Issues in Virtual Database Support for Decentralized Knowledge Discovery

Giuseppe Psaila, Davide Brugali
Università degli Studi di Bergamo Facoltà di Ingegneria Viale Marconi 5
I-24044 Dalmine (BG), Italy
e-mail: psaila@unibg.it e-mail: brugali@unibg.it

Abstract: Knowledge discovery processes require powerful computational resources, and specific expertise to extract knowledge from large amounts of data. Data, resources, and expertise are now available on the Internet. Thus, decentralization of knowledge discovery processes seems a viable solution. However, effective use of the Internet calls for technologies which allow distributed resources to be flexibly combined, and their activities coordinated.

In a seminal paper, we introduced the concept of Decentralized Knowledge Discovery. In this paper, we discuss issues concerning the decentralized execution of knowledge discovery activities. We show that data and computational resources can be assembled together through the Internet to build a virtual database for Decentralized Knowledge Discovery. In particular, we point out which issues are relevant (and must be investigated) for building such systems.

1. Introduction

The Internet connects people, resources and activities. It facilitates the exchange of information and supports the cooperative work of managers, buyers, sellers, analysts, engineers, technical operators, and workers. Thanks to its features, Internet opens new ways to the exchange of information and supports the cooperative work of

- This is the case of knowledge discovery and data mining problems, whose solutions require to elaborate large amounts of data from different sources, employing experts with competence in a variety of highly specific fields, and using powerful computational resources which support the execution of heavy computational tasks, such as data filtering, classification, clustering, and pattern extraction. However, in many application contexts knowledge discovery activities require the integration of several data sets (owned by different data holders) to be effective; for example in a supply chain data are owned by each enterprise involved in the chain, so that complete information about the production cycle are fragmented into several different databases. In a decentralized scenario, data holders may share their data sets, in order to build a virtual database; this can be exploited by experts to perform knowledge discovery activities, to find out useful information that cannot come out from within single data sets (for instance, enterprises involved in the same supply chain might share data to improve the overall quality of the process/products). In a seminal paper [13], we introduced the concept of Decentralized Knowledge Discovery. In this paper, we move on from this concept, and analyze the problem of building a virtual database system supporting decentralized knowledge discovery from several points of view. In particular, we discuss which are the basic elements that constitute a Knowledge Discovery Process (KDP) when this is viewed from a decentralized perspective. The results of this analysis will bring us to redefine knowledge discovery activities for decentralized environments. Furthermore, we will analyze technical issues related with the development of virtual database systems for Decentralized Knowledge Discovery. In the paper, we will also discuss how decentralized knowledge discovery processes and related systems might help enterprises and organizations, by making possible integrated analysis over dispersed data, or by allowing the cooperation of several knowledge discovery teams in a virtual and common environment.

The paper is organized as follows. Section 2 recalls the fundamental concepts of a KDP highlighting the roles of key actors and activities. Section 3 classifies typical application scenarios for decentralized KDPs. Section 4 revises the traditional KDP, while relevant mobility characteristics of a decentralized KDP are identified in Section 5. Section 6 introduces issues concerning workflow support for decentralized knowledge discovery and implementation issues for systems supporting the discussed concepts. Finally, Section 7 discusses previous related work, while Section 8 draws the relevant conclusions.

2. Knowledge Discovery Process

In this section, we give a general description of the Knowledge Discovery Process. This is necessary to understand which are the activities that are generally performed to analyze data and discover knowledge from within the data.

Once described the process, we figure out the actors involved in the process. This step is extremely useful, since it opens the way to understand the current and the next generation application scenarios.

2.1 The Classical Process

Activities concerning knowledge discovery can be really various, in that several algorithms and tools for data cleaning and data analysis may be necessary. However, in the past years it has been argued that all these activities can be assigned to specific categories, that constitute the so called Knowledge Discovery Process (KDP) (proposed in [5]). In this process, the user (the analyst) iterates through several steps, each one devoted to a well precise task. They are the following.

1. Comprehension of the context. At the beginning, it is necessary to identify some fundamental aspects, such as the application domain, the basic knowledge from which the process starts and the goals of the end user.
2. Selection of a significant data set. Given the overall data set, a sub-set of the data is obtained selecting only the actual data on which the process is applied.
3. Preprocessing of the data. Very often, the data set to analyze requires to be preprocessed, in order to remove noise and deal with incomplete data, or to adapt them to the available data mining tools.
4. Simplification of the data. The representation of the data is modified, in order to meet the goals of the process. This is done by means of suitable transformation methods, which generally reduce the number of variables (attributes) and identify common features in the data.
5. Data Mining. Then, after data are preprocessed and simplified, it is necessary to understand the exact tasks to be performed by the data mining algorithm, then choose the proper data mining method and apply it to the data.
6. Interpretation. The results obtained by the data mining step are analyzed, based on the previously available knowledge. This analysis can lead to the decision of repeating the previous steps modifying the choices.
7. Formalization of the discovered knowledge. The knowledge obtained at the end of the process is formally documented.
8. in order that it can be reused or simply shown to the interested people. During this phase, it is also necessary to
to identify conflicts w.r.t. the knowledge previously available, and remove them.

The Data Mining phase of the KDP requires the choice and the application of a data mining technique; chosen the technique, the
user is asked to drive the process, typically by selecting the features of the data that must be investigated. If the user chooses the extraction of association rules, such features are the attributes whose values are associated by rules, the minimum support, etc. (see [2,11]).

However, it may be difficult to understand both the data mining technique to use and how to set the parameters that drive the data mining tool. This means that the process is intrinsically iterative: based on the results obtained in each phase, the user might decide to modify the hypothesis made in the previous steps, in order, for instance, to modify the data selection or the preprocessing phase to obtain a data set more suitable for the chosen data mining tools. Not only, even if the choices were made correctly, the discovered knowledge may suggest to analyze the data changing the hypotheses, in order to compare different discovered knowledge. Finally, the process might be repeated with the same hypotheses simply because the data are updated.

2.2 The Actors

The basic idea behind knowledge discovery is simple: a bunch of data sets is available; then, by applying knowledge discovery techniques, and more specifically data mining tools, it should be possible to obtain useful information from the data.

However, to achieve this goal is not simple, because several Actors are involved in this process.

- **Data Holders**. Data holders are those who actually hold the data to analyze. In a traditional context, the enterprise that collected data through its information system is a data holder.

- **End Users**. End users are those who wish to take advantage of the (unexpected) knowledge that can be discovered from within the data. In a traditional context, the enterprise management expects that the discovered knowledge is exploited either to make decisions, or to set up behavioral rules (for example, an insurance company may define the fare based on the knowledge emerged from within the historical data).

- **Computational Resources Holders**. Knowledge discovery tools and systems which are able to operate on large volumes of data are necessary, and typically require to be executed on specialized and well equipped computers, especially as far as data storage units and fast CPUs are concerned. Then a computational resources provider is the one that holds such hardware and software, and makes it available for the knowledge discovery process.

- **Analysts**. Computational resources and knowledge discovery tools are useless without skilled human resources that are able to exploit them in the analysis process. To do that, it may be necessary to exploit both experts in the specific application domain (for instance, experts in marketing) and experts in the conduction of knowledge discovery processes (these are the technicians which are able to interpret users’ requests to adopt the proper knowledge discovery techniques and tools).

If some of the above identified actors are missing, the knowledge discovery process cannot be conducted. In the following section, we show how the identified actors may interact in different decentralized application scenarios.

3. Decentralized Application Scenarios

**Supply Chain Scenario**. Very often, several enterprises are involved in a unique supply chain: each enterprise performs a specific part of the overall production process, providing its specific competence. The effect is that the overall quality of the process/end product may be difficult to monitor; in fact, since data about each part of the process are fragmented into several data sets belonging to different enterprises, it makes sense to perform knowledge discovery tasks devoted to find out causes of defects on the process/end product only if all the data sets are available. Thus, the enterprises involved in the supply chain might create a virtual database: they share their data sets which are relevant for the supply chain; a pool of experts might then work on these data sets in a remote way. As far as the actors involved in the KDP are concerned, this means the following.

- **Data Holders and End Users** are the individual geographical distributed research centers which collect information about specific case studies.

- **The End User** is the research network as a whole and possibly the entire research community. In effect, intermediate results of complex activities are shared, in order to evaluate them and make decisions on how to carry on the overall research activities.

- **Knowledge Discovery Tools, Computational Resources, and Analysts** are in-house resources of every research center. In effect, we can figure out that research centers are equipped with the necessary computational resources, if they are specialized in performing specific analysis activities.

**Network of Excellence Scenario**. A research network (e.g. in biology) is made up of a number of distributed research centers, each one specialized for competence and research directions. They share a number of case studies that need to be analyzed in order to verify some theoretical hypotheses (e.g. the origin of an epidemic disease). As far as the actors involved in the KDP are concerned, this means the following.

- **The Data Holders** are the individual geographical distributed research centers which collect information about specific case studies.

- **The End User** is the research network as a whole and possibly the entire research community. In effect, intermediate results of complex activities are shared, in order to evaluate them and make decisions on how to carry on the overall research activities.

- **The Knowledge Discovery Tools, Computational Resources, and Analysts** are in-house resources of every research center. In effect, we can figure out that research centers are equipped with the necessary computational resources, if they are specialized in performing specific analysis activities.

4. Revising the Knowledge Discovery Process

The decentralized application scenarios previously introduced put in evidence a new way to intend the knowledge discovery process. The general model presented in Section 2 does not consider the decentralization of actors and resources. Hence, it is necessary to adequately the process to the new scenarios.

**Activities**. At first, we revise the set of activities which constitute the KDP.

- **Data Gathering**. In the decentralized scenarios, it is necessary to identify source data sets involved in the process.

- **Selection**. The selection activity is not overlapped to the Data Gathering activity. In fact, the previous activity identifies data sets and builds the initial virtual database with them. In contrast, the selection activity chooses the ones on which a specific subtask of the KDP is focused, among all available data sets.
In fact, the KDP might be composed of several subtasks, each one working on a different subset of the collected data sets. In any case, the selection activity extends the virtual database.

**Preprocessing.** As in the classical KDP, the preprocessing phase is necessary to remove noise and incomplete data from the data sets selected in the selection activity. Observe that specialized tools may be adopted; not only, it may involve several people with different skills (for instance application domain experts and data cleaning experts). At the end, new preprocessed data sets are added to the virtual database.

**Simplification.** The simplification activity consists in simplifying and transforming the data set to analyze, in order to make it suitable for the chosen data mining tool. As in the preprocessing activity, this activity may involve people with different skills (the application domain expert can provide indications on how to choose the features), but the expert of data mining tasks has the keys of the activity. This activity as well adds new data sets to the virtual database.

**Data Mining.** In terms of computational load, this might be the most expensive phase, where sophisticated tools are exploited. Furthermore, in this activity it is necessary to define the parameters that drive the data mining tool, so that the models (patterns) generated by the tool are significant (in the sense that the models/patterns are sufficiently synthetic but at the same time provide useful knowledge about the analyzed data set); thus, models and patterns constitute new data sets that are added to the virtual database.

**Evaluation.** Once the data mining activity produced significant patterns (i.e., potentially useful knowledge), it is necessary to evaluate such results. The goal of this phase is simple: if the extracted models/patterns are not considered useful or they are considered not accurate, the reasons why such patterns are not adequate should suggest how to modify the parameters featuring the execution of the previous activities. Otherwise, the generated models/patterns are made available for next activities; they constitute pieces of knowledge for the overall process.

**Knowledge Integration.** In the decentralized scenarios, an activity focused on knowledge integration is fundamental. Indeed, pieces of knowledge may be separately discovered by members of different teams involved in the process, but in order to achieve a full comprehension of the overall studied phenomena it is necessary to coherently integrate these pieces of knowledge. The result is the knowledge base of the overall process (that can be updated several times during the process). The integrated knowledge is itself a new data set added to the virtual database.

**Knowledge Delivery.** Finally, the knowledge is delivered to end users. Observe that this is not a trivial task, since not necessarily the complete knowledge base is of interest for end users, e.g., decision makers. In effect, the goal of this activity is to select the portion of discovered knowledge that really helps end users: this means that specific groups of users are interested in specific portions. The proper delivery of the discovered knowledge, w.r.t. end user’s interests, is the goal of this activity. Consequently, new data sets describing the portion of discovered knowledge to deliver are added to the virtual database and retrieved by end users within the virtual database.

Observe that this activity may produce another result: it may happens that new, previously unexpected, needs or ideas come out, suggesting to perform other knowledge discovery activities.

**Meetings.** are another kind of activity, which are usually important in the KDP, since they are the occasion to discuss the results and to make decisions. In a centralized environment, they are not considered in the KDP, because can be organized in any moment and without particular constraints. However, in a decentralized environment the organization of meetings is not a trivial task: in fact, if people have to physically meet in the same place it might result very difficult to organize the meeting (cost of travel, appointments of each single person, etc.); virtual meetings are a better solution, for example allowed by a video-conference equipment.

Furthermore, the results of meetings (reports, decisions, etc.) might be significant to move on the KDP, therefore they must be explicitly considered as a (special) activity for the revised knowledge discovery process.

**Tasks and Sub-tasks.** A new concept should be added to the knowledge discovery process: the concept of task. A Task is a sequence of knowledge discovery activities, performed with a specific goal; a task can contain specific sub-task. Each task or sub-task has a supervisor, which is responsible to move on the assigned task. This way, the knowledge discovery process is partitioned into possibly parallel processes, that can be performed by different teams (a very interesting perspective suggested by the decentralized scenarios). At the end, the supervisor of the main task collects the results, integrates and delivers the knowledge, possibly creating new sub-tasks.

It is clear that a knowledge discovery process based on the revised model should be adequately supported by a system. In particular, it is evident that the process is a sequence of activities and tasks; activities are well defined, as far as the general goal of each activity is concerned, while how the single activity has to be carried on depends on the specific situation.

Anyway, it is also clear that activities involve many people with different roles; hence, the support provided by the system must enhance cooperation among people and the definition of sub-goals.

Finally, since knowledge discovery activities are better performed when knowledge is effectively shared or retrieved from past activities, the system should facilitate information circulation and allow some kind of knowledge management and reuse of previous results/activities/processes.

**Parallelism and Decentralization.** Decentralization implies, in some sense, parallelism. Consider, for instance, the Network of Excellence scenario: in this context, knowledge discovery activities are necessarily decentralized, since they are delegated to each research center participating to the network; this means that each center is responsible to carry on specific activities, even complex, and therefore better performed inside the center (this fact implies mobility of activities and tasks, see below). However, there is no need that decentralized activities are necessarily sequential, but often they can be executed in parallel (for instance on different hosts).

However, we think that an excessive degree of parallelism is not suitable for the KDP: in fact, although it is true that certain activities are parallel, the intrinsic nature of knowledge discovery is sequential, since at each step it is necessary to reason about results obtained by the previous steps. Therefore, we think that a KDP is a sequence of steps, where each step allows the parallel execution of activities and sub-asks; until all parallel activities and sub-tasks are terminated, the KDP cannot evolve.
5. Mobility of the KDP

Traditionally, Knowledge Discovery tools are built around a central knowledge base and focus on shared data and local processes. With the introduction of the Internet however, some processes can easily be shared across the organization and operate mostly on local data. In fact, the idea of virtual database we are discussing in this paper can be intended in the broadest sense: the virtual database also supports virtual processes, such that each piece of computation can be executed in a decentralized way. Consequently, issues concerning mobility must be considered, in order to reinterpret data mining and knowledge discovery tools.

In this section we identify the patterns of distribution that characterize a decentralized knowledge discovery process in terms of distributed interactions between its key elements, i.e. the data (both the initial data sets and the knowledge extracted from them), the activities (the basic building blocks of the KDP), and the tools (the software applications that automate the knowledge discovery and data mining processes).

A fundamental constraint exists among these elements: an activity can be executed on a given network host, if input data, software tools, and the necessary computational and human resources can be made available at that host.

Consequently, the concept of virtual database on which the idea of decentralized knowledge discovery is built must be intended in the broadest sense: the virtual database for knowledge discovery should support virtual data sets, and virtual knowledge discovery processes; consequently, mobility is central in such a system, and may concern both data and code mobility.

- **Data Mobility**: When a given activity requires heavy computational resources to be executed, the input data must be transferred to the host where those resources are available. Depending on the volume of the input data and the iterative interaction between the data and the KDP, two alternative options should be considered:
  - Transfer of the remote input data / on site processing of the local copy.
  - Data streaming of the remote input data set to the local processing activity.

The activity’s output data are possibly disseminated to remote hosts.

- **Activity Mobility**: In the Network of Excellence Scenario, all of the distributed research centers own the computational resources to execute the KDP. Thus, there is no need to move input data from their origin host. What should be transferred from host to host are the activities that need to process remote data. An activity is created by a Task Supervisor on a specific network host. If the activity needs input data from a remote host, it suspends its state, migrate to that host, processes the input data locally, updates its state and move back to the origin host or forward to the next host where other input data are available.

- **Tool Mobility**: Task Supervisors and Activity Managers might decide to use software tools that are not available on their host. In this case, specialized KD tools might be made available for download from other network hosts. This situation is typical in the Network of Excellence Scenario.

6. Workflow Support

The revised KDP is based on concepts that are typical of workflow models. In effect, we can imagine the KDP as a special kind of workflow, in which the sequence of activities and subtasks is dynamically built.

This idea is different w.r.t. the idea behind classical workflow models; these are useful to model predefined processes, that organizations or automated systems must follow. However, this view is not suitable for knowledge discovery tasks: the sequence of activities to perform strongly depends on the partial results, thus it must be defined dynamically.

For the sake of space, we cannot discuss this issue in details. Anyway, we try to figure out the basic workflow concepts and implementation issues that a system supporting decentralized KDPs should provide.

We refer to Figure 1, which reports a possible graphical representation for a sample KDP process modeled as a dynamic workflow. Symbols in the figure have the following meanings. Thin ovals denote start symbols for tasks and sub-tasks, while thick ovals denotes stop symbols. Solid-line rectangles represent knowledge discovery activities, while dashed-line rectangles denote sub-tasks (inside them, we can find again start and stop symbols, activities, etc.). Triangles denote data sets, which are generated by activities and sub-tasks; dotted oval denote groups of data sets. Finally, diamonds represent convergence symbols, in which parallel activities or sub-tasks are synchronized.

The sample workflow denotes an on-going knowledge discovery task, since the stop symbol is not present; this means that a user, with the role of task supervisor, may decide to define new activities and sub-tasks; for each of them, the task supervisor defines a set of requirements (to instruct people involved in the activity/sub-task to properly perform it), defines the set of input data sets, assigns activities and sub-tasks to the proper working team.

At the beginning, a sub-task is empty; it must be defined by the person of the working team to which the sub-task has been assigned, i.e. the sub-task manager; similarly, for activities as well it is necessary to define the activity manager, which has the responsibility to carry on the activity. During the execution of single activities, the working team may exploit any kind of knowledge discovery tool suitable for the specific type of activity.

Sub-tasks and activities may be executed in parallel. For instance, this is the case of the two sub-tasks reported in the figure. When they finished, each of them produced a pool of data sets. Then, the general task is synchronized (diamond symbol) and all the data sets produced by the sub-tasks are made available for the whole main task.

Finally, all the data sets generated by the sub-tasks are used by the Knowledge Integration activity, which is responsible to integrate pieces of knowledge discovered by the two independent sub-tasks, generating new data sets which may constitute a first result of the knowledge discovery task (i.e. knowledge).

**Implementation Issues.** The dynamic workflow model previously discussed gives a global overview of the knowledge discovery task, without taking into account its decentralized nature. This is correct from the point of view of the task supervisor, since it is necessary, for him/her, to have a global view of the process. However, the implementation of a system should strongly take into account this issue. In terms of system functionality, this means that several hosts, located in different sites and connected through the Internet, should be registered into the system; then, the system is naturally distributed (since composed of several hosts). On each host, specific knowledge discovery tools can be installed and made available to the overall network.

Activities and sub-tasks can be freely moved from one host to another, in order to be executed on the proper host (for example, the host on which a specific tool is installed). Furthermore, data sets can be moved as well, or simply linked and remotely accessed. For example, a sub-task may produce data sets which resides on the host on which the task was carried on; they are made available to the main task. Them we can guess that the main task supervisor might decide to perform the Knowledge Integration activity on a specific host; to avoid an excessive access to the Internet during this phase, data sets might by moved, or copied, to the host on which this activity is performed.

Then we can conclude that a system for decentralized knowledge discovery really implements a very general idea of virtual database. In effect, data sets should be accessed independently of the host on which they are located, and they might be possibly moved; tools might move over the host network to meet the data to analyze; finally, the system implements a database of knowledge discovery
processes, which should be intrinsically virtual, since tasks might be moved around the network.

7. Related Work

A long series of papers and research works characterized the field of data mining and knowledge discovery in the last decade. In particular, it is possible to identify two main research lines: development of efficient algorithms, integration of data mining tools and databases. The first research line (development of efficient algorithms) demonstrated [1] the feasibility of knowledge discovery on large volumes of data. Efficient algorithms for a large variety of problems were developed, in particular for association rule extraction (see [2,8]), classification and clustering (see [14]) and many others. The new trend in research about algorithms, is the development of stream algorithms; an example can be [6].

On the other side, several researchers are addressing the problem of integrating data mining techniques, knowledge discovery processes and databases. Several works addressed this topic, in particular from a language perspective; in [11,7] different query languages for data mining based on the SQL syntax are proposed: the common idea is to extend SQL with specific constructs, in order to make the user able to specify in a declarative form data mining statements over relational databases. The main advantage of this proposals is the fact that data, patterns and mining statements belong to the same framework, i.e. the relational framework, where usually data are stored. The work in [12] tried to stress this idea, showing that the relational database framework could be effectively used to host several SQL-like data mining operators, thus obtaining a relational database mining framework.

A further step on this direction was the definition of the concept of Inductive Database, i.e. a unifying framework for data mining; this idea has been introduced for the first time in [9], but the work [3] formally defined this notion for the first time.

The concept of workflow originated several years ago as evolution of the concept of long transactions, which can last for a long time and do not have the classical ACID property typical of short transactions. Nowadays, several workflow models and systems were developed. Workflow Management Coalition (WFMC) is the official organization that drives research and developments in the field. Fundamental concepts and useful information can be found at [15].

The role of the Internet in the dynamic matching of business problems and available solutions has been analyzed in the literature mainly from the point of view of the automated mediation support offered by agent-based systems. In [4] a multi-agent middleware software framework is presented. The framework allows the development of specific mediating applications that sustain the complex task of dynamic switching on the Internet. In [10] a multi-agent system is presented that supports the dynamic formation of Virtual Enterprises.

8. Conclusions

In this paper, we present the concept of Virtual Database for Decentralized Knowledge Discovery Process, an idea to improve the exploitation of knowledge discovery techniques both in old and in new application scenarios.

The basic idea is that the knowledge discovery process is obtained as composition of activities that involve several actors and exploits several, and distributed, resources. Technical solutions based on the integration of these resources through the internet, which exploits the concept of process mobility, may give significant improvements to the adoption of knowledge discovery tools. In fact, the concept relies on the availability of distributed computing technologies such as code mobility, semantic interoperability, and software agents.

We discussed several issues related to the introduced concept. In particular, their implementation can be a challenging activity: issues concerning e-services, XML, data and code mobility, data streaming, software agents should be explored. For example, the connection among the hosts composing the decentralized system may be obtained through interfaces and protocols inspired to e-services; clearly, XML can play a significant role to achieve interoperability. Data mining and data analysis tools may be based on data streaming techniques: this solution will allow to obtain a reduced mobility of data resources around the system (this research area is just at the beginning). Finally mobility of code and workflow activities might be obtained by means of mobile software agents, which, based on specific criteria (e.g. load balancing techniques, negotiation techniques, etc.), might automatically migrate to the host which should provide the better support (in terms of execution environments, computational power, etc.). The reader can easily understand that these are only a few research lines that can originate from this work.

References