of relational networks: an co-affiliation network (G_{affi}) and a co-project network (G_{proj}) which similar to Lab and proj relational networks respectively.

Jin et al. proposed the *relation-identification* approach to extract target relational social networks [5]. Given a list of companies and target relations as input, the method extracts a social network of entities. To collect target relational information from the tops of web pages, it makes elaborate queries to emphasize a specific relationship, and applies text processing to those pages to form an inference of whether or not the relation actually exists. Subsequently, we extract two kinds of relational networks: an alliance network (N_{align}) and a lawsuit network ($N_{lawsuit}$).



(a) G_{cooc} (b) G_{jacc} (c) G_{over}

(d) G_{align} (e) $G_{lawsuit}$

Figure 1: Web-based social networks for companies with different relational indices or types.

Extracted networks for 312 companies are portrayed in Fig. 1.We can see that the social networks vary with different relational indices or types even thought they contain the same list of entities.

4. RANKING LEARNING

Given constructed multiple relational networks for a list of eitities, we can rank entities based on those given networks. Because these rankings are caused directly by relational network itself, we designate these as *internal ranking*. For i-th relational network, the internal ranking is indicated as $R^{(G_i)}$, and consequently call the target ranking *external ranking*, indicated as $\hat{R}$. Our task is to find the model for internal ranking which correlates most with the external ranking. We propose to use the following three methods to learn ranking based on networks.

4.1 Approach 1: Choosing the most predictive type of relation

With this method, we calculate some indices (such as centrality measures) based on the network for each type of relation. Although simple, it can be considered as an implicit step of social network analysis given multiple relations. We merely choose the type of relation that maximally explains the given ranking. We rank each type of relational network, then compare the *internal ranking* with the *external ranking*. Intuitively, if the correlation to the internal ranking $R^{(G_i)}$ is high, then the relation i as optimal parameter θ, which represents the important influences among entities for the given target:

$$\theta = \underset{i \in m}{\operatorname{argmax}}\, Cor(R^{(G_i)}, \hat{R}). \quad (1)$$

Several means can be used to rank network entities with different meanings of prominence and importance. In social network analysis, *degree centrality* (R_D), *Betweenness centrality* (R_B), and *Clossness centrality* (R_C) are often used to identify the prominence or importance of an actor. Other ranking methods such as *PageRank* (R_P) is defined as its steady-state visit probability on the Markovian network. These measures characterize some aspects of the local or global network structure, as seen from a given actor's embeddedness in the network. Considering several meanings of internal ranking, our method can be extended simply to find the parameter $\{i, j\} \in \theta$ ($i \in m$, $j \in n$), i-th network with j-th ranking indices, which maximize the coefficient between

internal ranking $R_j^{(G_i)}$ with target ranking $\hat{R}$.

$$\theta = \underset{i \in m, j \in n}{\text{argmax}}\, Cor(R_j^{(G_i)}, \hat{R}) \quad (2)$$

4.2 Approach 2: Learning ranking using a probabilistic model

Many existing algorithms related to ranking network entities specifically examine graphs with a single link type. However, multiple social networks exist in the real world, each representing a particular relationship type, each of which might be integrated to play a distinct role in a particular task. We combine several extracted multiple social networks into one network and designate this kind of social network as a *combined-relational network*. Our target is using combined-relational social networks, which are integrated by multiple relational networks extracted from the web, to learn and predict the ranking.

NetRank [1] is proposed to learn certain edge parameters of Markovian walks on network. The ideal random walk is likely to transit along edges of different types with different probabilities, and in which transition parameters are learned form given preference orders over pairwise nodes. Our model is based on NetRank idea, set each relation type i has a strictly positive weight $\beta(i) > 0$. In this case, transition matrix A is desingend as

$$A(y,x) = \begin{cases} \alpha \frac{[\beta(i(x,y)) \in E]}{OutWeight(x)} + (1-\alpha) r_y, & e \in V_0 \\ r_y, & \text{otherwise} \end{cases} \quad (3)$$

where $\text{OutWeight}(x) = \sum_y \beta(i(x,y))$. A is a function of the weights β, and we are looking for $\{\beta_i\}$ such that the p solves $p = Ap$ also satisfies given target ranking. Simply, we randomly sampling the weights for each relation type to combine a multi-relational network; we also do so based on the feedback of relevance of internal ranking with target ranking to tune the parameters.

4.3 Approach 3: Integrating multiple indices from social networks

The most advanced method in our research is to integrate multiple indices that are obtained from multiplex social networks. The goal of learning is to integrate all rankings from networks into a single ranking of the instances. They are expected to be useful to interpret a given target ranking most accurately.

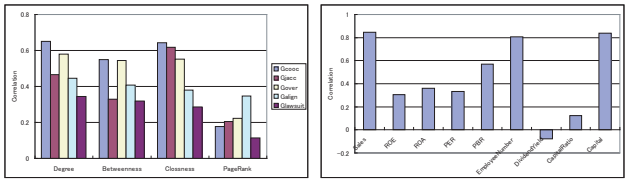
Intuitively, we integrate multiple indices from social networks, thereby combining several perspectives of importance for individuals from different relational structures. This integration is accomplished by regression of ranking based on various features. In this case, the purpose is to identify optimal parameters θ (i.e. weights for ranking features) in each network:

$$\theta = \underset{w_{i,j}}{\text{argmax}}\, Cor(w_{i,j} \cdot R_j^{(G_i)}, \hat{R}). \quad (4)$$

Therein, the score of $R_j^{(G_i)}$ represents the ranking score on i-th network with j-th ranking algorithm. Therefore, there have $m * n$ feature space to represent an actor. We try to identify the optimal combination weights $w_{i,j}$, $i \in m, j \in n$) and use SVM-regression to select optimal parameters.

5. EXPERIMENTAL RESULTS

Figure 2: Correlation between each rankings with external ranking of companies.



(a) Correlation with internal rankings

(b) Correlation with fundamental rankings

Table 1: Correlation between the external ranking with integrated internal ranking with different relation weights.

β_{cooc}	β_{jacc}	β_{over}	β_{align}	$\beta_{lawsuit}$	Cor
0.72	0.03	0.14	0.1	0.00	0.408

In this section, we describe results to clarify the effectiveness of ranking learning on extracted social networks. For the first dataset, we use 312 electrical-product-related companies listed on the Tokyo Stock Exchange[1] to predict ranking of companies. We extract social networks of five kinds from the web using a search engine (Fig.1): MSN[2]: Cooc network (G_{cooc}), jaccard network (G_{jacc}) and overlap network (G_{over}) networks are extracted using the co-occurrence-based approach; alliance network (G_{align}) and lawsuit network ($G_{lawsuit}$) are extracted using the relation-identification approach described in Section 3. For second dataset, we use 253 researchers from the University of Tokyo to predict ranking of researchers. We first extract three co-occurrence-based social networks (G_{cooc}, G_{jacc}, G_{over}), then label on the G_{jacc} network in order to extract co-affiliation network (G_{affi}) and a co-project network (G_{proj}) among researchers.

In our experiments, we conducted 3-fold cross-validation. In each trial, two folds of actors are used for training, one fold for prediction. The results we report in this section are those average over three trials.

5.1 Predict Ranking of Market-Cap among Companies

The target ranking of the companies is based on Market-Capmarket capitalization (Market-Cap)[3]. Market-Cap represents the market's valuation of all the equity in a company.

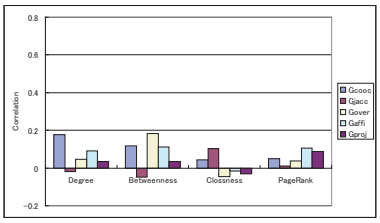
For the first approach, we simply choose the most relevant type of relation with Market-Cap. We rank each social network using various ranking methods introduced into Section 4.1. Fig. 2(a) shows all of ranking correlations between each internal ranking and external ranking. Rankings with degree centrality ($R_D^{(G_i)}$) and closeness centrality $R_C^{(G_i)}$ show the good correlation with external ranking, which means that big companies also readily co-occur with other companies on the web, and also they are nearly connected with other companies on the web. The G_{cooc} network produces good rankings. It is explainable that one company frequently co-occurring with other companies on the web would make the company well known. Consequently, the target ranking of company will be improved. As a comparison with relational

[1] http://profile.yahoo.co.jp/industry/electrical/electrical1.html
[2] http://jp.msn.com/
[3] Actually, we used log transformations of these market values as the quantities used for this study.

Table 2: Results of SVM-regression, respectively using relational indices, fundamental indices, and both indices as features.

$Cor(W \cdot R_j^{(G_i)}, \hat{R})$	$Cor(R^{(F)}, \hat{R})$	$Cor(R^{(BOTH)}, \hat{R})$
0.512	0.612	0.644

Figure 3: Correlation between each internal ranking in researcher networks with external rank.



indices produced from network, we use fundamental indices [4] which have been used traditionally for company valuation to measure the relevance with Market-Cap. Fig. 2(b) shows all of ranking correlation between rankings on fundamental indices and external ranking. Ranking companies with these carefully chosen fundamental indices have good relevance with extrenal ranking, because they are affected directly by the company profile itself, and it meets our intuition.

For the second approach, we randomly created transaction weights for each relation type (total of transition weight as 0.9) on a Markov network and apply NetRank to select optimal parameters. Table 1 shows the best parameter sets that generate 0.358 relevance between internal ranking and target ranking. The correlation is not so high. The possible reason is that the relations be aggregated into one network would lose some information of individual relations. However, we can see the G_{cooc} and G_{aling} create positive impact on the combined-relational network. impact.

Finally, we integrate multiple indices obtained from multiple social networks with several ranking indices as features of a company. We use the linear kernel SVM to combine these features. Table 2 shows the correlation between target ranking with estimated rankings using fundamental indices only, relational indices only, and both indices as features of a company. Fundamental indices yield a good estimation. However, combination of the fundamental indices with relational indices performs better than fundamental indices only. We can see that the relations and structural embeddedness of a company affect the company performance itself.

Results demonstrate that relations with powerful companies is more efficient than forming relations with small companies. For target ranking of companies with Mark-Cap, the rankings on cooc network is most relevant. Also the degree centrality and closeness centrality have good relevance with target ranking. It means that, big companies are readily co-occur with other companies on the web, and nearly connected with other companies.

5.2 Predict Ranking of Paper Productivity among Researchers

Academic papers are often the product of several researchers' collaboration. Therefore, a good position in a social network is derived through good performance. Is there any relation that is important to predict productivity?

Fig. 3 shows the ranking correlation between each kind of internal ranking and external ranking. Correlations between external rankings with each internal ranking are apparently not important. We then integrate multi-relational networks using transaction weights on different relational types. The best set of parameter $< \beta_{cooc}, \beta_{jacc}, \beta_{over}, \beta_{affi}, \beta_{proj} >$is $< 0.74, 0.00, 0.01, 0.14, 0.01 >$. Finally, we used a linear kernel SVM regression to train and learn the ranking. The correlation between predicted ranking from internal rankings and external ranking is 0.326, which is improved even though it is still not highly relevant.

We can see that the correlations between researcher's ranking with network-produced internal rankings are apparently not relevant. Perhaps the professor's publications (number of papers) are not highly related with our social networks. A further examination on finding which rankings are sensitive with relations will be done in the future.

6. CONCLUSION

This paper describes methods of learning the ranking of named entities from a social network mined from the web. Various relations pertain to our lives: their combinations and their aggregate impacts are influential to predict features of entities. These tasks include ranking or scores for target entities, i.e. external ranking. Based on that intuition, we constructed our ranking learning model from social networks to predict the ranking of other entities. We first extracted different kinds of social networks from the web. Then, we used these networks and a given target ranking to learn important relations ranking indices. We proposed three approaches to obtain the ranking model. Our approach suggests an interesting and important direction for advanced web mining.

7. REFERENCES

[1] A. Agarwal, S. Chakrabarti, and S. Aggarwal. Learning to rank networked entities. In *Proc. KDD'06*, 2006.

[2] B. Aleman-Meza, M. Nagarajan, C. Ramakrishnan, A. Sheth, I. Arpinar, L. Ding, P. Kolari, A. Joshi, and Tim Finin. Semantic analytics on social networks: Experiences in addressing the problem of conflict of interest detection. In *Proc. WWW2006*, 2006.

[3] H. Chang, D. Cohn, and A. McCallum. Creating customized authority lists. In *Proc. ICML2000*, 2000.

[4] Y. Jin, Y. Matsuo, and M. Ishizuka. Extracting social networks among various entities on the web. In *ESWC2007*, 2007.

[5] H. Kautz, B. Selman, and M. Shah. The hidden Web. *AI magazine*, Vol. 18, No. 2, pp. 27–35, 1997.

[6] Y. Matsuo, J. Mori, M. Hamasaki, K. Ishida, T. Nishimura, H. Takeda, K. Hasida, and M. Ishizuka. POLYPHONET: an advanced social network extraction system. In *WWW2006*, 2006.

[7] P. Mika. Flink: semantic web technology for the extraction and analysis of social networks. *Journal of Web Semantics*, Vol. 3, No. 2, pp. 211–223, 2005.

[8] Brian Uzzi. Social structure and competition in interfirm networks: The paradox of embeddedness. *Administrative Science Quarterly*, Vol. 42, pp. 35–67, 1997.

[4] ROE (return on equity), ROA (return on assets), PER (price earnings ratio), PBR (price to book value ratio), Sales, Assets, Asset Ratio, Dividend Yield, and Employee number