Social Mining as a Knowledge Management Solution

Mathijs Creemers and Mieke Jans

Hasselt University, Martelarenlaan 42, 3500 Hasselt, Belgium
mathijs.creemers@uhasselt.be

Abstract. We introduce knowledge management and identify common problems and challenges. We then introduce social mining, a technique from the field of process mining. We show that the techniques used in social mining could have added value for the knowledge management challenges. We also identify areas for improvement in social mining. The first of these concerns the joint activity metric. We argue that for this metric only correlation measures should be used and the distance measures should be re-evaluated. Next we propose the more detailed use of time information to add value to the existing metrics. Lastly we suggest a way of validating the connection in a social network. As a conclusion we find that social mining can be enhanced in these areas and then used as a tool to tackle knowledge management problems.

1 Introduction

Knowledge management is the process of capturing, distributing, and effectively using knowledge(Davenport,1994)[3]. A more formal definition is given by Duhon[5]:

> "Knowledge management is a discipline that promotes an integrated approach to identifying, capturing, evaluating, retrieving, and sharing all of an enterprise's information assets. These assets may include databases, documents, policies, procedures, and previously un-captured expertise and experience in individual workers."

We focus on the second part of this definition, meaning we consider the expertise and experience in individual workers. The purpose of this paper is to introduce some of the classical challenges in knowledge management and identify how process mining might be used to provide solutions.

Process mining is the bridge between data mining and process modelling. Starting from an event log we use data mining on this log to get process information. The main focus of process mining is process discovery, leading to an increasing amount of discovery algorithms. We can however also use it to find social networks, this is called social mining. Social mining is what we will use to establish the link with knowledge management. For this paper we use the

In: S. España, M. Ivanović, M. Savić (eds.): Proceedings of the CAiSE'16 Forum at the 28th International Conference on Advanced Information Systems Engineering, Ljubljana, Slovenia, 13-17.6.2016, published at http://ceur-ws.org

term social mining. If there is a chance for confusion with the practice of mining information off of social media, we suggest the term social process mining.

However, limitations of the current techniques are preventing a direct application of the social mining algorithms in the context of knowledge management. We discuss the potential areas of improvement to leverage the use of social mining for knowledge management challenges.

2 Knowledge management challenges

In knowledge management there are a few classic challenges[9,4]. We identify the ones that are interesting and later show techniques from social mining that might be applied to them.

A common activity in knowledge management is the generation of a knowledge map that shows where knowledge is located in the company. This might contain information about knowledge repositories but also about who knows what. This can be constructed using surveys but it's hard to keep such a map up to date. People might change what they work on or their contact information might change.

Another topic is team selection and the selection of the leader for such a team. For the latter activity, we require an up to date list of the skills of our employees. It might also be interesting to know who has worked together before or which employees already communicate often. To select a leader we need to find a person that is central in both the trust and the advice network. If both these networks are up to date it will be easier to find a leader.

Part of knowledge management is the creation of new ideas. An often used technique for this is fusion. This is a concept where people with a different background (but with some common ground) are put together in a room in a bid to generate new ideas. It's based on the idea that innovation occurs at the boundaries between mind-sets, not within. Again it appears to be important to know what skills people have and what they have recently been working on.

A big problem in knowledge management is brain drain. This is a phenomenon that appears when valuable knowledge leaves the company. The main solution for this problem as identified by knowledge management literature is making sure there is enough transfer of knowledge. This both includes appointing an experienced mentor to a novice entering the company and transcribing or codifying the knowledge of people leaving the company. Sharing experience is another aspect of this. If everyone shares their experiences this means that knowledge will propagate throughout the company.

3 The details of social mining

In general process mining focusses on the discovery of a process model for a given event log. Another branch of process mining uses the event log to create

social networks. This is called social mining. Since this concerns the creation of social networks and the analysis of people and the activities they perform, we argue it can be applied to knowledge management.

In process mining an event log is a log listing the events that have occurred. Here, each event is an atomic registration of an action in a process and contains at least a timestamp and a link to a case.

For the purpose of extracting a social network, a resource field for each activity is also mandatory. This resource field indicates the person (or machine) that undertook the activity. Using the resource field we can construct a social network based on the information in the log. In this paper we assume the resource field always indicates a unique employee. The terms resource, person and employee are used interchangeably.

A social network is a graph in which each vertex represents a person. The edges between these vertices then show the connections between these people. An example of such a social graph can be seen in Figure 1. There are multiple ways to create these social graphs based on bibliographical information, surveys or even mailing data. [7,8]

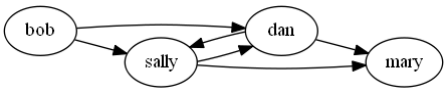


Fig. 1: An example social network

Case	Activity	Resource
1	Register	Bob
1	Negotiate	Sally
1	Sign	Dan
1	Sendoff	Mary
2	Register	Bob
2	Negotiate	Sally
2	Sign	Dan
2	Negotiate	Sally
2	Sendoff	Mary

Table 1: The log

The social network in Figure 1 is created by applying the handover of work technique to the log found in Table 1. This technique is a part of the possible causality metric of social mining.

Social mining was introduced by Song and Van Der Aalst[1,10,11]. They use the resource field and other information from the log to determine if there should be a connection between a pair of resources or not. There are different ways to calculate these connection, they are grouped in the following four metrics.

1. Possible causality (containing handover of work)
2. Joint cases
3. Joint activities
4. Special event type

The first metric assumes that when people handover cases to one another, they have a connection. This means that following events lead to a connection

between these resources in the social graph. Part of this metric is the handover of work metric which counts immediate handovers between people. Another example is subcontracting, which counts cases that get handed from person A to person B and then get handed back to person A. In a knowledge management context, this metric can be used to identify the flow of information between resources.

The second metric establishes a connection between two resources if they work on the same case.It is harder to establish relevance for this metric in the described knowledge management problems. However if we assume that a case corresponds to a project, this metric can be used to find out who worked on projects together. That information could then be used when we are selecting a team or a team leader.

The third metric yields a connection if two resources perform the same activities. Typically if a company is divided in functional departments this is the case.

This metric is promising for a lot of the knowledge management problems. This metric can be used to create profiles that relate people with the activities they perform, identifying the skills of a person. Knowing these skills can help, not only in team selection, but also when we are generating a knowledge map. It can also identify the risk of a brain drain if it is know that an activity is only performed by one employee. These profiles can also be used when trying to generate new ideas. Employees can be selected with both different skills and a few common skills to simulate fusion.

The last metric relies on special event types. Imagine an event that represents the activity of assigning a task to someone else. In this case it is known that the resource gives a task to someone else, so there is a connection between the two. Since it's so specific this metric does not have a clear knowledge management application.

The field of social mining dates back to 2004, having been originally introduced by Van Der Aalst[1].However no recent work builds on it. The last paper we could find that uses process mining as a tool of creating a social network stems from 2008[10]. We think this proves a chance to pick up an area of process mining that is now less used.

4 Social mining's areas of improvement

In the previous section we have given the details of social mining linked to the knowledge management challenges. There are opportunities for added value, however the current state of social mining still shows some room for improvement. There are a few areas which we first wish to improve before applying social mining to knowledge management.

We identify three main areas for improvement. The first area is of a smaller scale, concerning the distance measures used in the joint activity metric. As a next improvement we suggest that added value can be created by taking timestamps into account. A last improvement opportunity is the validation of the

connections in the social network. We suggest using email traffic as a way of verifying the connection between two people. We will now look at these possible improvements in more detail.

4.1 Joint activity metric

The joint activity metric assumes that there is a relation between people who perform the same activities. To calculate this metric, a performer-activity matrix is created. In this matrix every row corresponds to a resource and every column to an activity. The cells of the matrix represent the number of times the resource performs the activity.An example of such a matrix, related to the log of Table 1, can be found in Figure 2. To determine if there is a connection between two resources, the distances between the rows in the matrix are calculated. To do this calculation a few distance measures and a correlation measure are proposed [11]. The result of these measures is then used to determine if there should be a connection between the resources corresponding to these rows.

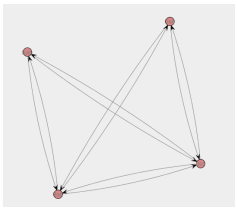
	Register	Negotiate	Sign	Sendoff
Bob	2	0	0	0
Sally	0	3	0	0
Dan	0	0	2	0
Mary	0	0	0	2

Fig. 2: An example of a matrix representing the resources and the activities they carry out

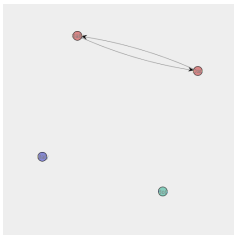
The measures proposed (based on the implementation of the metrics in ProM 6 [12] a process mining toolkit) are euclidean distance, Pearson correlation coefficient, Hamming distance and a similarity coefficient. However the original paper also mentions Minkowksi distance. The main improvement that can be made here is to add clarity for the end user. Using the current implementation, mining the same log with a different measure can lead to opposite results. An example of this can be found in Figure 3. Here we show the social networks mined from the same log, once using euclidean distance and once using the Pearson correlation coefficient. Using the euclidean distance connects every one in the network except for Dan and Sally whereas using the Pearson correlation coefficient only connects Dan and Sally.

The problem with using a mix of correlation and distance measures is that while a correlation measure yields a result in $[-1, 1]$ a distance measure yields a result in the range $[0, \infty[$. A threshold needs to be determined such that values above(or below) this threshold signify a connection. When using the Pearson correlation coefficient, a threshold value of 0.0 is suggested, but no threshold values are given for the distance measures [10].

Establishing a threshold value for the distance measures proves to be more difficult. Consider for example the Hamming and the euclidean distance. The



(a) the social network using euclidean distance



(b) The social network using the Pearson correlation coefficient

Fig. 3: Side by side comparison of both social networks

Hamming distance has a maximum that corresponds with the number of activities in the matrix. The euclidean distance however has a maximum depending on the amount of times activities are preformed. When applying these measure to the example from Figure 2, we can use these metrics to find the distance between Bob and Sally. The Hamming distance between them is 2 and the euclidean distance between them is $\sqrt{13} \approx 3.6$. Assume now that instead of 3, Sally performs the negotiate activity 10000 times. Using the Hamming distance as currently implemented in ProM, every non-zero value of the matrix is rescaled to one. This means the Hamming distance between both of them stays unchanged. However with the new values the euclidean distance between them becomes $\sqrt{(}100000004) = 26\sqrt{(}147929) \approx 10000$. This makes it very clear that both distance measures operate on a different scale.

Our suggestion is to drop the use of distance measures and instead focus on correlation measures. We should look at the different correlation measures to determine which ones are most appropriate.It is not guaranteed that the Pearson correlation coefficient is the most appropriate one. Using these measures there is already a possible threshold value at 0.0 and there is no danger of them operating on a different scale since they stay between -1 and 1. It will also make it easier for the end user to identify what measure to use since they will have to choose between the same sort of measures.

4.2 Timestamps

The original approach only uses timestamp information to order the activities. We believe the addition of more timestamp related information can add value to the results obtained. The addition of timestamps can be considered an improvement of a different scale since it affects more than one metric. There are a few reasons why the addition of this information to the metrics might add value.

We can assume that people who are performing the same activity at the same time have a better chance of knowing each other than people performing the same activity at another time of day. Similarly we can assume that when there

is an instant follow up, that this indicates a stronger connection. This means the handover of an activity from person A to person B with only 5 minutes in between corresponds to a stronger connection than the same handover with hours or even days in between.

Time related information is also interesting when it comes to joint activities or cases. It could be used to find out when a resource last performed an activity or when they were last in a case with another resource. This is important when trying to determine the skills of a person. A person might have done activity A a thousand times, but never in the last five years. Whereas another person might have done the same activity only five hundred times, but more recently. The latter person will probably be able to answer our questions about activity A more rapidly than the former.

We thus propose to examine all existing metrics and develop expanded metrics making use of the time related information. This information can then be used to add weight to the connections in the social network or to determine what happened more recently and thus is more important.

4.3 Validation of connections

The third area where we see room for improvement relates to the validation of connections in our social network. The different networks (created using the different metrics) are successful at showing the flow of information and the grouping of resources according to activities. It would be interesting to be able to conclude that there is a certain flow of communication based on the information in these networks. Right now this is not always the case. Consider for example two people who perform sequential activities, meaning a connection would be established when using the possible causality metric, but that are grouped in different departments and don't know one another.

Our proposal here is to enrich the social networks with information from the email system. Mining the emails sent in a company we can see actual communication between people [2][6]. This allows us to validate connections in our social network by checking if there is email activity between the connected people.

5 Conclusion

We started off by introducing knowledge management and some challenges from the domain. The main challenges are the brain drain, the creation of knowledge maps and team selection. We then proposed social mining as a possible way to add value. We did this by liking the social mining metrics with the knowledge management challenges.

We suggest some areas of improvement for social mining to be addressed before applying social mining to knowledge management. There are three areas of improvement. The first addresses the distance and correlation measures used in the joint activity metric. Here we suggest to drop the distance measures and focus on correlation metrics. A second suggestion concerns timestamp information.

This information is only used to order the events. We see chances to add value by using more details of this information. A last improvement lies in the validation of connections in the social network. Here we propose the use of existing email mining algorithms to validate the relation between connections in the social network and communication.

Brining these results together we have identified possible value improvement for knowledge management by using social mining techniques. We have also identified some steps to be taken first to improve the existing social mining techniques.

References

1. Van der Aalst, W.M., Song, M.: Mining Social Networks: Uncovering interaction patterns in business processes. In: Business Process Management, pp. 244–260. Springer (2004)
2. Culotta, A., Bekkerman, R., McCallum, A.: Extracting social networks and contact information from email and the web (2004)
3. Davenport, T.H.: Saving IT's Soul: Human-Centered Information Management. Harvard business review 72(2), 119–31 (1994)
4. Davenport, T.H., Prusak, L.: Working knowledge: How organizations manage what they know. Harvard Business Press (1998)
5. Duhon, B.: It's all in our heads. Inform 12(8), 8–13 (1998)
6. Farnham, S., Portnoy, W., Turski, A.: Using email mailing lists to approximate and explore corporate social networks. In: Proceedings of the CSCW. vol. 4 (2004)
7. Fisher, D., Dourish, P.: Social and temporal structures in everyday collaboration. In: Proceedings of the SIGCHI conference on Human factors in computing systems. pp. 551–558. ACM (2004)
8. Kazienko, P., Brdka, P., Musial, K., Gaworecki, J.: Multi-layered social network creation based on bibliographic data. In: Social Computing (SocialCom), 2010 IEEE Second International Conference on. pp. 407–412. IEEE (2010)
9. Koenig, M.E.: What is KM? Knowledge Management Explained (May 2012)
10. Song, M., van der Aalst, W.M.P.: Towards comprehensive support for organizational mining. Decision Support Systems 46(1), 300–317 (Dec 2008)
11. Van Der Aalst, W.M., Reijers, H.A., Song, M.: Discovering social networks from event logs. Computer Supported Cooperative Work (CSCW) 14(6), 549–593 (2005)
12. Verbeek, H.M.W., Buijs, J., Van Dongen, B.F., van der Aalst, W.M.: Prom 6: The process mining toolkit. Proc. of BPM Demonstration Track 615, 34–39 (2010)