4Q1-2in2

Deploying Exploration in Proximity Indices for Link Collection Problem

Guoxi Zhang[*1] Hisashi Kashima[*2]

[*1] [*2]Graduate School of Informatics, Kyoto University

Link analysis has attracted attention from wide range of disciplines as it reveals latent knowledge from relational data. In applications like drug discovery or reactome analysis, identifying unobserved relations from data may reduce cost of wet experiments. In this paper, we study the problem of link collection. Given fixed set of nodes and a set of links, an algorithm for this problem iteratively outputs pair of nodes as queries. Queries are labeled by an annotator and links among them are utilized in further query calculation. We examine the effect of deploying several exploration strategies into proximity indices. In our experiments modifications improve performance of proximity indices on several datasets. Our work provides insights about how exploration can be deployed into proximity indices for graph mining tasks.

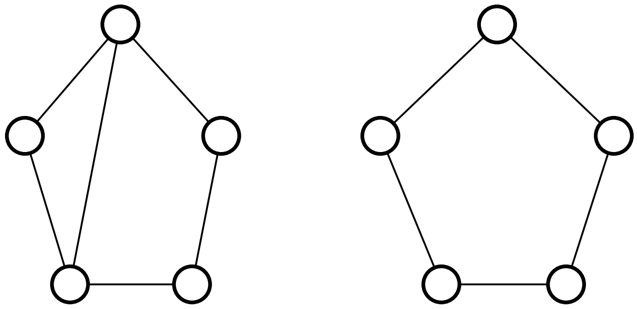

Figure 1: Example for situation where greedy policy can be suboptimal.

1. Introduction

Relational data analysis reveals latent knowledge from structure of data. For example, in bioinformatics by representing the interactome as undirected acyclic graphs allows prediction of new interactions. Many proximity indices have been proposed in literature. For instance, the common neighbor coefficient is defined to be the number of common entities in neighborhood of two nodes. In the case of social network, this index can be interpreted as common friends, and two individuals are likely to know about each other if they have a lot of common friends. In literature, proximity indices have been thoroughly studied and widely applied to various applications, because of their effectiveness.

The construction of relational datasets is also an important task in relational data analysis. Expertise and experiment expense are often necessary in constructing domain specific datasets. In this paper, the problem of link collection is considered. Given the set of entities V, a subset of links E_{ini}, and an oracle O, the agent iteratively picks a query from a set of node pairs P whose connectivities are unobserved. If a query is labeled as connected by O, it will be added into the set of links. The object is to maximize the number of links collected within T iterations. This problem setting is of benefit to applications in which the budget of data construction is limited.

An obstacle in making use of proximity indices to collect link is the exploration-exploitation dilemma. Proximity indices identify apparently best queries in an iteration, but naively selecting the currently best query may result in a suboptimal overall outcome. For example, suppose a network is initialized as shown in Figure 1. Naively applying common neighbor index results in deterministically querying for all unconnected pairs in the left pentagon first, in regardless of true distribution of links.

In our approach, the deployment of exploration is carried out from two aspects. On the one hand, two scores are proposed for selecting queries, which measure the value of neighboring pairs. The agent is encouraged to explore information in the neighborhood of a pair of nodes. On the other hand, proposed scores are combined with proximity indices stochastically. Experiments are carried out on four network datasets with three proximity indices. Our results show that stochastic combination can outperform using proximity indices only.

2. Related Work

Active learning paradigm has been applied in scenario to reduce cost of obtaining data in literature of link prediction problem. For example, in [MCG10] and [KNM14] the authors applies sampling heuristics to protein-protein prediction task and to protein-compound response prediction task, respectively. In active learning an agent needs knowledge about unconnected pairs, so it tends to query for it if necessary. On the contrary, in link collection problem an agent should avoid querying for unconnected pairs to maximize number of links obtained in T iterations.

In [KKB$^+$15], the authors address the problem of link collection in multi-relational setting, and learn a predictive score to select queries by solving an optimization problem. In [KXO16] the authors propose a probabilistic factorization approach for selecting queries that could benefit from path structure in a graph. To our knowledge, though proximity indices become a important line of approach in literature of link prediction [LNK07], they have not been applied to link collection problem yet.

Contact: Guoxi Zhang, guoxi@ml.ist.i.kyoto-u.ac.jp

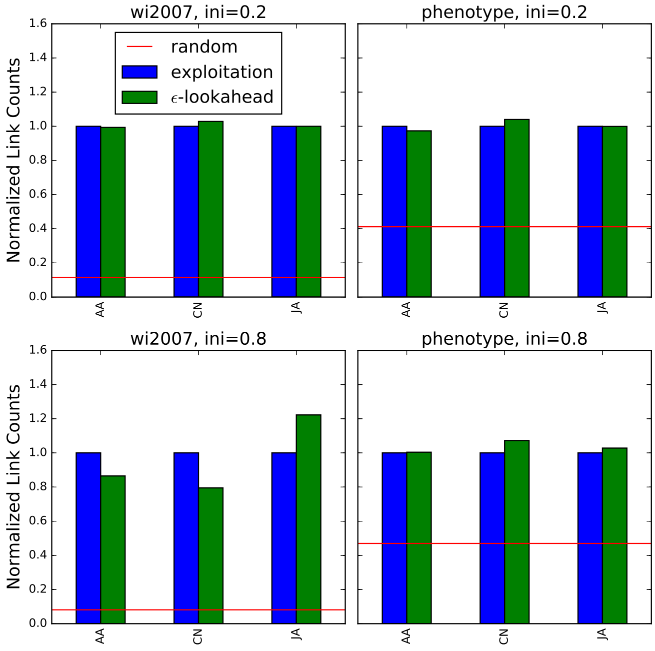

Figure 2: Results of link collection.

3. Approach

3.1 Proximity Indices

Various proximity indices are proposed in literature. In this paper three of them are considered: Common Neighbor (CN), Adamic Adar (AA) and Jaccard(JA). Denote nodes in a node pair x as x_0 and x_1. Directions are not considered. Denote set of nodes that are directly connected to some node as $\Gamma(\cdot)$. The definitions of proximity are:

$$\begin{aligned} \mathrm{CN}_E(x) &= \left|\Gamma(x_0) \cap \Gamma(x_1)\right|, \\ \mathrm{JA}_E(x) &= \frac{|\Gamma(x_0) \cap \Gamma(x_1)|}{|\Gamma(x_0) \cup \Gamma(x_1)|}, \\ \mathrm{AA}_E(x) &= \sum_{w \in \Gamma(x_0) \cap \Gamma(x_1)} \frac{1}{\log(|\Gamma(w)|)}, \end{aligned} \tag{1}$$

where subscript E refers to the current set of links.

3.2 Scores for Neighborhood Exploration

Proximity indices are calculated from information of set of nodes that are one or two hops away from x_0 and x_1. Denote set of all possible unconnected pairs formed by nodes in such set as $\Delta(x)$. In real network x and elements in $\Delta(x)$ are often correlated. For example, in the case of co-authorship network, collaboration of two scientist can induce further collaboration between members of their groups. Thus for some proximity index f, $\mathrm{f}_{E\cup\{y\}}(x)$ can also be used as proximity index, where $y \in \Delta(x)$. Subscript $E \cup \{y\}$ refers to temporarily add y to E, the current set of link, and remove it after calculation. Compared to f, $\mathrm{f}_{E\cup\{y\}}$ utilizes correlation between x and $\Delta(x)$. By enumerating y, information in neighborhood can be explored. Based on this idea, we propose the following score:

Table 1: Statistics of Datasets

dataset	# of nodes	# of links	type
wi2007	1496	1714	protein-protein
phenotype	912	22738	protein-protein

$$\mathrm{lookahead}_E(x) = \frac{\sum_{y\in\Delta(x)} \mathrm{f}(y)\mathrm{f}_{E\cup\{y\}}(x)}{\sum_{y\in\Delta(x)} \mathrm{f}(y)}. \tag{2}$$

3.3 Exploration and Exploitation

In each iteration, an agent selects best node pairs either according to proximity indices or according to score defined in Eq. 2. The former action is referred to as exploitation. In the latter case, for efficiency reason, an agent only select query from pairs with top-ten proximity values in P. The idea of ϵ-policy is adopted. An agent selects queries according exploration score with probability ϵ, and selects query according to proximity indices with probability $1 - \epsilon$.

4. Experiments

4.1 Data

We use "wi2007 " and "phenotype", two protein-protein interaction datasets available on webset of Center for Cancer System Biology [SRC+09]. All directions and loops are removed.

4.2 Setting

Radom selection is used as baseline, in which queries are selected uniformly at random from P. The number of links collected by exploitation, random selection and ϵ-info-gain are divided by the number of links collected by exploitation of the corresponding proximity index. This gives a normalized performance measure for each policy.

Experiments are carried out with E_{ini} containing 20% and 80% of total links. By comparing results from them, we can examine impact of structural information on link collection. In each iteration, 100 node pairs are given to an agent, which is randomly sampled from $V \times V \setminus E$. V is the set of nodes. Collection procedures are executed for 10,000 iterations. ϵ is set to 0.1.

4.3 Results

Figure 2 shows result of link collection on four datasets. Both exploitation policy and ϵ-lookahead policy are more than five times better than random selection on "wi2007" dataset. Influence on exploitation and ϵ-lookahead of amount of structural information is observable. This is because more unconnected pairs will be assigned high value of proximity if more links are available. In the case of "wi2007", performance of the two policies being comparing diverge when 80% links are used in initialization, though they are almost the same when 20% links are used in initialization. In the case of "phenotype", performance of ϵ-lookahead is improved when more links are used in initialization.

5. Conclusion

In this paper we present experiment results of deploying exploration in link collection with proximity index on four datasets. These preliminary results implies that pure exploitation, or greedy policy, can be sub-optimal one if the object is to maximize total links collected.

References

[KKB+15] Hiroshi Kajino, Akihiro Kishimoto, Adi Botea, Elizabeth Daly, and Spyros Kotoulas. Active learning for multi-relational data construction. In *Proceedings of the 24th International Conference on World Wide Web*, pages 560–569. ACM, 2015.

[KNM14] Joshua D Kangas, Armaghan W Naik, and Robert F Murphy. Efficient discovery of responses of proteins to compounds using active learning. *BMC bioinformatics*, 15(1):1, 2014.

[Kun13] Jérôme Kunegis. Konect: the koblenz network collection. In *Proceedings of the 22nd International Conference on World Wide Web*, pages 1343–1350. ACM, 2013.

[KXO16] Dongwoo Kim, Lexing Xie, and Cheng Soon Ong. Probabilistic knowledge graph construction: Compositional and incremental approaches. In *Proceedings of the 25th ACM International on Conference on Information and Knowledge Management*, pages 2257–2262. ACM, 2016.

[LNK07] David Liben-Nowell and Jon Kleinberg. The link-prediction problem for social networks. *journal of the Association for Information Science and Technology*, 58(7):1019–1031, 2007.

[MCG10] Thahir P Mohamed, Jaime G Carbonell, and Madhavi K Ganapathiraju. Active learning for human protein-protein interaction prediction. *BMC bioinformatics*, 11(1):1, 2010.

[SRC+09] Nicolas Simonis, Jean-François Rual, Anne-Ruxandra Carvunis, Murat Tasan, Irma Lemmens, Tomoko Hirozane-Kishikawa, Tong Hao, Julie M Sahalie, Kavitha Venkatesan, Fana Gebreab, et al. Empirically controlled mapping of the caenorhabditis elegans protein-protein interactome network. *Nature methods*, 6(1):47–54, 2009.