Collaborative Judgement

Ewa Andrejczuk[1,2(✉)], Juan Antonio Rodriguez-Aguilar[1], and Carles Sierra[1]

[1] Artificial Intelligence Research Institute (IIIA-CSIC), Barcelona, Spain
[2] Change Management Tool S.L., Barcelona, Spain
{ewa,jar,sierra}@iiia.csic.es

Abstract. In this paper we introduce a new ranking algorithm, called Collaborative Judgement (CJ), that takes into account *peer opinions* of agents and/or humans on objects (e.g. products, exams, papers) as well as *peer judgements* over those opinions. The combination of these two types of information has not been studied in previous work in order to produce object rankings. We apply CJ to the use case of scientific paper assessment and we validate it over simulated data. The results show that the rankings produced by our algorithm improve current scientific paper ranking practice based on averages of opinions weighted by their reviewers' self-assessments.

1 Introduction

In many areas of our lives we are used to the process of assessing and being assessed. We pass exams at the University, we go through job interviews, we undergo research project reviews, we are evaluated by our employers, etc. Artificial Intelligence research has focused on the assessment process for long and a number of algorithms have been developed to assist in assessing the performance of humans or artificial agents. Indeed large number of trust and reputation models have been proposed [3,12,15–17].

Surprisingly, to our knowledge, no significant effort has been put in the development of algorithms that use *judgement* information over such assessments. We consider exam marks unjust, interview outcomes biased, and review reports unfair, and we normally comment about these opinions on our performance with friends and relatives. We think that this kind of information is very important as it can be key to build the reputation of assessors. A bad assessor can be detected by the assessing community if they were allowed to simply express their opinions about the bad assessor. Actually, in many social networks this kind of information is collected ("was this recommendation useful to you?"), and they present this information to users but how the sites use this information to rank recommendations is never clearly explained if it is used at all.

Similarly, in the area of multiagent systems, agents' performance is key to build teams and coalitions [10]. Team formation and coalition formation are key for many applications related to multiagent cooperation, e.g. RoboCup rescue team [8,13], Unmanned Aerial Vehicles (UAVs) operations [5], team formation in social networks [7]. Both team formation and coalition formation focus on

Q. Chen et al. (Eds.): PRIMA 2015, LNAI 9387, pp. 631–639, 2015.
DOI: 10.1007/978-3-319-25524-8_46

forming the *best* possible group of agents (be it either a team or a coalition) to accomplish some tasks of interest given some limited resources. Hence, it is key in these processes to count on an assessment of the *expected capabilities* of the agents to recruit. With this aim, many trust models have been developed in the past to model agent behaviour [9,11], but judgements have again never been used to our knowledge.

In this paper we present an algorithm, called *Collaborative judgement* (CJ), which wants to go a step further in the use of judgements. CJ takes into account judgements on opinions to build reputation values on assessors and then use them as the basis to aggregate the opinions of a group of assessors. In current recommender systems the opinions about an object are aggregated using weights or not. When no weights are used, the final opinion is just an average of all the opinions (e.g. Amazon, TripAdvisor). When they are used the aggregated opinion is a weighted average using self-assigned weights. This is very common in Conference Management Systems like Confmaster or Easychair. In this paper we will compare CJ with the standard algorithm that weights opinions with the assessors self-assessments. We will call this simple algorithm *Self-Assessment Weighted Algorithm* (SAWA).

Here we will particularize the problem of peer judgement to the case of Conference Paper reviewing. The need to improve the way conferences (and to some extend journals) assess papers is key for scientific progress and its pitfalls have been discussed recently, see for instance the NIPS experiment: http://blog.mrtz.org/2014/12/15/the-nips-experiment.html. Some researchers have been trying to ameliorate the situation by improving the paper assignment process [1]. However, there is a growing phenomenon in which reviews are not made nor supervised by the expert member of the program committee but by someone to whom the reviewing task was delegated (e.g. a PhD student). This would invalidate this potential improvement. Here we propose to adapt CJ to detect those non-expert reviewers and dismiss their opinions from the final decision on accepting a paper. Henceforth, the notation we will use will be based on the ontology of a conference: papers, reviewers, marks, ...

In Section 2 we present the ranking algorithm that we benchmark in Section 4 against SAWA, presented in Section 3. Then, in Sections 5 and 6 we discuss the results and summarise our main achievement and outline our future work.

2 Collaborative Judgement

We first introduce the notation (focused on the case of paper assessment), which we will use in the rest of the paper.

Definition 1. *A conference is a tuple* $\langle P, R, E, o, v \rangle$*, where*

- $P = \{p_i\}_{i \in \mathcal{P}}$ *is a set of papers.*
- $R = \{r_j\}_{j \in \mathcal{R}}$ *is a set of reviewers.*
- $E = \{e_i\}_{i \in \mathcal{E}} \cup \{\bot\}$ *is a totally ordered evaluation space, where* $e_i \in \mathbb{N}$ *and* $e_i < e_j$ *iff* $i < j$ *and* $\bot$ *stands for the absence of evaluation.*

- $o : R \times P \to E$ *is a function giving the opinions of reviewers on papers.*
- $v : R \times R \times P \to E$ *is a function giving the judgements of reviewers over opinions on papers.*[1] *Therefore, a judgement is a reviewer's opinion about another reviewer's opinion.*

In general we might have different dimensions of evaluation, that is a number of E spaces over which to express opinions and judgements. For instance, originality, soundness, etc. but for simplicity reasons we will assume that the evaluation of a paper is made over a single dimension. Actually, the 'overall' opinion is what is aggregated in real systems.

The steps of the CJ algorithm applied over a conference $\langle P, R, E, o, v\rangle$ are as follows:

1. Compute the *agreement level* between reviewers r_i and r_j, $a : R \times R \to [0,1] \cup \{\bot\}$. If the reviewers had some papers to review in common then the judgements on the opinions, in case they exist, and the similarity between the opinions over the common papers are combined as follows:

$$a(r_i, r_j) = \begin{cases} \frac{\sum_{p_k \in P} s(r_i, r_j, p_k)}{|P_{ij}| \cdot d} & \text{if } P_{ij} = \{p_k \in P | o(r_i, p_k) \neq \bot, o(r_j, p_k) \neq \bot\} \neq \emptyset \\ \bot & \text{otherwise} \end{cases}$$

   where d is the maximum distance in the evaluation space and:

$$s(r_i, r_j, p_k) = \begin{cases} v(r_i, r_j, p_k) & \text{if } v(r_i, r_j, p_k) \neq \bot \\ Sim(o(r_i, p_k), o(r_j, p_k)) & \text{otherwise} \end{cases}$$

   and *Sim* stands for an appropriate similarity measure. When no explicit judgements are given, the similarity in opinions is considered a good heuristic for them. The more similar a review is to my opinion, the better I judge that opinion.
2. Compute a complete *Trust Graph* as an adjacency function matrix $C = \{c_{ij}\}_{i,j \in R}$.

$$c_{ij} = \begin{cases} a(r_i, r_j) & \text{if } a(r_i, r_j) \neq \bot \\ \max\limits_{h \in chains(r_i, r_j)} \prod\limits_{(k,k') \in h} a(r_k, r_{k'}) & \text{otherwise} \end{cases}$$

   where $chains(r_i, r_j)$ is the set of sequences of reviewer indexes connecting i and j. Formally, a chain h between reviewers i and j is a sequence $\langle l_1, \ldots, l_{n_h}\rangle$ such that $l_1 = i$, $l_{n_h} = j$, and $a(r_k, r_{k+1}) \neq \bot$ for each pair $(k, k+1)$ of consecutive values in the sequence. To compute this step we use a version of Dijkstra's algorithm that instead of looking for the shortest path (using $+$ and min) it looks for the path with the largest arc product (using $\cdot$ and max). The running time of the Dijkstra algorithm can take $O(n \log n)$, where $n = |R|$, if using priority queues [2].

[1] In tools like ConfMaster (www.confmaster.net) this information could be gathered by simply adding a private question to each paper review, answered with elements in E, one value in E for the judgement on each fellow reviewer's review.

3. Compute a *reputation* for each reviewer in R, $\{t_i\}_{i \in R}$, by using Eigentrust [6]. In order for this to be applicable we need to guarantee that the graph C is aperiodic and strongly connected. In this step we obtain a global trust value for each reviewer. In vectorial notation, the trust vector is assessed as $\bar{t} = \lim_{k \to \infty} \bar{t}^{k+1}$ with $\bar{t}^{k+1} = C^T \bar{t}^k$ and $\bar{t}^0 = \bar{e}$ being $\bar{e}_i = 1/|\bar{e}|$. The complexity of the Eigentrust algorithm used in this step is $O(n^2)$. In our case, we cannot force the values in a row of C to add up to 1, as required by the Eigentrust algorithm, so we do normalize the trust vector as generated after each step to guarantee convergence.
4. Compute the *final opinion* on objects as a weighted average of the opinions of those that expressed an opinion. The weights are the reputation of those expressing an opinion:

$$o_{CJ}(p_j) = \frac{\sum_{i \in \{i \in R | o(r_i, p_j) \neq \perp\}} \bar{t}_i \cdot o(r_i, p_j)}{\sum_{i \in \{i \in R | o(r_i, p_j) \neq \perp\}} \bar{t}_i}$$

3 The SAWA Algorithm

We will benchmark CJ against the algorithm used by the conference management systems mentioned in the introduction, which we call in this paper SAWA. We assume there is a function $r : R \times P \mapsto [0, 1]$ that keeps how confident each reviewer feels about her opinion on a paper. So the aggregated opinion on a paper is computed as:

$$o_{SAWA}(p_j) = \frac{\sum_{i \in \{i \in R | o(r_i, p_j) \neq \perp\}} r(i, p_j) \cdot o(r_i, p_j)}{\sum_{i \in \{i \in R | o(r_i, p_j) \neq \perp\}} r(i, p_j)}$$

4 Evaluation

In this section we validate the algorithm via simulation. We show that CJ behaves according to expectations with respect to SAWA. The hypotheses we are interested in are:

H1 **CJ rankings get closer to the true quality of a paper when the number of good reviewers increase.**[2]
H2 **Ceteris paribus, the better the reviewers, the larger the improvement of CJ with respect to SAWA.**
H3 **The overall trust on reviewers positively correlates with the number of good reviewers.**

We next explain the experimental setting and three experiments providing support to these hypotheses.

[2] See next subsection for our representation of a good reviewer.

4.1 Experimental Setting

We assume a set $P = \{p_1, \ldots, p_n\}$ of papers and a function for their true quality in a range $[0, 1]$,[3] $q : P \to [0, 1]$. We use the following evaluation space $E = \{1, 2, 3, 4, 5, 6, 7, 8, 9, 10\}$, which is rather common in the context of paper reviewing. We assume two types of reviewers: good and bad, with the following behaviour:

- *Good Reviewer.* She provides fair opinions and fair judgements. Her opinion on any paper p_k is always close to its true quality $q(p_k)$. We assume the absolute value of the difference between the opinion of a reviewer and the true quality (as a percent) follows a beta distribution, $Beta(\alpha, \beta)$, very positively skewed, for instance with $\alpha = 1$ and $\beta = 30$. For each paper p_k reviewed by a good reviewer, we sample the reviewer's associated beta distribution for a percentage difference, apply it to the paper quality $q(p_k)$ (up or down randomly) and round the result to fit an element in E. Her judgements on someone's opinion are close to 0 if the opinion is far from the true quality of the paper, and close to 1 otherwise. We implement this as the following function:

$$v(r_i, r_j, p_k) = 1 - |o(r_j, p_k) - q(p_k)|$$

  and self-judgements from $Beta(5, 2)$, slightly negatively skewed.
  We assume that when a good reviewer judges a bad reviewer she samples a value in E from a beta distribution rather positively skewed: $Beta(2, 40)$. The intuition is that good reviewers poorly mark bad reviews.
- *Bad Reviewer.* She provides unfair opinions, because she is incompetent, but provides reasonable judgements as she can interpret the opinions of others as being informative or not. Thus, we sample opinions from $Beta(20, 12)$ — rather central with a slight negative skew, judgements for good reviews and self-judgements from $Beta(5, 2)$ as for good reviewers —negatively skewed, and judgements on bad reviews from $Beta(2, 5)$ —slightly positively skewed. The overall idea is that bad reviewers stay mostly in the central area of the evaluation space.

We use $Sim(x, y) = (|E| - 1 - |\tau(x) - \tau(y)|)/(|E| - 1)$ as a simple linear similarity function where τ is a function that gives the position of an element in the ordered set E.

Experiment 1: We set an increasing percentage of good reviewers, from none to 100%. We plot the improvement, that is, the error reduction, using the Mean Absolute Error of the values generated by the two ranking methods (CJ and SAWA) and the true quality of the papers. That is, we plot $(1 - (MAE(O_{CJ}, q)/MAE(O_{SAWA}, q))) \cdot 100$ where

$$MAE(f, g) = \frac{\sum_{p_j \in P} |f(p_j) - g(p_j)|}{|P|}.$$

[3] Assessing the true quality of an object may be difficult and it is certainly a domain dependent issue.

In Figure 1 we see this improvement for 10 runs of the algorithms. We observe that CJ improves SAWA and the improvement becomes larger than 10% and statistically significant for percentages of good reviewers between 20% and 80%. These results support H1.

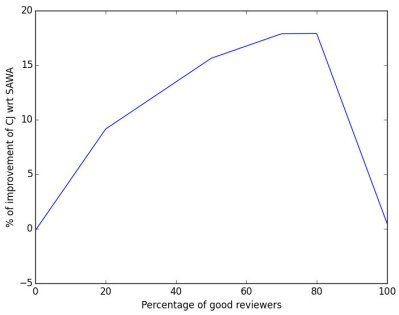

Fig. 1. Percentage of error improvement of CJ over SAWA with the error measured as the Average Mean Absolute Error with respect to the true quality of papers for increasing percentages of good reviewers. The parameters are those explained in the experimental setup.

Experiment 2: As mentioned before, we model good reviewers' opinions with a $Beta(\alpha, \beta)$ very positively skewed from which we sample the difference between the reviewer's opinion and the true quality. With $\alpha = 1$ and $\beta > 30$ the expert is frequently telling the true quality in her opinions (specially because we discretise the sampled values into our evaluation space —i.e. almost all the distribution mass is rounded to a distance of 0 with respect to the true quality). In figure 2 we plot the improvement of CJ with respect to SAWA for $\alpha = 1$ and increasing values of β (better reviewer behaviour). We observe that the improvement asymptotically grows to 16%, and hence this supports Hypothesis H2.

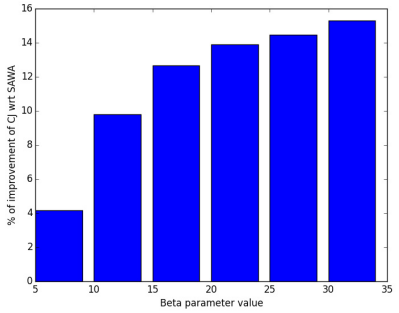

Fig. 2. Improvement of CJ over SAWA as the reviewers' quality increases (with fixed $\alpha = 1$ and increasing β values). This plot is for a population with 50% good reviewers and 50% bad reviewers.

Experiment 3: In Figure 3 we see the increasing mutual trust between reviewers as the average of all values in matrix C with respect to an increasing percentage of good reviewers. This supports Hypothesis H3.

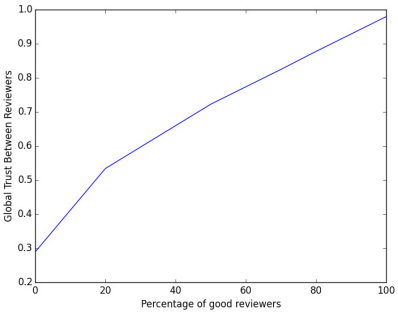

Fig. 3. Increasing mutual trust of reviewers for an increasing percentage of good reviewers.

5 Discussion

One issue worth discussing is the feasibility of getting real data to model $q(\cdot)$. We mentioned before that this is obviously a domain dependent issue and that it can be difficult to obtain. For instance, in the case of paper review, what is the true quality of a paper? It seems impossible to answer this question. We could get data on impact of papers and assume that impact relates to quality. This can be done for the papers that were accepted and published, but not for those that were rejected. Therefore, the validation of the algorithm results will necessarily be partial. This will always be controversial as the use of any quality metric would always be debatable. It is in this context that our algorithm contributes since the key assumption of our algorithm is: *when there is no clear-cut method to determine the quality of an object, then the true quality can be determined by the social acceptance of the opinions expressed by experts.*

Another issue worth mentioning is that reviewer quality depends on the particular subarea of a conference. In general, our opinions are fair or not depending on our competences. Thus, CJ should consider this dimension as many existing trust models do [10,14]. The inclusion of a semantic dimension on trust and reputation requires defining an ontology of the domain and semantic distances between the elements in the vocabulary. This represents no technical problem and will basically increase the complexity of the computation proportionally to the granularity of the vocabulary.

6 Conclusions and Further Work

In this paper we introduced CJ. It is a new ranking algorithm that takes into account *peer opinions* of agents and/or humans as well as *peer judgements* over those opinions. We applied CJ to the use case of scientific paper assessment and we validated it over simulated data. The results show that the rankings produced by this new algorithm (under (reasonable) assumptions on reviewer behaviour) improve current scientific paper ranking practice. The use of this algorithm in the context of agent team formation is key as it will provide a sound method

to assess the *capabilities* of agents by observing peer opinions and judgements made by agents and humans.

Part of the future work is centred on evaluating CJ over real data. We are planning to get data from a commercial bank about the skills of team members. That is, employees work in teams to solve tasks that require specific skills. Team members record opinions on their team-mates' skills and judgements on the opinions after anonymisation. We are also discussing the extension of functionalities of a major conference management system so that we can get data on judgements in conferences in the near future.

At simulation level we want to further explore the sensitiveness of the results for varying parameter settings, including the impact of similarity functions as these have not been relevant in the reported experiments. Finally, the modelling of *malicious* reviewers (those who know the quality of a paper and deliberately lie about it) will be considered. We expect that our method might help in detecting those reviewers.

In many settings, including conference paper rankings, the actual numerical value is not the key element but the order between alternatives. CJ produces a *partial ranking* among alternatives, that is, there can be ties between objects (e.g. papers). We plan to compare CJ vs SAWA using the generalisation of the Kendall Tau distance proposed in [4] to compare partial rankings.

Finally, this algorithm is an important milestone on our path to develop methods to build agent and human teams to solve complex tasks that balance capabilities and mutual relationships.

Acknowledgments. The first author is supported by an Industrial PhD scholarship from the Generalitat de Catalunya. This work is also supported by the CollectiveMind project (Spanish Ministry of Economy and Competitiveness, grant number TEC2013-49430-EXP) and the COR project (TIN2012-38876-C02-01).

References

1. Charlin, L., Zemel, R.S., Boutilier, C.: A framework for optimizing paper matching. CoRR, abs/1202.3706 (2012)
2. Cormen, T.H., Stein, C., Rivest, R.L., Leiserson, C.E.: Introduction to Algorithms, 2nd edn. McGraw-Hill Higher Education (2001)
3. de Alfaro, L., Shavlovsky, M.: Crowdgrader: Crowdsourcing the evaluation of homework assignments. Thech. Report 1308.5273, arXiv.org (2013)
4. Fagin, R., Kumar, R., Mahdian, M., Sivakumar, D., Vee, E.: Comparing and aggregating rankings with ties. In: Proceedings of the Twenty-Third ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems, PODS '04, pp. 47–58. ACM, New York (2004)
5. Haque, M., Egerstedt, M., Rahmani, A.: Multilevel coalition formation strategy for suppression of enemy air defenses missions. Journal of Aerospace Information Systems **10**(6), 287–296 (2013)
6. Kamvar, S.D., Schlosser, M.T., Garcia-Molina, H.: The eigentrust algorithm for reputation management in p2p networks. In: Proceedings of the 12th International Conference on World Wide Web, WWW '03, pp. 640–651. ACM, New York (2003)

7. Lappas, T., Liu, K., Terzi, E.: Finding a team of experts in social networks. In: Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '09, pp. 467–476. ACM, New York (2009)
8. Nair, R., Tambe, M., Marsella, S.C.: Team formation for reformation in multiagent domains like RoboCupRescue. In: Kaminka, G.A., Lima, P.U., Rojas, R. (eds.) RoboCup 2002. LNCS (LNAI), vol. 2752, pp. 150–161. Springer, Heidelberg (2003)
9. Osman, N., Gutierrez, P., Sierra, C.: Trustworthy advice. Knowl.-Based Syst. **82**, 41–59 (2015)
10. Osman, N., Sierra, C., McNeill, F., Pane, J., Debenham, J.K.: Trust and matching algorithms for selecting suitable agents. ACM TIST **5**(1), 16 (2013)
11. Osman, N., Sierra, C., Sabater-Mir, J.: Propagation of opinions in structural graphs. In: Coelho, H., Studer, R., Wooldridge, M. (eds.) ECAI 2010–19th European Conference on Artificial Intelligence, Lisbon, Portugal, 16–20 August 2010. Frontiers in Artificial Intelligence and Applications, vol. 215, pp. 595–600. IOS Press (2010)
12. Piech, C., Huang, J., Chen, Z., Do, C., Ng, A., Koller, D.: Tuned models of peer assessment in moocs. In: Proc. of the 6th International Conference on Educational Data Mining (EDM 2013) (2013)
13. Ramchurn, S.D., Farinelli, A., Macarthur, K.S., Jennings, N.R.: Decentralized coordination in robocup rescue. Comput. J. **53**(9), 1447–1461 (2010)
14. Sierra, C., Debenham, J.K.: Trust and honour in information-based agency. In: Nakashima, H., Wellman, M.P., Weiss, G., Stone, P. (eds.) 5th International Joint Conference on Autonomous Agents and Multiagent Systems (AAMAS 2006), Hakodate, Japan, 8–12 May 2006, pp. 1225–1232. ACM (2006)
15. Walsh, T.: The peerrank method for peer assessment. In: Schaub, T., Friedrich, G., O'Sullivan, B. (eds.) ECAI 2014–21st European Conference on Artificial Intelligence, 18–22 August 2014, Prague, Czech Republic - Including Prestigious Applications of Intelligent Systems (PAIS 2014). Frontiers in Artificial Intelligence and Applications, vol. 263, pp. 909–914. IOS Press (2014)
16. Wu, J., Chiclana, F., Herrera-Viedma, E.: Trust based consensus model for social network in an incompletelinguistic information context. Applied Soft Computing (2015)
17. Zhang, J., Ghorbani, A.A., Cohen, R.: A familiarity-based trust model for effective selection of sellers in multiagent e-commerce systems. Int. J. Inf. Sec. **6**(5), 333–344 (2007)