A Conceptual Model of Actors and Interactions for the Knowledge Discovery Process

Lauri Tuovinen
Biomimetics and Intelligent Systems Group, University of Oulu,
P.O. Box 4500, FI-90014 Oulu, Finland

Keywords: Knowledge Discovery in Data, Process Model, Computer-supported Collaboration, Intelligent Systems.

Abstract: The knowledge discovery process is traditionally viewed as a sequence of operations to be applied to data; the human aspect of the process is seldom taken into account, and when it is, it is mainly the roles and actions of domain and technology experts that are considered. However, non-experts can also play an important role in knowledge discovery, and furthermore, the role of technology in the process may also be substantially expanded from what it traditionally has been, with special software facilitating interactions among human actors and even operating as an actor in its own right. This diversification of the knowledge discovery process is helpful in finding tenable solutions to the new problems presented by the current deluge of digital data, but only if the process model used to manage the process adequately represents the variety of forms that the process can take. The paper addresses this requirement by presenting a conceptual model that can be used to describe different types of knowledge discovery processes in terms of the actors involved and the interactions they have with one another. Additionally, the paper discusses how the interactions can be facilitated to provide effective support for each different type of process. As a future perspective, the paper considers the implications of intelligent software taking on responsibilities traditionally reserved for human actors.

1 INTRODUCTION

Twenty years ago, the seminal process model for knowledge discovery in data (KDD) was presented (Fayyad et al., 1996). In this model, the process is divided into five steps: selection, preprocessing, transformation, data mining, and interpretation/evaluation. Today, there are a number of KDD process models offering various augmentations and alternative perspectives to the seminal one, but the big picture has not changed radically, with the sequence of steps laid out by Fayyad et al. still being found at the heart of many of the more recent proposals. In fact, it is arguably the case that even if it is not explicitly acknowledged, this model is present in every KDD effort, since it prescribes a set of operations that are always required to get from raw data to actionable knowledge, even if they do not generally occur in a neat waterfall-style sequence.

The problem with the established KDD process model is not that it does a poor job of representing what it is intended to represent, but rather that what it represents is not the whole picture of KDD but just one aspect of it. The human aspect - the actors who participate in the KDD process and the interactions among them - is not addressed at all, and although there are some models that do account for it in some way, they are increasingly unsuitable for representing the variety of human-human and human-machine interactions that may be required in order to successfully complete a KDD effort. There is, therefore, a need for a collaborative process model that covers the full spectrum of actors and interactions involved.

A truly collaborative model of the KDD process is an important goal because traditional process models and methodologies are not designed to adequately respond to the challenges posed by the data deluge we are currently experiencing. According to a report published in 2012, 1.8 zettabytes ($1.8 * 10^{21}$ bytes) of data were generated globally in 2011 (Federal Big Data Commission, 2012), and the rate is likely to be even higher today. Even the volume of data available on the Web to any Internet-connected individual is overwhelming, and it is relatively effortless and inexpensive to collect more for one's own purposes using various sensors, such as those we already carry with us everywhere in our smartphones. The question, therefore, is not how we can obtain data, but how we can benefit from it.

Traditionally, the interactions that drive the KDD

Tuovinen, L.
A Conceptual Model of Actors and Interactions for the Knowledge Discovery Process.
DOI: 10.5220/0006045902400248
In *Proceedings of the 8th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management (IC3K 2016) - Volume 1: KDIR*, pages 240-248
ISBN: 978-989-758-203-5

process take place between domain experts and technology experts; these two actor types collaborate to develop an understanding of what data is available and what knowledge could be gained from it, define a problem, design a solution and evaluate the results. Sometimes, however, the experts rely on non-expert actors for some crucial resource, which makes the interactions more complicated. Furthermore, there is now a new trend where the KDD process is driven by the needs of a non-expert actor, effectively reversing the traditional roles of experts and non-experts. The process model of KDD needs to be able to represent the dynamics of all these different collaborative relationships that the practice of KDD may entail.

This paper presents a conceptual model to serve as the basis of understanding the actors of the KDD process and the collaborative interactions among them. Besides human actors, the model includes technology as a special type of actor; traditionally, the role of technology in KDD has been to serve as a tool to be applied by technology experts, but advances in the field of artificial intelligence are making it increasingly feasible for autonomous software to carry out KDD tasks that have previously required a human actor. As the data deluge continues to multiply in volume, such software will be an essential part of any effort to make sense of it, and therefore its role in the KDD process must be defined by the process model just as the roles of the human actors are.

The remainder of this paper is organized as follows: Section 2 reviews literature on KDD process models and other relevant topics, showing how the new work presented in the paper advances beyond the current state of the art. Sections 3 through 5 present the main components of the proposed conceptual model, with Section 3 focusing on the actors, Section 4 on the interactions, and Section 5 on different ways of facilitating the interactions. Section 6 examines the relative strenghts and weaknesses of the proposed model and other KDD process models, and suggests a way of applying the model in practice. Section 7 discusses some noteworthy future perspectives, and Section 8 concludes the paper.

2 BACKGROUND

Following the publication of the KDD process model of (Fayyad et al., 1996), a number of derivative models were developed, as described in the survey of (Mariscal et al., 2010). In 2000, the CRISP-DM model (Wirth and Hipp, 2000) was published, intended as a standard model for the KDD process and incorporating elements or influences from several of the previously proposed models. This was followed by another set of derivative models, as well as another standardization effort, although ultimately it appears that the development of CRISP-DM 2.0 has ceased without ever yielding any concrete output (McCormick, 2007).

The idea of identifying the actors of the KDD process, their roles in the process and their relationships with one another can be found already in some early work such as the model proposed in (Brachman and Anand, 1996). The idea carried over to the Internet-enabled model of (Büchner et al., 1999) – a derivative designed specifically for knowledge discovery from Web data – and also to CRISP-DM, which acknowledges the importance of understanding the problem at hand from the business perspective of the customer as a prerequisite for examining it from the technical perspective of the KDD analyst.

Among relevant work published after CRISP-DM, particularly interesting are the RAMSYS methodology for remote collaborative KDD efforts (Moyle and Jorge, 2001), the knowledge exchange perspective of (Diamantini et al., 2006), and the knowledge fusion model of (Horeis and Sick, 2007). The ASUM-DM methodology (Haffar, 2015) of IBM is also relevant, being an augmented version of CRISP-DM. However, the first process model that properly acknowledges the role of non-expert actors in the KDD process was proposed by the author (Tuovinen, 2014).

The author's model builds upon previous research on issues related to the involvement of non-experts in KDD; for instance, there was already a significant body of work on dealing with the potential ethical problems arising from the use of personal data in KDD. However, such issues were previously always considered in isolation, as technical problems rather than something to be accounted for on the level of the KDD process. The author's model, in contrast, adopts the perspective that non-experts should be treated by the process model as actors with important contributions that the process should facilitate, and as stakeholders with legitimate expectations that the process should aim to satisfy.

In this paper, the author's previously published work on modeling the KDD process in terms of actors and interactions is extended in the following ways:

- In addition to the four actor types inherited from previous work, the paper defines four interaction types that, together with the actor types, can be used to construct different types of KDD processes.
- Using these basic elements, the paper presents a more detailed and formal account of three previously identified KDD process types, and addition-

ally of one type not covered by previous work, where the process is driven by a non-expert actor.

- Having identified and described the four different process types, the paper discusses possible ways of supporting them by using technology to facilitate key interactions.

These extensions constitute the novel scientific contribution of the paper.

3 TYPES OF ACTORS

Following the lead of previous research, we begin by identifying three principal types of human actors in the KDD process: domain experts, technology experts and non-experts. Additionally, interaction with computing technology is an essential part of any KDD process, so we shall consider that the fourth type of actor. We can now visually represent the relationships between these different actor categories using a triangle, with the human actors at the vertices and technology at the center, as shown in Figure 1.

The roles played by the actors vary depending on which specific type of KDD process we are examining. Different types of processes can be characterized based on which of the three human actor types are mainly involved and which of them interact with one another; thus we can visualize KDD process types by highlighting different sections of the triangle as in Figure 2. Here we identify four different process types, which we shall now look at in more detail.

The standard version of the KDD process is the one where only the base of the triangle, representing the collaborative relationship between domain experts and technology experts, is involved (Figure 2(a)). The relationship, as already outlined in the introduction, consists of interactions for specifying objectives, developing a solution and evaluating the outcome; interactions with technology are required to execute the necessary computations and examine their results. This process type is already quite well covered by existing process models such as CRISP-DM.

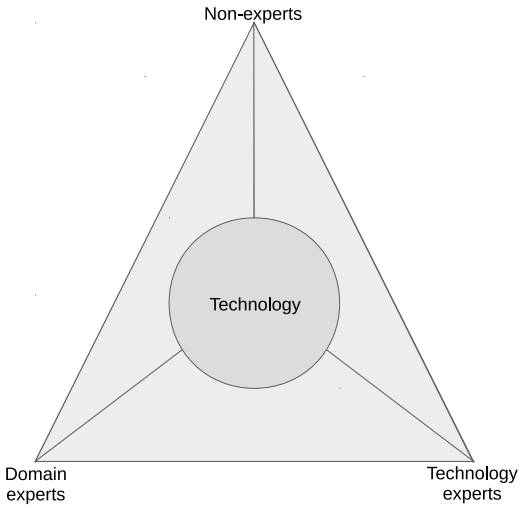
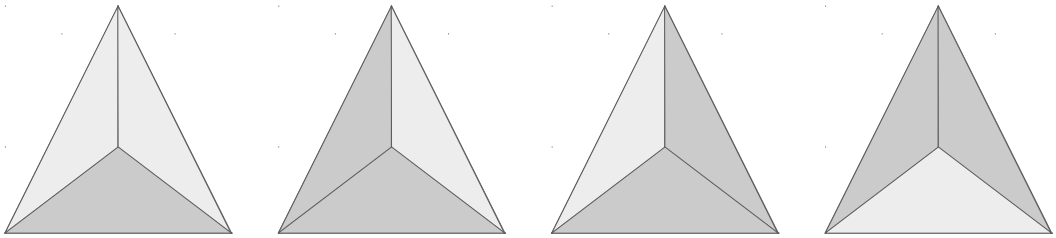


Figure 1: The four main types of actors in the KDD process. Technology is both a mediator of interactions among the three types of human actors (vertices of the triangle) and an actor in its own right.

Involving the third vertex of the triangle, representing non-expert human actors, in different ways yields different variations or special cases of the standard KDD process. Assuming that the process is still driven by the expert actors, there are two ways in which the non-experts can be involved: they can be a source of data that will be analyzed in the KDD effort, or they can provide computing resources to be used in the effort for data analysis. In the former case, their main interaction is with the domain experts, with whom they agree on the terms under which their data may be used (Figure 2(b)); in the latter case, a form of KDD known as volunteer computing, the main interaction is with the technology experts, who take care of the logistics of recruiting volunteers and distributing tasks (Figure 2(c)).

In the volunteer computing case, the technology actor has a role that differs somewhat from its role in the standard KDD process. The data processing in this case is done on the computers of the volunteers, and to coordinate this sub-process, there needs to be a system that divides the data to be analyzed into partitions, sends them out to be processed and collates the results. Thus, in addition to its usual role, technology acts here in a mediating role between the expert and non-expert human actors. Perhaps the most famous example of volunteer computing is the SETI@home project (Korpela et al., 2015), but there are many others; in a variation, it is the volunteers themselves and not their computers who perform the analysis, e.g. by classifying galaxy images (Lintott et al., 2008).

In both the above cases, it is important to ensure that the rights and legitimate expectations of the non-expert actors are respected. Unlike the experts, the non-experts may have no immediate interest in the outcome of the KDD effort, but they do typically have interests of their own at stake; the right to privacy, in particular, may be threatened when KDD is done on personal data, and there is a substantial body of research on how violating it can be avoided without compromising the objectives of the KDD effort (Shah and Gulati, 2015; Terzi et al., 2015). Volunteer computing has not been studied as much from this perspective, but it also has some potential ethical issues to be addressed (Tuovinen and Röning, 2009).



(a) Standard KDD process (b) KDD using personal data (c) KDD using volunteer computing (d) KDD driven by a non-expert actor

Figure 2: KDD process types illustrated. A shaded triangle section signifies that the corresponding actors are involved in the process and interact with one another. Refer to Figure 1 for a mapping of triangle vertices to actors.

It is also possible for non-expert actors to take charge of analyzing their own data (Figure 2(d)); this is happening, for example, with the quantified self movement, where some knowledgeable individuals have begun to use various wearable sensor devices and other data sources as means to improve their personal health and well-being (Swan, 2013). In this case, the non-expert relies on the experts' contributions rather than the other way around, and the interactions among them may be more indirect, with the non-expert e.g. consulting sources written by the experts instead of collaborating with them directly. Also, the technology actor needs to partially assume the role of technology expert by assisting the non-expert in the construction of KDD workflows and the evaluation of results.

It is clear from looking at these different variations of the KDD process that introducing the category of non-expert human actors creates a considerable amount of diversity in the roles and relationships of all the actors. Since established KDD process models do not account for the non-expert category, they cannot adequately represent all this diversity, and therefore cannot fully support the planning and execution of KDD processes that differ from the standard one. This point becomes even clearer in the next section, which views the KDD process as a sequence of interactions and compares how the sequence plays out in different process types.

4 TYPES OF INTERACTIONS

In order to be able to build a conceptual model of the KDD process in terms of interactions among the actors identified in the previous section, we first need to define a set of interaction types that we can use to represent the dynamics of the process. Based on the interactions that take place between domain experts and technology experts in the standard version of the process, we can identify the following four basic types:

- *Negotiation:* the actors establish the terms of their collaboration, agreeing on what each partner will contribute and on what conditions.
- *Analysis:* the actors establish a shared understanding of a domain of mutual interest to enable them to define a further course of action
- *Assignment:* the actors agree on a specific task to be carried out by some of them to achieve an established objective
- *Delivery:* the actors who carried out a task hand over the result to those expecting them and help them understand it.

Besides interactions between human actors, assignment and delivery can be used to represent the interactions of technology experts with technology as they set up and execute computations and examine the results. Thus, using these four interaction types, we can visualize the standard KDD process using sequence diagram notation, as shown in Figure 3.

The different variations of the standard process discussed in Section 3 can similarly be depicted as sequences of negotiation, analysis, assignment and delivery interactions. Figure 4 shows extracts from the sequence diagrams for the variations, focusing on those parts of the process where they differ notably from the standard case. In the case where non-expert actors are employed as a source of personal data, the most significant difference is that once the experts have completed the analysis interaction and thus established their data requirements, interactions with the non-experts are required to obtain the data (Figure 4(a)). Once the experts have gained access to the data, the process can continue in the standard manner.

In the volunteer computing case, where the non-expert actors are employed as providers of computing resources, the main difference is in how the computations are done (Figure 4(b)). Whereas in the standard version of the KDD process the technology experts interact with technology for this purpose, in volunteer computing both the experts and the non-experts interact with technology, and additionally there is a large

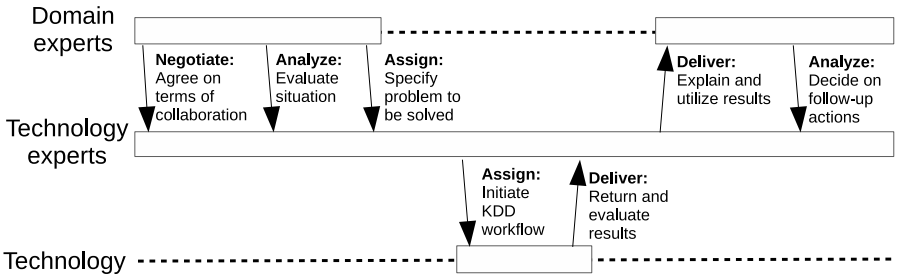


Figure 3: Depiction of the standard version of the KDD process as a sequence of interactions among domain experts, technology experts and technology.

number of interactions between individual technology actors - the computing server that distributes the jobs and the clients that process them. Not shown in the figure are the community interactions characteristic of volunteer computing efforts, where, for instance, the experts share information on the progress of the effort with the volunteers.

Finally, in the case where the KDD process is driven by a non-expert actor, the entire process differs considerably from the standard version, with the role of technology substantially expanded and the roles of human experts diminished accordingly (Figure 4(c)). To compensate for the non-expert's lack of essential domain knowledge and technical skills, it is especially important in this case for the interactions to be adequately supported, so that the non-expert can obtain the required knowledge and apply it to set up a KDD workflow. However, also in the other three cases supporting the process is largely a matter of supporting the interactions.

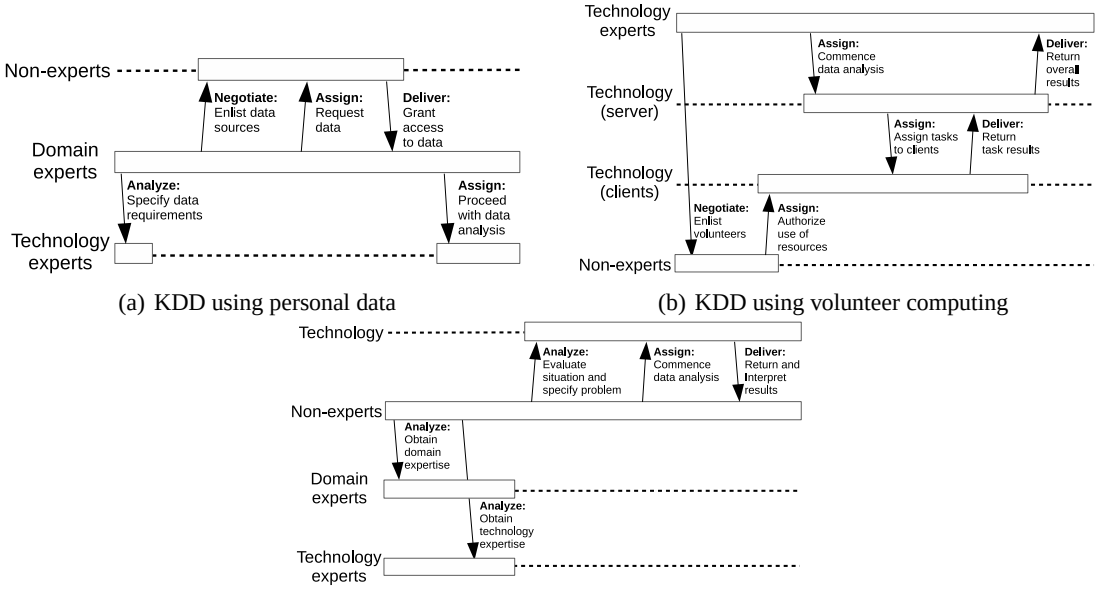
Supporting the various types of interactions that the KDD process in its various incarnations may involve is a vast, complex and multidisciplinary issue that cannot be addressed here in much detail. However, if we limit our discussion to a technological perspective, we can discern two principal types of support that computing technology can provide: we can use computers to facilitate interactions between human actors, or the computers themselves can interact with humans (or other computers) in more advanced ways. The next section takes a closer look at these possibilities.

5 FACILITATING THE INTERACTIONS

The most obvious way in which technology can facilitate collaborative interactions among the actors of the KDD process is by providing them with a communication channel. A wide range of options is already available for this purpose, from basic text-based channels such as email and instant messaging services to more advanced teleconferencing solutions with features such as live audio/video and screen sharing. However, while undoubtedly useful, well-established and generic technologies such as these are not particularly interesting in this context; we are more interested in facilitating communication in the KDD process specifically, and especially in supporting interactions involving non-expert actors, which are not as well understood as the interactions of the standard process.

One possible bottleneck in interactions involving non-experts is setting up the interaction: before collaboration can take place, the collaborators need to be introduced to one another somehow. Pioneering work addressing this problem has been done in the area of volunteer computing, where there are frameworks such as BOINC (Anderson, 2004) that make it simpler both for experts to set up projects and for non-experts to contribute to them. Various crowdsourcing platforms in general are relevant here in that they establish a convenient way for people in need of a service to find and negotiate with people who are able and willing to provide the service, be it data processing or, for example, transport as in the case of Uber. A similar approach would conceivably work in other forms of KDD as well, with a special software platform bringing the actors together and providing them with easy-to-use tools for specifying what they require and what they offer in exchange for it.

In terms of the four interaction types defined in the previous section, the platform outlined above would facilitate negotiation, assignment and delivery interactions among human actors, but not analysis interactions. In most versions of the KDD process, analysis is where the domain and technology experts interact to establish a shared understanding of the problem to be solved and the data and methods available for tackling it. This interaction cannot be modeled as an exchange of service and compensation; it requires closer collaboration among the actors, and therefore



(a) KDD using personal data

(b) KDD using volunteer computing

(c) KDD driven by a non-expert actor

Figure 4: Variations of the standard interaction sequence of the KDD process.

facilitating it requires a different approach, one that acknowledges that the result of the analysis is constructed not by one actor for the benefit of another but by both actors together.

Again, there are general-purpose solutions for collaborative work that are technically relevant but not particularly interesting; for instance, cloud services such as Google Docs and Office Online, which enable real-time collaborative editing of shared documents, could be used to develop artifacts representing the outcome of the analysis. In fact, collaborative editing is conceivably a viable paradigm for the facilitation of analysis interactions in the KDD process, but the format of the artifacts and the tools provided by the editor would have to be much more specialized than those found in generic office software suites. For example, by enabling the actors to specify available datasets, represent their internal semantics and identify potentially important relationships among them, the editor could help the actors understand what they could accomplish using the available data and which datasets they would need to combine to solve a given KDD problem.

So far, we have only considered the facilitation of interactions between human actors. KDD having been traditionally subsumed under computer science, it is probably true that interactions with technology are currently better understood than interactions between humans, which require a more multidisciplinary approach. Nevertheless, interacting with technology is far from trivial, and the importance of supporting these interactions effectively will increase as the data deluge gets more and more difficult to control, so we should take here some time to look at how they could be facilitated better.

In the standard version of the KDD process, human actors interact with technology via assignment and delivery interactions where the role of the technology actor is limited to executing the computations specified by the human actor and returning the results. Both the assignment and the delivery can be made more collaborative to improve this sub-process. To facilitate assignment, the specification of KDD workflows can be partially automated; there is already a substantial body of interesting research in this area, as demonstrated by a 2013 survey of KDD software systems capable of providing at least some level of intelligent assistance to the user (Serban et al., 2013). As one would expect, the ideal intelligent assistant does not yet exist: the systems reviewed are generally limited in some way, and those that are not require the user to be experienced enough to configure the algorithms and compose the workflow without assistance. The paper notes the limitations and proceeds to offer some future directions toward more advanced intelligent assistants, including the observation that they should be based on a collaborative model of KDD.

To facilitate delivery, new interactive visualization techniques can be developed to make it easier for the human actor to grasp the significance of the results. The need for closer integration of information visualization with KDD has given rise to the research field known as visual analytics. Several state-of-the-art surveys of this field have been carried out in recent

years; academic research is reviewed in (Sun et al., 2013) and commercial software systems in (Zhang et al., 2012). (Holzinger, 2013) also reviews a considerable amount of relevant work, advocating the integration of KDD with human-computer interaction in general. An interesting example of a technology that may prove useful for this purpose in the future is provided by (Donalek et al., 2014), where the possibility of using virtual reality platforms for information visualization is explored.

In summary, there are two principal ways in which new technology can help facilitate interactions in the KDD process:

- Interactions among human actors can be facilitated by matching potential collaborators and providing them with collaboration tools designed specifically for KDD tasks.
- Interactions between human actors and technology can be facilitated by developing intelligent software capable of taking on a more active role in the execution of KDD processes.

Although the role of technology in the KDD process would be somewhat expanded by these developments, negotiation and analysis interactions would still take place exclusively between human actors, with technology acting only as a facilitator. Section 7 discusses future developments that would transform the process in a more fundamental manner.

6 NOVELTY AND UTILITY

The main advantage of the proposed model over existing KDD process models is that it presents a broader view of the KDD process by including the category of non-expert actors. As a consequence, the proposed model is more suitable for representing process flow in variants of the KDD process where non-experts play a substantial part. The “quantified self” variant, in particular, is difficult to reconcile with established models, which are built on the assumption that the process is driven by expert actors.

Having a broader perspective on who the participants of the KDD process are also results in a broader perspective on what the objectives of the process are. In the traditional view of KDD, the purpose of the process is to extract value from data for the owner of the data, but this is just one aspect of the big picture, representing the point of view of one participant; the more participants there are, the more purposes there are, and some of the purposes may conflict with one another. In such cases it is essential to steer the process such that the nature of the conflict is identified and measures to resolve it are taken.

In its present state, the proposed model is mainly a conceptual one, which limits its utility as a practical tool for the management of KDD efforts. In comparison with, for example, CRISP-DM, the model lacks an overarching task structure and detailed guidance for the completion of each task. Some amount of structure and guidance is provided by the classification of interactions and the description of KDD process types as interaction sequences, but there is a considerable amount of practical verification and refinement to be done here.

In fact, probably the most effective way to apply the proposed model in its current form is to adopt the task structure of CRISP-DM – or some other established KDD process model – and use the interaction model to complement it. For example, the top-level task structure of CRISP-DM consists of six phases: business understanding, data understanding, data preparation, modeling, evaluation and deployment. If we substitute “business understanding” with “domain understanding”, the phases are completely generic and not in any way dependent on who the process actors are.

We can thus conceive a meta-process for combining the two models, where the first step is to identify the actors taking part in the KDD process. It is then possible to map each phase of the standard process model to the interactions among the actors that take place in each phase. Now, for guidance on how to execute the process, we can refer to both the documentation of the standard model and the ideas discussed here on how the interactions can be facilitated.

7 DISCUSSION

In KDD efforts driven by non-expert actors, it is necessary for the technology actor to assume a more active role in analysis interactions than it does in more traditional versions of the process. Currently there is software available that can do this in a limited manner, using preconfigured algorithms to extract knowledge from predefined data sources (e.g. a specific type of activity monitor). If the non-expert wishes to transcend such limitations, they need to learn to use a more generic KDD software suite, effectively becoming a technology expert as a result.

For us to be able to take a more decisive step toward this new type of KDD process, the technology actor needs to become more intelligent so that it can, in effect, play the role of both domain expert and technology expert and thus assist the non-expert human actor with the task of turning their requirements into

a specification for a KDD solution. The construction of the solution and the interpretation of the results – human-technology interactions in expert-driven KDD processes – would be handled autonomously by the technology actor. The role of the human actor in this case would be to collect the data, specify what they want to discover from it, and, having received the results, decide on next steps.

Clearly the ability of the technology actor to contribute to analysis interactions as an active participant would also substantially change the nature of KDD processes driven by expert human actors. With the technology actor taking over some of the responsibilities of the human actors, the latter could spend more of their time on higher-level analysis and planning activities. They would thus be able to address KDD problems on a higher level of abstraction, allowing them to tackle more complex problems and making the data deluge more manageable.

Taking yet a step further, we can consider the possibility of collaboration among technology actors. This would involve autonomously operating units of KDD software interacting with one another by requesting and providing services; thus, for example, a technology actor engaging in an analysis interaction with a human actor could enlist the help of another technology actor capable of offering special expertise relevant to the problem at hand. To a certain extent, this is already happening in various schemes for agent-based KDD such as those mentioned in (Klusch et al., 2003), but so far, the role of the agents has been subject to the same limitations as the role of technology in the KDD process in general, the agent-based approach mainly serving as a strategy for distributing computations.

In the future, we can expect KDD agents to even have the ability to engage in negotiation-type interactions with one another and with human actors. Such agents would serve as autonomous proxies for their owners, negotiating and performing the exchange of services and compensation without requiring human intervention as long as the terms offered are within the limits of their authorization. The benefits of agent-based KDD, particularly those pertaining to distribution of computational operations, would thus be expanded to cover a wider range of KDD tasks, and the range of options available for managing the data deluge would be similarly expanded.

In terms of the triangle diagrams of Section 3, the logical next step is a KDD process where all the vertices of the triangle are equally active, with each individual KDD effort having the optimal combination of actors as determined by the requirements of the problem at hand and the availability of resources. With a sufficiently powerful remote collaboration platform, geographical distribution of the actors is not a problem; what may become one, however, is coordination of the interactions among the actors, as there is no upper limit to the complexity of the KDD processes that can be constructed in this way. Designing a process model capable of coping with this complexity is necessary for us to be able to cope with the increasing complexity of KDD problems.

8 CONCLUSION

Digital data is potentially a highly valuable resource for a wide range of users, from individuals to large organizations. However, realizing this potential depends on finding the most appropriate human and technological resources for the task and using them effectively. This, in turn, involves an intricate web of collaborative interactions that the established KDD process model is not very well equipped to represent or support. The model particularly fails to account for the expanding roles of computers and non-expert humans in the process, both of which are necessitated by the new challenges arising from the big data phenomenon.

In this paper we presented a conceptual model intended to pave the way for a more detailed KDD process model that represents the collaborative aspect of the process better than the established model. The conceptual model describes the basic actor and interaction types that occur in KDD and shows that several different KDD process types can be derived from these. Practitioners of KDD can use the model to identify the actors involved in a given KDD effort, the role of each actor and the nature of the interactions between each pair of actors; this knowledge is required for the practitioner to be able to support the process, which is largely a matter of facilitating the interactions effectively.

ACKNOWLEDGEMENTS

The work presented here extends the author's dissertation; the author would like to thank Prof. Alan F. Smeaton for discussions that inspired this follow-up.

REFERENCES

Anderson, D. P. (2004). BOINC: A system for public-resource computing and storage. In *Proceedings of*

the Fifth IEEE/ACM International Workshop on Grid Computing, pages 4–10.

Brachman, R. J. and Anand, T. (1996). The process of knowledge discovery in databases. In *Advances in Knowledge Discovery and Data Mining*, pages 37–57. American Association for Artificial Intelligence.

Büchner, A. G., Mulvenna, M. D., Anand, S. S., and Hughes, J. G. (1999). An internet-enabled knowledge discovery process. In *Proceedings of the 9th International Database Conference*, pages 13–27.

Diamantini, C., Potena, D., Domenico, and Smari, W. W. (2006). Collaborative knowledge discovery in databases: A knowledge exchange perspective. In *Proceedings of the AAAI Fall Symposium on Semantic Web for Collaborative Knowledge Acquisition*, pages 24–31.

Donalek, C., Djorgovski, S. G., Cioc, A., Wang, A., Zhang, J., Lawler, E., Yeh, S., Mahabal, A., Graham, M., Drake, A., Davidoff, S., Norris, J. S., and Longo, G. (2014). Immersive and collaborative data visualization using virtual reality platforms. In *Proceedings of the 2014 IEEE International Conference on Big Data*, pages 609–614.

Fayyad, U., Piatetsky-Shapiro, G., and Smyth, P. (1996). The KDD process for extracting useful knowledge from volumes of data. *Communications of the ACM*, 39(11):27–34.

Federal Big Data Commission (2012). Demystifying big data: A practical guide to transforming the business of government. Technical report, TechAmerica Foundation.

Haffar, J. (2015). Have you seen ASUM-DM? Blog entry, retrieved 16 Sep, 2016. https://developer.ibm.com/predictiveanalytics/2015/10/16/have-you-seen-asum-dm/.

Holzinger, A. (2013). Human-computer interaction and knowledge discovery (HCI-KDD): What is the benefit of bringing those two fields to work together? In *Availability, Reliability, and Security in Information Systems and HCI: Proceedings of the International Cross-Domain Conference and Workshop*, pages 319–328.

Horeis, T. and Sick, B. (2007). Collaborative knowledge discovery & data mining: From knowledge to experience. In *Proceedings of the 2007 IEEE Symposium on Computational Intelligence and Data Mining*, pages 421–428.

Klusch, M., Lodi, S., and Moro, G. (2003). Agent-based distributed data mining: The KDEC scheme. In Klusch, M., Bergamaschi, S., Edwards, P., and Petta, P., editors, *Intelligent Information Agents: The AgentLink Perspective*, pages 104–122. Springer Berlin Heidelberg.

Korpela, E. J., Siemion, A. P. V., Werthimer, D., Lebofsky, M., Cobb, J., Croft, S., and Anderson, D. (2015). The next phases of SETI@home. In *Proceedings of SPIE 9606, Instruments, Methods, and Missions for Astrobiology XVII*.

Lintott, C. J., Schawinski, K., Slosar, A., Land, K., Bamford, S., Thomas, D., Raddick, M. J., Nichol, R. C., Szalay, A., Andreescu, D., Murray, P., and Vandenberg, J. (2008). Galaxy Zoo: Morphologies derived from visual inspection of galaxies from the Sloan Digital Sky Survey. *Monthly Notices of the Royal Astronomical Society*, 389(3):1179–1189.

Mariscal, G., Marbán, O., and Fernández, C. (2010). A survey of data mining and knowledge discovery process models and methodologies. *The Knowledge Engineering Review*, 25(2):137–166.

McCormick, K. (2007). CRISP-DM 2.0. Blog entry, retrieved 16 Sep, 2016. http://keithmccormick.com/crisp-dm-20/.

Moyle, S. and Jorge, A. (2001). RAMSYS - a methodology for supporting rapid remote collaborative data mining projects. In *ECML/PKKD01 Workshop on Integrating Aspects of Data Mining, Decision Support and Meta-Learning*, pages 20–31.

Serban, F., Vanschoren, J., Kietz, J.-U., and Bernstein, A. (2013). A survey of intelligent assistants for data analysis. *ACM Computing Surveys*, 45(3):article 31.

Shah, A. and Gulati, R. (2015). Contemporary trends in privacy preserving collaborative data mining - a survey. In *Proceedings of the 2015 International Conference on Electrical, Electronics, Signals, Communication and Optimization.*

Sun, G.-D., Liang, R.-H., and Liu, S.-X. (2013). A survey of visual analytics techniques and applications: State-of-the-art research and future challenges. *Journal of Computer Science and Technology*, 28(5):852–867.

Swan, M. (2013). The quantified self: Fundamental disruption in big data science and biological discovery. *Big Data*, 1(2):85–99.

Terzi, D. S., Terzi, R., and Sagiroglu, S. (2015). A survey on security and privacy issues in big data. In *Proceedings of the 10th International Conference for Internet Technology and Secured Transactions*, pages 202–207.

Tuovinen, L. (2014). *From machine learning to learning with machines: Remodeling the knowledge discovery process*. PhD thesis. University of Oulu, Finland.

Tuovinen, L. and Röning, J. (2009). Everybody wins: Challenges and promises of knowledge discovery through volunteer computing. In *Proceedings of the 8th International Conference on Computer Ethics: Philosophical Enquiry*, pages 821–842.

Wirth, R. and Hipp, J. (2000). CRISP-DM: Towards a standard process model for data mining. In *Proceedings of the 4th International Conference on the Practical Applications of Knowledge Discovery and Data Mining*, pages 29–39.

Zhang, L., Stoffel, A., Behrisch, M., Mittelstadt, S., Schreck, T., Pompl, R., Weber, S., Last, H., and Keim, D. (2012). Visual analytics for the big data era - a comparative review of state-of-the-art commercial systems. In *Proceedings of the 2012 IEEE Conference on Visual Analytics Science and Technology*, pages 173–182.