DATA MINING THROUGH INFORMATION ASSOCIATION: A Knowledge Discovery Tool for Materials Science

Krishna Rajan[1] and Mohammed Zaki[2]
Faculty of Information Technology
[1]Dept. of Materials Science and [2]Dept of Computer Science
Rensselaer Polytechnic Institute, Troy NY USA

ABSTRACT

The recent announcements of the discovery of materials exhibiting superconductivity (MgB_2 – and a conjugated polymer, regioregular poly(3-hexylthiophene))[1,2] serve to highlight the fact that materials discovery is still a process that is often governed by empiricism and accidental discoveries. While incremental progress is made in specific technological areas of interest, we need to have a means of exploring vast combinations of structure-property relationships combined with the appropriate search tools that can accelerate the discovery process. Once a "discovery" is made , the usual and important questions are how can the property can be improved, reproduced and the potential reasons for such behavior. An understanding of the mechanisms governing superconducting behavior for instance can provide guidelines for developing new materials and associated processing approaches to enhance properties. On the other hand, if an unexpected discovery is made, then classical theories have to be reexamined resulting in new theoretical formulations accompanied by experimental studies which establish guidelines for new materials development. In this paper we outline some of the basic operating principles of a tool known as "data mining" which when applied to appropriate datasets (both theoretically and experimentally derived) can be used to *anticipate* the existence of materials attributes.

INTRODUCTION

The connection between advances in materials and changes which define social and economic progress has always existed in human history. The "Stone Age", "Iron Age" and "Bronze Age" are all terms that we recognize as describing stages of human evolution in terms of technological "eras". It is no coincidence that the dawn of our present generation, which may be classified as the "Information or Computer Age" was initiated with critical discoveries and engineering developments in the field of semiconductors and other materials. Traditionally, computers have been used as storage mediums for large volumes of data and as tools for carrying out extensive numerical computations and simulations. Recently, computers have started to take a more active role in guiding the scientist through the research and discovery process with the help of data mining methods for automatic discovery of patterns in large volumes of data. The challenge is now to make these data mining methods ubiquitous and an integral part of the data collection and verification process. Materials Science offers a unique challenge in data mining due to the variety of data types, and their complex interconnections. During the material discovery process, there is a need to integrate multiple, heterogeneous databases to reach new and even unexpected conclusions as well as to use databases actively to design new processing strategies. This complex coupling of data models, data analysis methods and physical methods offer a unique computing challenge that has not yet been addressed sufficiently in both information technology research as well as materials science research.

KNOWLEDGE DISCOVERY in DATA BASES

Knowledge Discovery in Databases (KDD) or data mining is a new interdisciplinary field merging ideas from statistics, machine learning, databases, and parallel and distributed computing. It has been engendered by the phenomenal growth of data in all spheres of human endeavor, and the economic and scientific need to extract useful information from the collected data. The key challenge in data mining is the extraction of knowledge and insight from massive databases. It takes the form of discovering new patterns or building models from a given dataset. Furthermore, implementation of data mining methods in high-performance parallel and distributed computing environments is crucial for ensuring system scalability and interactivity as datasets grow inexorably in size and complexity. This is an extremely important function as many of the paradigms in materials science and engineering are in fact often derived through empirical correlations between observed behavior. Theoretical constructs of these observations have often been established or developed afterwards in order to achieve an explanation of existing data. This is coupled with advances in what is termed computational materials science permitting the simulation of ever-increasing complex phenomena which has effectively provided another source of data generation besides traditional experiments. However the process of sequentially building on experiment / theory and now simulation to iterate a search for a solution to a particular engineering problem has essentially not changed. At present even with the advances in computational materials science, we can only propose theories and suggest possibilities of remedies or on the other hand search for direct experimental evidence for a specific case study. The task of searching all possible levels of data sets and looking for correlations is a daunting if not impossible task.

A key aspect of developing an "informatics" approach to materials discovery is the need to establish the critical array of descriptors of materials attributes that may be subsequently input into a database. Having physically meaningful descriptors is key to developing and searching for associations between apparently disparate or disjointed datasets[3,4]. Establishing such an association through a variety of datasets (experimentally and/ or computationally derived) can serve as a means of searching for *anticipated* structure-property relationships in materials. Data mining is a tool to exploit the masses of available data to accelerate the discovery of these relationships and possible new associations. For us the "association" is between structure and property of materials. Data mining acts as a descriptive tool for hypothesizing relationship between structures and materials that are interpretable by the material scientist.

Materials science data bases (whether derived from experiments or simulation) are *phenomenological* specific and not generally developed around the final *functions* which the material will be used. Hence the usual approach in materials design is to identify expected materials requirements and through a combination of experiment and simulation develop an iterative process of refinement until the desired result is achieved. Even then there may be unexpected events as highlighted by failures in service, which adds to the knowledge base. Collectively, the combination of scientific principles and engineering experience provide the basis by which the *knowledge discovery process* for new materials is developed

DATA MINING TECHNIQUES

There are numerous types of data mining techniques, each appropriate for different types of queries for a given set of data. These include:

- Predictive modeling (classification, regression)
- Segmentation (clustering)
- Dependency modeling (graphical models, density estimation)

- Summarization (associations)
- Change and deviation detection

These can be described briefly as follows:

- Classification and regression: assign a new data record to one of several predefined categories or classes. Regression deals with predicting real-valued fields. Also called supervised learning.
- Clustering: partition the dataset into subsets or groups such that elements of a group share a common set of properties, with high within group similarity and small inter-group similarity. Also called unsupervised learning.
- Association rules: detect sets of attributes that frequently co-occur, and rules among them, e.g. 90% of alloys which are fatigue resistant, also exhibit stress corrosion resistance (60% of all alloys exhibit both properties)
- Sequence mining (categorical): discover sequences of events that commonly occur together, .e.g. In a set of stacking sequences ABCABC is followed by ABAB after a gap of 9 lattice spacings, with 30% probability

In this paper we will focus specifically on Association mining as we feel that it is particular appropriate to address the question of : "*How do we combine heterogeneous databases so that we can discover interesting patterns from them*?"

As in any field, the data of experience accumulates and one can learn from this experience, provided patterns are sought after and examined carefully. Traditionally, the link between fundamental materials science and the practice of materials engineering, has been guided by phenomenological insight supported by detailed theoretical and experimental studies in specific parametric bounds. Expanding beyond these bounds is often limited by simply the vastness of the potential database and searching such large arrays is not often feasible. This is where data mining techniques which provide the interface between statistics and computer science offers a means of truly overcoming this challenge[5].

Materials science databases often contain data from different lengths of scale, from simulations as well as experiments. For instance, how does information on phase stability extracted from thermochemical data compare with information derived from electronic structure calculations? From these independently organized databases, we can compare and search for associations and patterns that can lead to ways of relating information among these different datasets. The new patterns to be discovered should reflect the complex relationships that exist in both spatial and temporal dimensions. Such a pattern search process can potentially yield associations between seemingly disparate data sets as well as establish possible correlations between parameters that are not easily studied experimentally in a coupled manner.

ASSOCIATION RULE MINING (ARM) and MATERIALS DATABASES

Consider a general data mining process that consists of: finding the frequently occurring patterns in the experiment results and parameter space, generating experiments for unexplored controllable sub-patterns of the discovered patterns and inputting the results in a database, checking the frequency of the patterns after the experiment. This process can benefit from database technology in several ways. First of all, many associations discovered in a large variety of databases have to be stored in an efficient structure for later use. Furthermore, we have to store the sub-patterns to be explored and patterns that have already been explored in a data-structure designed to optimize region intersection/difference type operations. Finally, we need to generate experiments based on the physical constraints of the problem (e.g. melting point of material, environment of operation of materials etc.) as well as the underlying feasibility of the experiments(e.g. ease of processing techniques used for materials synthesis).

Since its inception, association rule mining has become one of the core data-mining tasks and has attracted tremendous interest among researchers and practitioners. ARM is undirected or unsupervised data mining for variable-length data, and it produces clear, understandable results. It has an elegantly simple problem statement: to find the set of all subsets of items or attributes that frequently occur in many database records or transactions, and additionally, to extract rules on how a subset of items influences the presence of another subset. Although ARM has a simple statement, it is computationally and I/O intensive. Because data is increasing both in terms of the dimensions (number of items) and size (number of transactions), one of the main attributes needed in an ARM algorithm is scalability: the ability to handle massive data stores. Sequential algorithms cannot provide scalability, in terms of the data dimension, size, or runtime performance, for such large databases. Therefore, we must rely on high-performance parallel and distributed computing.

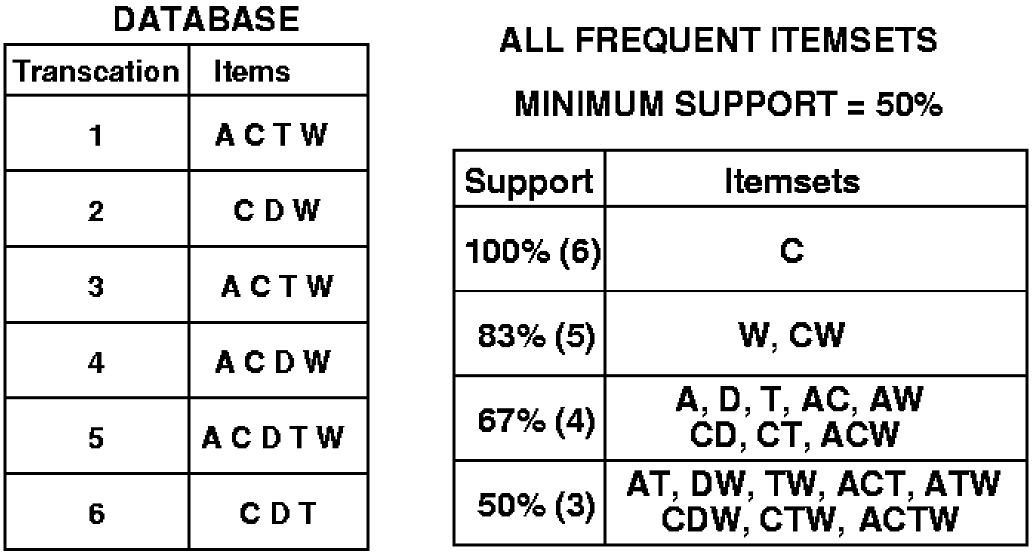
ARM METHODOLOGY

Association mining works as follows. Let I be a set of items, and D a database of transactions, where each transaction has a unique identifier (*tid*) and contains a set of items called an *itemset*. An itemset with k items is called a k-itemset. The *support* of an itemset X, denoted $\sigma(X)$, is the number of transactions in which that itemset occurs as a subset. A k-subset is a k-length subset of an itemset. An itemset is frequent or large if its support is more than a user-specified *minimum support (min_sup)* value. F_k is the set of frequent k-itemsets. . A frequent itemset is maximal if it is not a subset of any other frequent itemset.An *association rule* is an expression $A \Rightarrow B$, where A and B are itemsets. The rule's support is the joint probability of a transaction containing both A and B, and is given as $\sigma(A \cup B)$. The confidence of the rule is the conditional probability that a transaction contains B, given that it contains A, and is given as $\sigma(A \cup B)/\sigma(A)$. A rule is frequent if its support is greater than *min_sup*, and strong if its confidence is more than a user-specified minimum confidence *(min_conf)*.

Data mining involves generating all association rules in the database that have a support greater than *min_sup* (the rules are frequent) and that have a confidence greater than *min_conf* (the rules are strong). This task has two steps:

- Find all frequent itemsets having minimum support. The search space for enumeration of all frequent itemsets is 2^m, which is exponential in m, the number of items. However, if we assume the transaction length has a bound, we can show that ARM is essentially linear in database size.
- Generate strong rules having minimum confidence, from the frequent itemsets. We generate and test the confidence of all rules of the form X-$Y \Rightarrow$ Y, where $Y \subset X$, and X is frequent. Because we must consider each subset of X as the antecedent, the rule-generation step's complexity is $O(r \cdot 2^l)$, where r is the number of frequent itemsets, and l is the longest frequent itemset.

Consider the example database shown in Figure 1. There are five different items $I = \{A, C, D, T, W\}$. The database comprises six "transactions" or sets of measurements containing a subset of these five items. Figure 1 shows all the frequent itemsets contained in at least three transactions (*min_sup* = 50%). Figure 2 also shows the set of all association rules with *min_conf* = 80%.



DATABASE

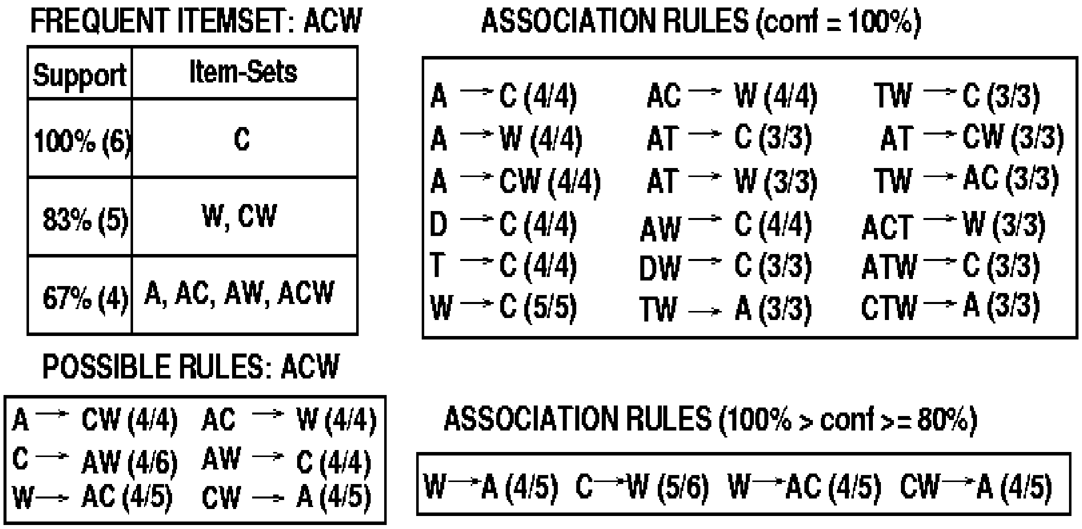
Transcation	Items
1	A C T W
2	C D W
3	A C T W
4	A C D W
5	A C D T W
6	C D T

ALL FREQUENT ITEMSETS

MINIMUM SUPPORT = 50%

Support	Itemsets
100% (6)	C
83% (5)	W, CW
67% (4)	A, D, T, AC, AW CD, CT, ACW
50% (3)	AT, DW, TW, ACT, ATW CDW, CTW, ACTW

Figure 1: a) Example Database, b) Frequent Patterns (min_sup = 50%)



FREQUENT ITEMSET: ACW

Support	Item-Sets
100% (6)	C
83% (5)	W, CW
67% (4)	A, AC, AW, ACW

POSSIBLE RULES: ACW

A → CW (4/4)	AC → W (4/4)
C → AW (4/6)	AW → C (4/4)
W → AC (4/5)	CW → A (4/5)

ASSOCIATION RULES (conf = 100%)

A → C (4/4)	AC → W (4/4)	TW → C (3/3)
A → W (4/4)	AT → C (3/3)	AT → CW (3/3)
A → CW (4/4)	AT → W (3/3)	TW → AC (3/3)
D → C (4/4)	AW → C (4/4)	ACT → W (3/3)
T → C (4/4)	DW → C (3/3)	ATW → C (3/3)
W → C (5/5)	TW → A (3/3)	CTW → A (3/3)

ASSOCIATION RULES (100% > conf >= 80%)

W → A (4/5)	C → W (5/6)	W → AC (4/5)	CW → A (4/5)

Figure 2: Establishment of “Strong Rules” (min_conf = 80%)

As a simple illustration of this logic as applied to a materials database problem, we show the example of the utilization of “structure maps” which are based on searching for crystallographic descriptors for association of crystal stoichiometry with structure type. A recent example is that proposed by Pettifor using descriptors based on a sequencing pattern of the Mendeleev number[6]. As an example of association of vastly different length and time scales, Figure 3 shows a mapping of the association of structure type with an engineering scale metric describing the quality of welding or joining dissimilar materials. Clearly there is no ab-initio theory to correlate such vastly different types of data sets, but this provides an example of the disparate forms of data sets that exist in materials science.

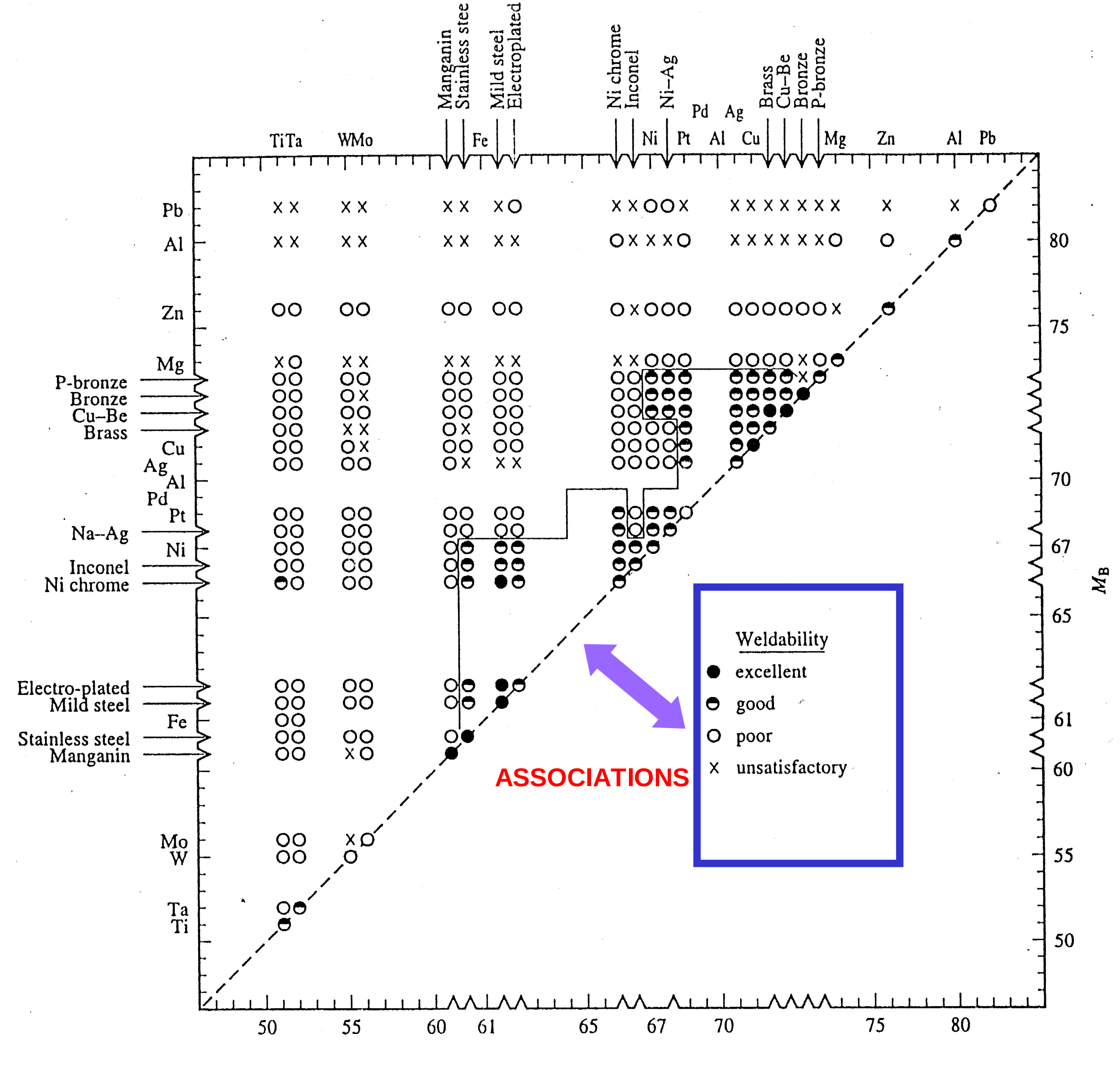


Figure 3: Pettifor structure map graphically showing associations between weldability performance data and weld chemistry (From Pettifor (1988))

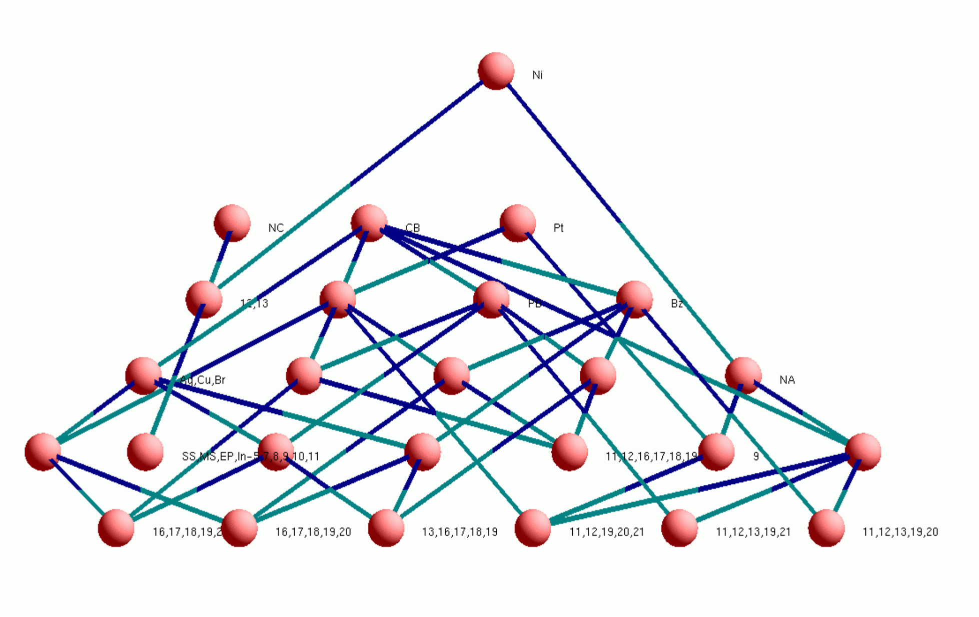


Figure 4: Frequency tree calculated from structure map data of Figure 3 using association mining algorithms

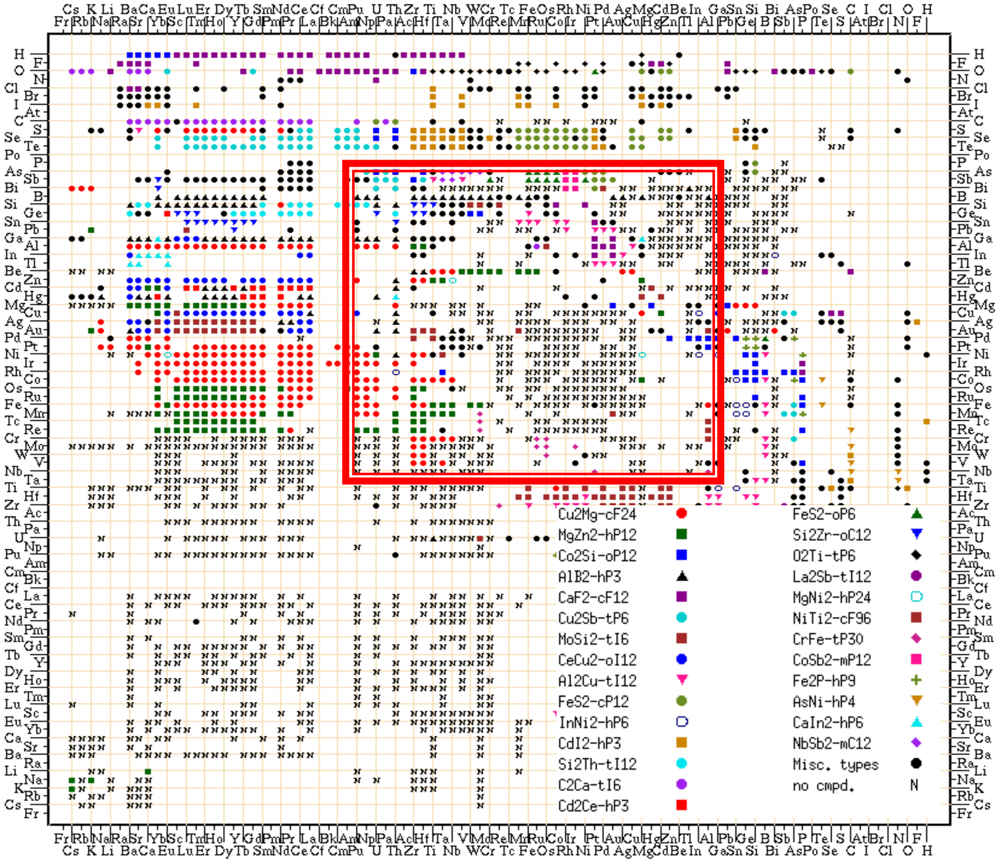


Figure 5: Structure map for AB_2 stoichiometry highlighting region mapped in Figure 3. Note that regions of "good weldability" appear to match welding combinations that do *not* produce AB_2 compounds (adapted from J.Rodgers – private communication and Pettifor, 1988)

Figure 4 shows the graphical representation of an hierarchial tree of associations based on the data from Figure 3. This form of data association representation is based on classifying the frequency of occurrence of combinations of materials demonstrating excellent or good weldability. For example :

- Ag / Cu / Brass / Cu-Be : Frequency 9
- Bronze / Cu-Be / Pt : Frequency 9
- Ni-Ag / Ni / Pt: Frequency 8
- Stainless steel / Mild steel / Electro-plated / Inconel / Ni chrome / Ni: Frequency 8…….etc

When this data is fed back into our general structure map database, we find an interesting association between the occurrence of good welds with the absence of compound formation. This statistical inference is physically plausible as it would suggest that brittle compounds formed at a weld could contribute to poor joining characteristics. Hence, while the database on crystallographic structure type is totally independent of weldability databases, association rules can provide valuable insight across heterogeneous databases. While for the purpose of example, a small dataset was used here, the data mining algorithms can handle terabytes of information permitting vast and multidimensional associations.

CONCLUSIONS

Information association forms a key aspect of knowledge discovery in science and engineering. This association usually is built around low dimensional datasets and help to guide theoretical and experimental studies which either try to expand those datasets or develop a deeper fundamental understanding of the phenomena. The challenge , especially in materials science which covers vast arrays of length and time scales, is to search for meaningful associations , especially when the databases are disparate and heterogeneous. In this paper we hav introduced the importance of taking advantage of new developments in the computer science community in the field of data mining as being an enabling tool to meet this challenge. The example of associating crystal structure maps to weldability has for example permitted one to infer or anticipate possible rules for designing good welds. While the example was presented in a two dimensional manner, the frequency association mining algorithms are scalable to vast data sets.

REFERENCES

1. J. Nagamatusu, N.Nakagawa, T. Muranaka, Y. Zenitani, and J. Akimitsu : Superconductivity at 39K in magnesium boride, Nature **410** 53-65 (2001)
2. J.H.Schon, A. Dodabalpur, Z.Bao, Ch.Kloc, O.Schenker and B. Batlogg: Gate induced superconductivity in a solution-processed organic polymer, Nature **410** 189-192 (2001*)*
3. D.Rodgers and A.D. Mighell, “The Use of Lattice and Emperical Formula in the Registration of Crystalline Materials” J. Chem. Inf. Comp. Sci, **21** 42 (1981)
4. R.Car and M. Parrinello “Unified Approach for Molecular Dynamics and Density Functional Theory” Phys.Rev. Lett. **55** 2471 (1985)
5. Mohammed J. Zaki, Scalable Algorithms for Association Mining, IEEE Transactions on Knowledge and Data Engineering, Vol. 12, No. 3, pp 372-390, May/June 2000.
6. D.G.Pettifor “Structure Maps for Pseudobinary and Ternary Phases” Materials Science and Technology **4** 675-691 (1988)