

DATA MINING DISCUSSION GROUP REPORT

1 Introduction

The data mining discussion group comprised a diverse set of investigators, ranging from those focused on core data mining concepts and techniques to those focussed on applications of data mining. The group included one chemist, two artificial intelligence oriented researchers, one cooperative work researcher, and several researchers interested in image and video data mining, in addition to researchers focused on core database and data mining techniques.

The results of our deliberations are presented in three parts. First, we lay out the scope of the problem being addressed, its importance, and the state of current solutions. Second, based on the above observations, we suggest a number of possible new research directions, keeping in mind that this list is not exhaustive. Finally, we propose a funding model to encourage fruitful collaboration with application domain research, maximizing the fruits of interdisciplinary collaboration while encouraging advance in each discipline.

2 Status Check

2.1 Scope of Problem

The data mining task is a multistep process that comprises:

1. Data Acquisition
2. Data Cleaning
3. Feature Extraction, possibly including Data Reduction and Dimensionality Reduction.
4. Pattern Extraction and Discovery
5. Visualization
6. Evaluation of Results

Research has tended to focus on step 4, and perhaps rightly so, as it is a crucial step, likely to consume the most computational resources, and also the step most likely to benefit from scientific advances. In fact, techniques developed for this step have been proposed recently as an aid to performing some of the earlier steps, such as steps 2 and 3. Nonetheless, it is important to bear in mind that the entire process has to be supported if data mining is to be practically useful. As such, appropriate investigation into the other steps of this process should also be encouraged.

An orthogonal defining characteristic of data mining is that it applies to large data sets. While useful information can be gleaned from small data sets also, their small size permits the use of many different techniques, including human inspection, that just do not apply to large data sets. The challenge of size is central to the data mining task.

Finally, feedback is typically required to complete the data mining process successfully. Interaction between the human analyst and the data mining system could be human driven (e.g. data exploration) or system driven (e.g. relevance feedback). In either case, it is typically expected to be an iterative process that converges to the desired solution.

2.2 Current State

One can think of a four-level hierarchy of abstraction, as suggested by Fig. 1. At the most generic level are the concepts and principles of data mining. These are broadly applicable across a wide range of applications. The next level has data mining techniques, for instance, the A-priori algorithm for association-rule detection. These techniques too tend to be relatively broadly applicable. The third level comprises data mining tools. Whereas commercial vendors for such tools may proclaim broad applicability, it appears that there is still much technical progress possible in this regard. Finally, the fourth level has the data mining applications, which of necessity are application specific.

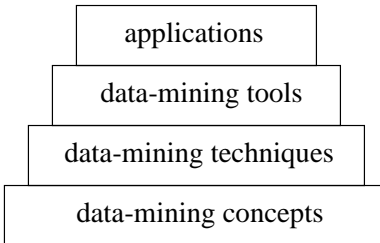


Figure 1: Hierarchy of data-mining issues

The central goal of data mining research is to address issues of scale and performance at the level of concept, technique, and tool. While applications can indeed drive research into developing appropriate new tools, and possibly techniques and even concepts, application-specific investigation is not the only way to make progress. There is clearly room for, and a need to encourage, development of new concepts, techniques, and even tools, independent of the application. A paucity of publicly available large, non-main-memory, data sets is currently a roadblock to testing the efficacy of new data mining techniques and algorithms.

Data mining has already a huge impact on society; numerous articles in the popular press loudly attest to this fact. In fact, if anything, popular media may have caused expectations to run so high in some circles that we run the risk of a backlash for failing to deliver. NSF funding has been a key enabler for the development of data mining technology so far. Additional funding is critical if the computer science community is to make a credible attempt to meet the current high expectations that society has for data mining.

3 Research Directions

We present below, in no particular order, some fruitful research directions identified by our group. This list is not exhaustive, and is not meant to suggest that other research directions are not worth pursuing.

- Mining stream data or online data mining. Develop techniques for extracting patterns from data in one pass, with only a limited amount of local storage available.
- Mining the dynamics of data: finding trends and detecting qualitative change points in sequence data.

- Mining distributed, heterogeneous, and/or non-integrated data. Problems here could range from data normalized into multiple tables within a single centralized database to data drawn from autonomous sources on the internet.
- Interactive, semiautomatic data mining: reusing discoveries to make more discoveries.
- Domain-appropriate modifications of tools and techniques. For example, can different algorithms be selected for the same task based upon data statistics? What would be material statistics for this purpose?
- Cleaning data – What is the model for “dirt”? How can we convince a domain scientist that we did not lose valuable information in the data cleaning process?

4 Recommendations

Data mining is a valuable area – there is much technical progress possible, and much benefit to be obtained in a variety of application domains. It is an ideal area in which to devote government funding, with a high likelihood of obtaining highly leveraged returns on this investment. Indeed, there is considerable application-driven interest in data mining, from a variety of domains, ranging from business to medicine to the natural sciences. While application-driven research can often be valuable, and interdisciplinary research can be a driver of research in each of the disciplines involved, it is important to ensure that core data-mining technology development get the support that it deserves. As such, we strongly recommend additional funding for data-mining work.

However, much data-mining work, when done in concert with application domains, is interdisciplinary. We have a proposal with respect to a new model for encouraging high-quality interdisciplinary research.

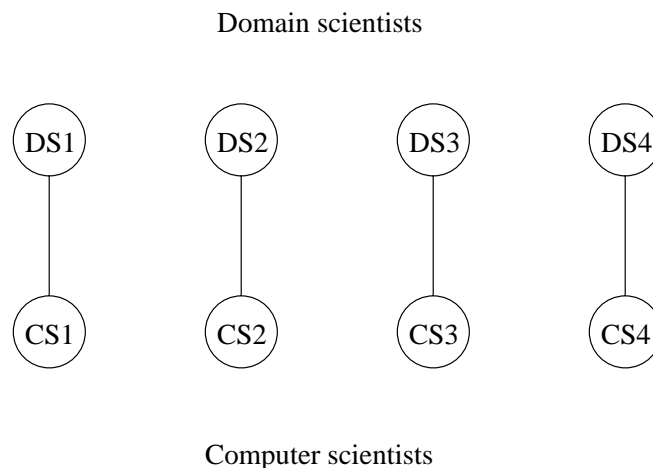


Figure 2: The inefficient “KDI” model for CS/domain interaction

The current model, followed for instance in the KDI funding, is shown in Fig. 2. There is a pairing of computer scientists investigating data mining issues and application domain experts. Of necessity, these pairings are limited by mutual acquaintance. There is no systemic assistance in developing such

partnerships. Worse, the partnerships that do develop are not necessarily the ones that are technically the most viable. More typically, partnerships are based solely on geographical proximity.

Partnerships formed in this way are unlikely to be optimal. For instance, when a domain scientist pairs up with a computer scientist colleague investigating a particular class of data mining techniques, the scientist may not have the computing knowledge to determine whether this class of techniques is appropriate for the particular problem domain. It may well be that alternative approaches, proposed by computer scientists elsewhere, would be more suitable. If true computer-science research is involved, then it may well be the case that no one involved in the project can predict in advance the most appropriate computational solution, or even if such a solution exists.

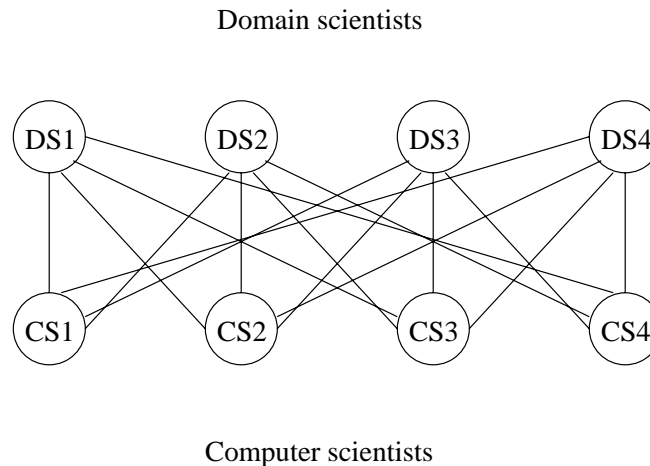


Figure 3: Many-many pairing offers new opportunities, but is too inefficient to manage

Rather than a one-to-one pairing, one would like to have a multiway connection of the sort depicted in Fig. 3. Each computer scientist could attempt to apply his or her new techniques to a variety of applications and find the ones most suitable. Similarly, each domain scientist could try out several different mining approaches and choose the ones most appropriate for their application. Unfortunately, there is no practical way to create such a rich network of connections.

Our proposal is that NSF serve as a “broker” as shown in Fig. 4. The idea is to create a “bank” of application-domain problem descriptions and data sets, and perhaps also a “bank” of data mining tools. Each scientist putting in a tool or an application problem into the bank makes a commitment to “support” it appropriately — helping others understand the problem description for example, or attending to bug reports in a tool. An extension to the Irvine machine-learning repository is one model of what we have in mind, should that effort become more focused on non-main-memory data sets and on incorporating data sets from scientists on demand.

Computer scientists should be funded for conducting research in data mining and creating and supporting tools that could be used by domain scientists. Domain scientists should be funded for research in their domains, and for creating and supporting problem definitions in the bank. This funding on each side can take place independently, and at volumes compatible with NSF allocations for the respective areas.

In addition, when a match is found between a domain scientist and a computer scientist, we believe that this pairing should be furthered by supplemental grant, perhaps to support a pair of students, one

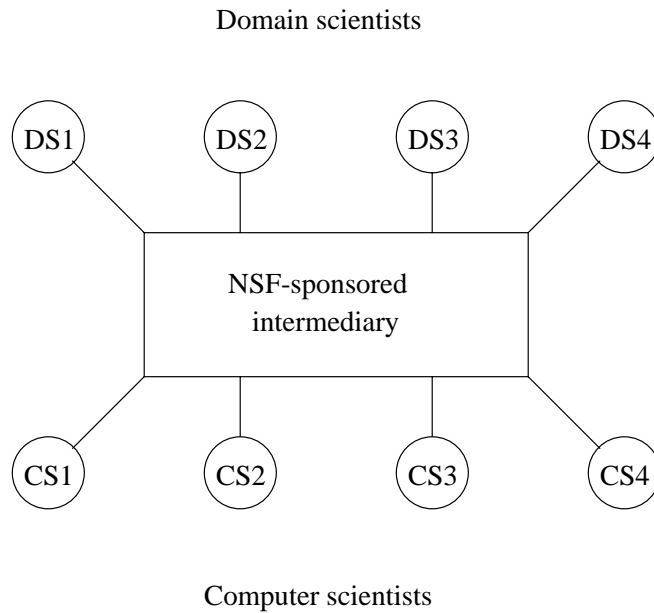


Figure 4: NSF should broker interactions between domain scientists and computer scientists to get the best data-mining techniques working on the best problems

from each side of the problem, collaborating on a solution. We propose that CISE take the leadership in providing this supplemental funding. Thus, interdisciplinary work would be encouraged, and rewarded, but the partnerships formed can be drawn from a wide, national pool, rather than each being limited to small circle of local acquaintances.