Tell me what you Need, I Can Manage the Competition in Crowdsourcing systems

Iheb Ben Amor, Salima Benbernou, Mourad Ouziri

Tell me what you Need, I Can Manage the Competition in Crowdsourcing systems

Iheb Ben Amor
Université Sorbonne Paris Cité
Paris Descartes
iheb.ben-amor@etu.parisdescartes.fr

Salima Benbernou and Mourad Ouziri
Université Sorbonne Paris Cité
Paris Descartes
firstname.lastname @ParisDescartes.fr

ABSTRACT

Crowdsourcing is a practice of allowing enterprise to use the creativity and the imagination of Internet users to provide a solution of an issue i.e. web content. However, the task to be accomplished is more and more complex and therefore it needs to find unknown collaborative and competitive group of solvers. In this paper,we propose a crowdsourcing system called *SocialCrowd* allowing the composition of collaborative team that should provide the best solution and treat that solution as a trade secret avoiding data leak between competitive teams due to reward behind the outsourcing of the complex task

Keywords

Crowdsourcing, Social network,privacy

1. INTRODUCTION

Crowdsourcing is emerging as a powerful paradigm in human problem solving techniques to the Open World. It helps to solve a wide range of tedious and complex problems in various enterprise applications. It allows enterprise to use the creativity and the imagination of Internet users to provide a solution for an issue. We introduce a type of crowdsourcing called *Socialcrowd* based on the efficiency of social network to outsource a task to be performed by people on demand on the social/professional network instead of an open world as Mechanical Turk is doing. As related work one can cite [3] which studies the max/top-k and grouping database problems in the crowd sourcing setting. where the criteria used for grouping and ordering are difficult to evaluate by machines but much easier to evaluate by the crowd. In [1], we developed an approach discovering hidden relationships between crowd members in social network and building only one team that may solve the outsourced problem, in [2], we extended our previous work in order to discover a set of collaborative et competitive team compositions answering the business call. In this paper we propose to discuss a data leak free team composition method which is based on heuristics.

2. SOCIALCROWD: AN OVERVIEW

The *SocialCrowd* architecture is composed of three main components as depicted in Fig.1:

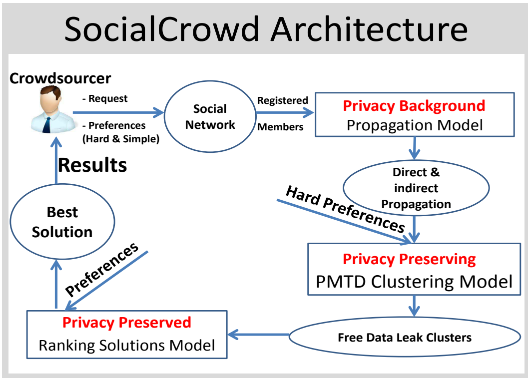

Figure 1: *SocialCrowd* architecture

- Data propagation component: In order to compose data leak free teams, the component computes the propagation of data in the social network.
- PMTD Clustering Model: To discover competitive and collaborative teams, we provide **PMTD** algorithm based on a specific distance representing the propagation for clustering crowd workers that capture data leaks.
- Ranking Solutions Model: Based on the possible team compositions provided by the PMTD model, this module will use the requester preferences to choose the best solution, it will combine the preferences in order to find the solutions that satisfy the majority of them.

3. PROBABILISTIC METAHEURISTIC BASED TEAM COMPOSITION

In this section we introduce our approach SocialCrowd by detailing the Data propagation component and the PMTD Clustering Model.

3.1 A Markov chain-based approach for data propagation

We propose an algorithm to calculate the optimal probability of data propagation from a given member to all the members of the social network in order to capture the possible data leak between crowd members. This algorithm is based on an energy function we define as follows:

DEFINITION 1. *(Energy function).*
The energy function p_i is the probability that data is propagated to the member m_i. Since in our model data is propagated following a Markov chain:$p_i = \underset{m_k \in N_{m_i}}{Max} (p_k \times p_{ki})$, *where, N_{m_i} is a set of direct friends of m_i, p_k is the energy function of m_k and p_{ki} is the probability of propagating data from m_k to m_i.*

The maximum function is the suitable aggregation function as the purpose of the approach is to estimate the maximum risk of data propagation. The aim is to compute the data probability propagation p_{ij} and p_{ji} of all the pairs of members (m_i, m_j). The proposed algorithm computes the optimal energy functions $P_{o*} = (p_{o1}, p_{o2}, ..., p_{on})$, which represent probabilities of data propagation from the member m_o to members $\{m_i\}_{i \in [1,n]}$. It processes as follows:

- Initialisation:$p_{owner} = 1, \forall i \neq owner \ p_i = 0$. That is only the owner has the data.
- Iterations: At each iteration, the algorithm computes p_i for $m_i \in N_{m_i}$using the energy function.
- Stops: The algorithm stops when the maximum probability of each member is reached.

3.2 Data leak aware heuristic-based composition

The composition model is based on data propagation technique defined in the previous section. In contrast to existing clustering algorithms (such as k-means [5] which generates a single clustering solution, our proposed approach generates all possible compostion solutions while preserving their respective data. As this is a hard problem, we use the Simulated annealing (SA) algorithm to compute near optimal solutions within a reasonable computation time [4]. The PMTD starts at a high temperature T_0 and the temperature is slightly lowered under a certain mechanism which is called cooling schedule when the procedure proceeds. In this paper, we use the exponential cooling schedule,$t_i = \alpha \times t_{i-1}$,where $\alpha \in (0, 1)$ is the temperature decreasing rate. The fundamental idea is to generate a new user sequence λ a random rule from the neighbourhood of incumbent sequence x. The new solution λ is assessed by a mechanism called acceptance criterion to decide whether accept the λ or reject it. The neighbourhood generating function provide new solutions neighbour to the latest accepted solution. The main core of the function is to check the admissibility of the new neighbour solution by checking it's occurrence and validate that is respecting hard preferences of the caller, that means, the mandatory criterion of the solution accepted by the requester. As mentioned before, in the single-objective SA, the variation is calculated $\Delta C = f(\lambda) - f(x)$. If $\Delta C \leq 0$, solution λ is accepted. Otherwise, solution λ is accepted with a probability equal to $P_r = \exp\{-\Delta C \div t_i\}$. Last but not least, t_i an operator is employed to generate a neighbour solution λ from the current candidate solution x by making a slight change in it. This operator performs so as to avoid producing infeasible solutions. In this paper, we take into consideration three different move operators :Swap,Single relocation or New Team creation.The objective function of the PMTD algorithm is based on the privacy, the homogeneous competency and promoting teamwork.
$T(\Phi) = Var(Connected\ Member\ in\ each\ team) + Var(Competency) + Privacy\ Computation$

4. SOLUTION RANKING

According to the large number of solutions returned by the PMTD algorithm, in this section, we use Soft Constraint Satisfaction Problem (SCSP) to rank the provided solutions using preferences. In order to answer a crowdsourcing request, the clustering operations to discover teams aware dara leak may give many solutions that fulfil the privacy conditions. In this case, we integrate the preferences of the requester (how launched the call) to choose the suitable solution based on the privacy criterion and the preferences. Let CS = $\langle S, D, V \rangle$ be a constraint system and P = $\langle def, con \rangle$ be the problem to be resolved, where V = con = {TeamPropagationThreshold *(tpt)*, TeamSnPropagation *(tsp)*, ExpertLevel *(el)*, Competency *(com)*}, D = $\{\{tpt_1, tpt_2, tpt_3, tpt_4\}$, $\{tsp_1, tsp_2, tsp_3, tsp_4\}$, $\{el1, el2, el3, el4\}$, $\{com1, com2, com3\}\}$, $S_p = \langle [0, 1], max, min, 0, 1 \rangle$, $C = \langle c_1, c_2, c_3, c_4 \rangle$. For simplicity, variables and their domains have been written in the same order. The solution with highest rank is chosen first.

5. CONCLUSIONS

In this paper, we discussed the data propagation approach for the problem of friendship propagation of crowd members in social/professional networks. We also proposed a composition algorithm using simulated annealing heuristic model for team discovering process while avoiding data leaks between composed teams.

6. REFERENCES

[1] I. B. Amor, S. Benbernou, M. Ouziri, M. Nadif, and A. Bouguettaya. Data leak aware crowdsourcing in social network. In *WISE*, 2012.

[2] I. B. Amor, M. Ouziri, S. Sahri, and N. Karam. Be a collaborator and a competitor in crowdsourcing system. In *MASCOTS*, 2014.

[3] S. B. Davidson, S. Khanna, T. Milo, and S. Roy. Using the crowd for top-k and group-by queries. In *ICDT*, 2013.

[4] S. Kirkpatrick, C. D. Gelatt, and M. P. Vecchi. Optimization by simulated annealing. *SCIENCE*, 220(4598):671–680, 1983.

[5] J. MacQueen et al. Some methods for classification and analysis of multivariate observations. In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability.* California, USA, 1967.