T-RecS: Team Recommendation System through Expertise and Cohesiveness

Anwitaman Datta, Jackson Tan Teck Yong, Anthony Ventresque*
School of Computer Engineering
Nanyang Technological University, Singapore
{anwitaman, jacktty, aventresque}@ntu.edu.sg

ABSTRACT

Searching for people by exploration of social networks structure is an interesting problem which has recently gathered a lot of attention. Expert recommendation is an important but also extensively researched problem. In contrast, the generalized problem of team recommendation has not been studied a lot. The purpose of this demo is to show a *multidisciplinary team search and recommendation* prototype. While the current demo uses specific (NTU academic) dataset, the framework is generic, and can be extended for other domains subject to availability of suitable information.

Categories and Subject Descriptors

H.3.3 [**Information Systems**]: Information Storage and Retrieval; H.3.4 [**Information Systems**]: Information Interfaces and Presentation—*User profiles and alert services*

General Terms

Theory, Algorithms, Design

Keywords

Expert Search, Team Recommendation, Social Network Analysis

1. INTRODUCTION

Recommendation and referral systems are used in many walks of life, boosting productivity or user experience. A specific class of recommendation which can be useful for head-hunting as well as to find the appropriate person(s) with required expertise in very large organizations or in virtual communities where knowing everyone well and personally is not realistic is *expert recommendation* [4, 3]. A generalization of the expert recommendation problem is the *team recommendation* problem, where a composite team needs to be formed to carry out a task requiring multiple skills.

A naïve scheme to form teams would be to identify experts for each of the necessary skills and put together a group which between them cover all the required skills. Obvious refinements of such a strategy would be to define some ways to rank teams - possibly based on the degree of competence of the team members [6]. However, another important factor determining the quality of a team is how well a team can function together, which in turn depends on, among other things, the degree of cohesiveness among the team members [1]. To this end, we could leverage on the ample digital footprints of social networks [2, 3, 7] which are also deployed at workplace (for example, IBM's Lotus Connections[1]). Social network analytics sounds very promising to connect people in cohesive teams.

This paper considers the context where people have skills in certain topics, and are connected together through some social interactions. Then, given a multidisciplinary task, the objectives of our approach are to define (i) the degree of expertise of individuals for specific topics, and (ii) a composite mechanism to exploit the different elements of individuals and community given by the expertise and connections. Most work in the literature has been done on the expertise retrieval problem but to the best of our knowledge, nothing on the particular task of forming teams of individuals who could be selected together to perform a multi-skilled task.

While our approach is general enough to incorporate other domains, the current implementation of *T-RecS*[2] is an application leveraging on academic knowledge networks to recommend and help users explore multidisciplinary scientific teams. We use T-RecS both as a vehicle for validation and exposition of our ongoing works on the general theory and algorithms of team recommendation [1], as well as to implement and demonstrate a framework on how such individual components can be stitched together to build a working prototype. We have instantiated and deployed this prototype for local usage at NTU Singapore, using publicly available academics web sites and publications records. Although the case study is restricted to academic networks, and uses only public domain information, the T-RecS framework can be easily extended, possibly with addition of new elements to the modular architecture (Figure 1) for team recommendation in other settings, using other adequate information to drive the recommendation engine. Availability of such information is an orthogonal and complex issue, since it includes, privacy concerns, location, willingness, etc., nevertheless, applications like SmallBlue [3] indicate their emergence and acceptance in recent years. But once these specific modules are instantiated, we assert that T-RecS framework for team visualization and manipulation works for any other domain.

*Authors would like to thank Krzysztof Rzadca, for his early groundwork. The work was funded by A*STAR grant number 072 134 0055.

WWW 2011, March 28–April 1, 2011, Hyderabad, India.
ACM 978-1-4503-0637-9/11/03.

[1]http://www.ibm.com/lotus/connections
[2]http://sands.sce.ntu.edu.sg/T-RecS/

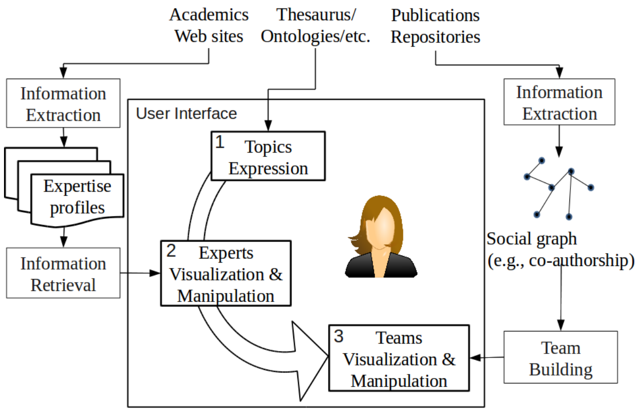

Figure 1: T-RecS system overview.

2. FINDING THE EXPERTS

2.1 Competence of a Researcher

Staffing (creating teams) starts with a measure of staff expertise. It is done either by topic extraction or by information retrieval from user topic expression. We use the latter in T-RecS, because it seems to fit better with the academic dynamics of topic trends.

2.1.1 *Expert Profiling*

Academic researchers' scientific profile is very often, both as self-declarations as well as publicly available publication and patent repositories. Thus, for the purpose of validation of our framework, we have chosen NTU research staff records. Specifically, we consider NTU publication record, and utilize NTU academics' profiles as displayed on their web site[3]. The publication record of NTU academics provides the title of papers, coauthorships and venue. We have weighted the importance of conference and journals appearing using Microsoft Academic Search journal and conference ranking[4], which gives the ratio of citations per paper in a venue. We consider this as the indicator of importance of conferences and journals, which in turn determines the degree of expertise of the academics. We note here that any other metric may be used instead.

2.1.2 *Query*

An important goal in our system is to disambiguate users' queries in order to find the best query formulation for their needs. For instance, a user may try words *strategy* and *auction* while she has in mind *game theory*, which is more accepted for the branch of mathematics. Instead of looking for keywords by using some well-known techniques (statistical techniques or domain specific ontologies), we have decided to use Wikipedia categories, which turns out to be accurate enough. We used *Lucene*[5] with default settings as the back-end data indexing system.

2.1.3 *Validation*

In order to find out the practicality of our expertise retrieval system, we asked domain "experts" from NTU to judge our system. They are called "judges" in this section. The judges were 7 academics from NTU (professors and post-docs) from 4 different schools. We proposed 15 topics ("Game Theory", "Water Quality", etc.) representing the various fields of expertise of the different schools of NTU; and for each topic we displayed 20 to 25 academics. We have put the top-ranked experts of our system together with some random people, and asked the judges to input the relevance of the individuals as experts on the considered topic (inter-experts correlation is 86%).

The web sites and papers information which we have from the record database are structured in fields, and Lucene provides a boost mechanism to increase the importance of some fields. We eventually find a 85% of precision and 65% of recall with best correlated boost values.

[3]http://research.ntu.edu.sg/expertise/
[4]http://academic.research.microsoft.com
[5]http://lucene.apache.org/

2.2 Competence Coverage in a Team

The aim of this section is twofold: (i) find possible teams, i.e. those covering all the needed skills, and (ii) rank teams according to a scoring function based on members' competence for the required skills, i.e. the team's *competence score* (Equation 1).

Let $R = \{r_1, \dots, r_m\}$ be the set of researchers and $S = \{s_1, \dots, s_n\}$ the set of possible skills. We define two functions $e: R \times S \rightarrow \mathbb{R}$, which gives the expertise value of a researcher, and $E: R \rightarrow \mathcal{P}(S)$, which links a researcher to its set of competences. It means that for a given threshold th: $e(r_i, s_j) > th \iff s_j \in E(r_i)$.

Let $T =< R, S >$ the set of teams. Let $t_m =< R_m, S_m >$ be a team; t_m (simply) covers its desired skills: i.e., there is at least an expert for every skill in this team: $c(t_m) = true \iff \forall s_i \in S_m \exists r_j \in R_m$ such that $s_i \in E(r_j)$.

If a set of people do not provide complete coverage, then they can not form a legitimate team by themselves. But there may also be different combinations of people who together do achieve a legitimate team instance. Nevertheless, such teams may (not) be satisfactory, and a way to score and rank the teams is needed. Many different scoring functions for a team are possible with respect to the coverage of the requisite set of skills. In T-RecS, we currently score a team based on any one of the max, min or average score across the different skills needed (i.e., $\Omega \in \{\max, \min, \text{avg}\}$), customizable according to individual user's preferences.

$$score(t_m) = \sum_{j=1}^{|S_m|} \Omega_{i=1}^{|R_m|} e(r_i, s_j) \quad (1)$$

Maximum is interesting if one is looking for at least one good expert of each skill in a team. Presence of specialists is more important than the number of people in the team. Similarly, minimum is useful to identify the so called "weakest link" in the team, avoiding people with very poor skills in some competences. Average reflects that teams with comparable numbers of experts and non-experts in the various topics could be more desirable.

Definition of properties like minimality (t_m is minimal $\iff c(t_m) = true$ and $\nexists\, t_n =< R_n, S_m >, c(t_n) = true$, and $|R_n| < |R_m|$), maximality (the opposite relation), or minimal covering ($\mathfrak{c}(t_m) = true \iff \forall s_i \in S_m \exists! r_j \in R_m$ such that $s_i \in E(r_j)$) can be easily given in our system. The properties, together with some values like the value $|R_m|$ for a team t_m, could help the user to customize to her needs the team recommendation process.

3. COHESION OF TEAMS

The cohesiveness of a team depends on how well the members bond with each other, as pairs as well as collections. Generally speaking, individuals may be related to each other in many different ways. For instance, in academic networks, people may be related to each other as: co-authors, members of the same department, participants in the same event or conference. Each link provides different kinds of acquaintance (undirected, weighted) graphs. People also may cite each others' works, which again induces a (directed, weighted) graph that can be inferred as some sort of "trust" graph. Any or many of these multiple social networks can be superimposed to build a multi-dimensional social network [1]. However, there exist a great interest in the literature for social network analysis of co-authorship networks [5]; because unlike co-authorship, other kinds of relations among researchers may not be as explicit, and also complex to comprehend or even to obtain. In current T-RecS implementation, we mainly focus on publicly available publication records (thus, only on a single-dimensional social network) for calculating cohesiveness (as described next).

Four different type of graphs can be used to represent the co-authorship network: simple graph, weighted graph, bipartite graph and hypergraph. The simple graph does not take into account the strength of collaborative ties. Now, it is obvious that this tie is not uniform among all the authors. For instance, academics that often publish together are far more linked than academics that did it only once. Weighted graph captures this variation in the intensity of the interaction. Bi-partite graph and hypergraph [8] are definitely very relevant to describe how teams evolve, because they are not ego-centric as graphs and weighted-graphs. But they are useless in the current context, since we do not address either dynamics of teams nor semantics of links between academics. We eventually choose the weighted graphs representation of co-authorship, because it can better give the intensity of co-authorship links. Even if richer weighting patterns can be found [5], we only use the number of publications in T-RecS, as it is the straightforward and meaningful way to describe the strength of the relation between academics (we don't consider here age for links strength).

In the current T-RecS deployment, we use a dataset composed of 1223 researchers (all members of NTU) who have published in total 11651 papers within the period of 2007-2009. We studied the ratio of papers per author, of authors per papers, and of the number of collaborators per author and find that there are power law distributions, with only the classical anomalies of such short periods analysis. The graph characteristics resemble what is found in literature [5].

But it is worthwhile noting that any of the previous assumptions, on either the kind of social network(s) considered, the graph used and the dataset can easily be changed in T-Recs. Indeed, our system is open and our implementation, based on interface definition, can be reused with other assumptions, structure or data. For instance, in current implementation we already allow users to visualize and navigate the above mentioned multi-dimensional social network to make their own decisions as well as custom build teams and compare them with T-RecS's default recommendations.

Let $G(R,P)$ be a graph, where R represents researchers and P represents co-authorship between any two researchers r_i and $r_j \in R$. An edge $p_{ij} \in P$ exists iff r_i and r_j have published together. We can even consider a link between every researchers, since $num(p_{ij})$ for an edge $p_{ij} \in P$ denotes the number of publications co-authored by r_i and r_j together.

A team t_m has a graph representation in the social graph: $G^{t_m} = (R_m, P^{t_m})$ which is a projection on $G(R,P)$. In general, members of a team may not have any direct relations between themselves. In fact, for a multidisciplinary team it is rather likely. We thus determine the connectivity between any two members by determining the strength of the shortest path between the same pair in the original graph $G(R,P)$. Thus, even if an edge did not exist between two people r_i and r_j in the original graph, in the projection $G^{t_m} = (R_m, P^{t_m})$ we may include such an edge, i.e., $p^{t_m}_{ij} \in P^{t_m}$. P^{t_m} is then defined as the virtual edges between any two researchers of R_m.

The strength of an edge p_{ij} in P^{t_m} is calculated according to Equation 2. The formula, albeit ad-hoc and thus with room for future refinement is based on the following rationale. A path between any two nodes $r_i, r_j \in R$ is a set of edges: $\{p_{12}, p_{23}, \cdots, p_{(n-1)n}\}$ such that $r_1 = r_i$, $r_n = r_j$. In P^{t_m}, p_{ij} is a virtual edge denoting the shortest path between r_i and r_j in R. We attach two values to p_{ij}. $\sigma_{p_{ij}}$ is the sum of weights on edges $e_{12}, e_{23}, \cdots, e_{(n-1)n}$ of P. $L_{p_{ij}}$ is the number of edges in this path (basically, n).

Thus $L_{p_{ij}} \times \max_{k,l\in R} \sigma_{p_{kl}}$ corresponds to the maximum possible strength of a path of the same length as p_{ij}. Dividing $\sigma_{p_{ij}}$ by this quantity gives a normalized strength of a path between r_i to r_j in G. But it does not take enough into account the impact of length of path. So we introduce a power $(1+\alpha)$, with $\alpha \in [0..1]$, which magnifies the impact of the path length in the formula (shorter paths mean stronger tie). The number of publications does not seem to be relevant beyond certain limits. Arguably, 20 papers together denotes a very good relation, and yet it looks insignificant w.r.t. a pair who have authored 50 papers together. We use the square root to mitigate this partly.

$$score^{t_m}(r_i, r_j) = \sqrt{\frac{\sigma_{p_{ij}}}{(L_{p_{ij}})^{1+\alpha} \times \max_{k,l\in R} \sigma_{p_{kl}}}} \quad (2)$$

We define a local clustering coefficient for any vertex $r_i \in R_m$, adapting the classical measure [9]. This measure essentially represents how well an individual belongs to the rest of the team.

$$C_i = \frac{2 \times \sum_{j,k\neq i} score^{t_m}(r_j, r_k)}{(|R_m|-1) \times (|R_m|-2)} \quad (3)$$

The global clustering coefficient for the team is then the average of local coefficients of all vertices in T, and represents the whole team's cohesiveness:

$$C_{t_m} = \frac{\sum_{i=1}^{|R_m|} C_i}{|R_m|} \quad (4)$$

Thus, for any composition of a team, we can determine its *social cohesiveness score* using Equation 4.

4. USER INTERFACE

A team recommendation interface has to be well designed, because it gives many options to the user, and it is necessary not to confuse her. A good interface should allow the user to enter several topics, modify their expression, see experts descriptions and social network information, filter experts lists, navigate teams, modify expertise/cohesiveness parameters.

In current implementation of T-RecS, user has to click first on "experts" to give some info on experts, and then on "teams" to display teams according to the desired skills she has suggested. We have designed a tab panel (see Figure 2, A) for expertise retrieval. Each tab corresponds to a skill entered by the user. The tree view of categories (D) allows the user to (i) disambiguate her queries, by selecting the correct expression and (ii) explore Wikipedia categories and filter information, together with some other filtering options (C) (faceted search). Each time she modifies anything, the experts list is refreshed (B). User has also the possibility to see a full description of each academic and the social network centered on this academic.

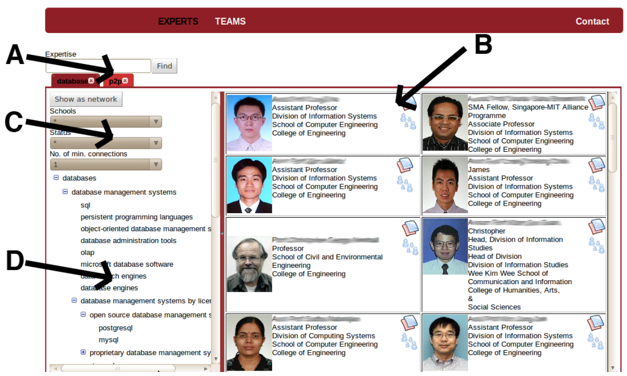

Figure 2: Expertise retrieval page.

The user can then discover the list of teams suggested by the system with current team recommendation parameters (see Figure 3). She can modify the relative importance of each topic, e.g. increase importance of one skill, as it is central for the task (E). She can also play with some parameters (F). Another interesting option, e.g for an ego-centric search (i.e. to find people to work with) is to add an academic to every team (G). Thus, the user can see who to contact regarding a particular project. Finally she sees the different teams, and can navigate a graphical/networked version of each team, in order to figure out current parameters and how people are linked (H).

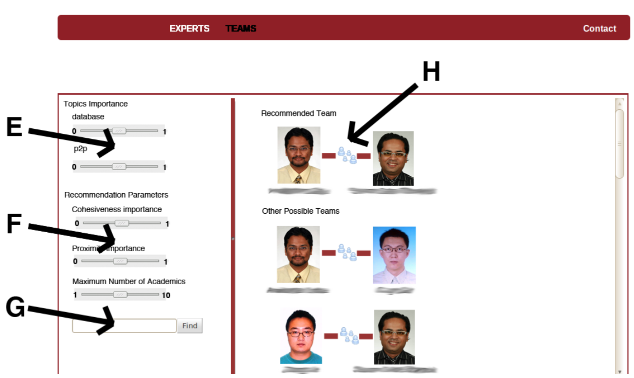

Figure 3: Team Recommendation page.

5. DEMONSTRATION

We will demonstrate the several elements of the multidisciplinary team recommendation system, starting from the expertise retrieval, to the team recommendation and exploration. We will play out two scenarios: A third party looking for the best multi-disciplinary scientific team to carry out a task or project, as well as a ego-centric search of exploring multi-disciplinary team personalized to an individual researcher's needs and social connections. Note that such teams indeed often need to be identified and formed when looking for consultants or for applying for research grants, and T-RecS, even as a prototype, is a practical utility to help explore such potential candidates.

We will show how easily and efficiently we can find experts in T-RecS, and how a query can be refined (using faceted search and other filters, as well as query expansion mechanisms) or personalized (e.g., ego-centric exploration). We will see that user can easily disambiguate her query and find the correct topic expression. The expert list for each skill is a valuable indicator of who is important in the future team, while the experts visualization interface aims at providing a better understanding of an expertise network in the university and is an help to find potentially good team.

We will further demonstrate, that given different needs (high level expertise, very cohesive team, maximum number of individuals per team, etc.) the results can be reranked easily, providing customization of the results. In the demo, besides being exposed to the general framework and the underlying enabling algorithms, the audience can explore teams to understand impacts of some decisions and use T-RecS' deductions as feedback to dynamically modify team formation parameters.

6. REFERENCES

[1] A. Hupa, K. Rzadca, A. Wierzbicki, and A. Datta. Interdisciplinary matchmaking: Choosing collaborators by skill, acquaintance and trust. In *Computational Social Networks Analysis: Trends, Tools and Research Advances*, 2009.

[2] M.-T. Le, H.-V. Dang, E.-P. Lim, and A. Datta. Wikinetviz: Visualizing friends and adversaries in implicit social networks. In *ISI'08*, 2008.

[3] C.-Y. Lin, N. Cao, S. X. Liu, S. Papadimitriou, J. Sun, and X. Yan. Smallblue: Social network analysis for expertise search and collective intelligence. *ICDE'09*, 2009.

[4] D. W. McDonald and M. S. Ackerman. Expertise recommender: a flexible recommendation system and architecture. In *CSCW'10*, 2000.

[5] M. Newman. Who is the best connected scientist? a study of scientific coauthorship networks. *Complex Networks*, pages 337–370, 2004.

[6] S. Rodrigues, J. Oliveira, and J. de Souza. Competence mining for team formation and virtual community recommendation. *CSCWD'05*, 2005.

[7] P. Serdyukov and D. Hiemstra. Being omnipresent to be almighty: The importance of the global web evidence for organizational expert finding. In *SIGIR 2008 Workshop on Future Challenges in Expertise Retrieval (fCHER)*, 2008.

[8] C. Taramasco, J.-P. Cointet, and C. Roth. Academic team formation as evolving hypergraphs. *Scientometrics*, 14, 2010.

[9] D. Watts and S. Strogatz. Collective dynamics of small-world networks. *Nature*, 393, 1998.