The EDAM Project: Mining Atmospheric Aerosol Datasets [1]

Raghu Ramakrishnan, James J.Schauer, Lei Chen, Zheng Huang, Martin M. Shafer

University of Wisconsin-Madison
e-mail: raghu@cs.wisc.edu, jschauer@engr.wisc.edu, chenl@cs.wisc.edu, zhuang@cs.wisc.edu, mmshafer@facstaff.wisc.edu

Deborah S. Gross, David R. Musicant

Carleton College
e-mail: dgross@carleton.edu, dmusican@carleton.edu

Abstract

Data mining has been a very active area of research in the database, machine learning, and mathematical programming communities in recent years. EDAM (Exploratory Data Analysis and Management) is a joint project between researchers in Atmospheric Chemistry and Computer Science at Carleton College and the University of Wisconsin-Madison that aims to develop data mining techniques for advancing the state of the art in analyzing atmospheric aerosol datasets.

There is a great need to better understand the sources, dynamics, and compositions of atmospheric aerosols. The traditional approach for particle measurement, which is the collection of bulk samples of particulates on filters, is not adequate for studying particle dynamics and real-time correlations. This has led to the development of a new generation of real-time instruments that provide continuous or semi-continuous streams of data about certain aerosol properties. However, these instruments have added a significant level of complexity to atmospheric aerosol data, and dramatically increased the amounts of data to be collected, managed, and analyzed. Our ability to integrate the data from all of these new and complex instruments now lags far behind our data-collection capabilities, and severely limits our ability to understand the data and act upon it in a timely manner.

In this paper, we present an overview of the EDAM project. The goal of the project, which is in its early stages, is to develop novel data mining algorithms and approaches to managing and monitoring multiple complex data streams. An important objective is data quality assurance, and real-time data mining offers great potential. The approach that we take should also provide good techniques to deal with gas-phase and semi-volatile data. While atmospheric aerosol analysis is an important and challenging domain that motivates us with real problems and serves as a concrete test of our results, our objective is to develop techniques that have broader applicability, and to explore some fundamental challenges in data mining that are not specific to any given application domain.

Keywords: Atmospheric aerosols, ATOFMS, association rules, Database Systems, Data Mining, query optimization, classification, clustering, frequent itemsets, mass spectra, multi-step mining, regression, subset mining, support vector machines

1. Introduction

Increasing concern over the role of atmospheric particles (aerosols) on global climate change (IPCC 96; NRC 96), human health and welfare (NRC 98) and the Earth's ecosystem (Baker 97) has created a great need to better understand the composition, origin, and influence of atmospheric pollutants. In order to develop control strategies that can mitigate the onset of climate change, as well as the degradation of the environment and our

[1] This work is supported through National Science Foundation ITR grant IIS-0326328.

quality of life, there is a great need to develop tools to better assess the origin and impact of atmospheric aerosols. The traditional approach for particle measurement, which is the collection of bulk samples of particulates on filters (Chow 95), is not adequate for studying particle dynamics and real-time correlations. This has led to the development of a new generation of real-time instruments (Turpin 90, Landis 02, Moosmuller 01), including aerosol mass spectrometers (Prather 94, Noble 96, Gard 97, Suess 99), which provide continuous or semi-continuous streams of data about certain aerosol properties. However, these instruments have added a significant level of complexity to atmospheric aerosol data, and dramatically increased the amounts of data to be collected, managed, and analyzed. Our ability to integrate the data from all of these new and complex instruments now lags far behind our data-collection capabilities, and severely limits our ability to understand the data and act upon it in a timely manner.

Data mining has been a very active area of research in the database, machine learning, and mathematical programming communities in recent years, and there is a wealth of techniques that can be brought to bear on atmospheric aerosol datasets. In particular, we show how a powerful class of analysis techniques developed for analyzing customers' purchase histories can, unexpectedly, be brought to bear on mass spectrometry data by preprocessing it appropriately. Unfortunately, while some of these techniques are available in commercial data analysis products, many of the most useful ideas are of very recent origin, and are at the research stage. Further, atmospheric aerosol analysis raises a number of challenges for which there is currently no satisfactory solution. These range from how to incorporate scientists' domain knowledge to data provenance, data validation and collaboration support. Large datasets that are gathered in real-time require robust quality-assurance protocols to ensure data reliability, and improved data management tools are a necessary component to achieve this goal.

1.1 Objectives

The objectives of the EDAM project can be summarized as follows. We aim to apply and advance the state of the art in data mining in the following main ways:

- **Applying Currently Available Data Mining Techniques:** A number of currently available techniques can be applied to real-time and semi-continuous atmospheric aerosol data streams to greatly mitigate pressing bottlenecks. Time-series analysis, clustering, and decision trees are well-known techniques for which robust software is available (both in commercial tools, and in freely distributed source code form). Rather surprisingly, a broad class of techniques (association rules and sequential patterns) developed for analyzing customer transactions is also applicable—we show how mass spectrometry data can be approximated in a form that mimics customer transactions. However, for many aerosol data analysis tasks, it is not clear what existing techniques (if any) are applicable, and how best to apply them. The Carleton and UW groups both include computer scientists as well as domain scientists (i.e., chemists, environmental engineers, and atmospheric scientists) because we anticipate that the key to solving such problems will be close, day-to-day interdisciplinary collaborations.
- **Developing Novel Mining Paradigms:** There is no framework that enables scientists to create **multi-step analyses** using one or more mining techniques, and to focus the patterns generated by these techniques by incorporating domain knowledge into the analysis. We aim to generalize and adapt existing algorithms, or develop new ones when necessary, to create a suite of algorithms for traditional analysis tasks that can be easily combined and trained with a variety of additional knowledge. A common pattern of multi-step mining arises when we want to find correlations between (parts of) different datasets, and is motivated by problems arising in combining mass spectrometry and environmental monitoring data. We are developing a stylized framework for such analyses, and we believe that this framework, which we call **subset mining**, will find broad applicability in a number of other domains.
- **Monitoring Complex System State:** Analyzing real-time streams is an active area of research, and many algorithms have been proposed to maintain different kinds of data mining models in real time. Clearly, these results are of great relevance to us. We will focus on how to allow environmental scientists to describe complex "states" that they are interested in monitoring (e.g., the correlation between levels of reactive mercury and particulate sulphate ion), and combine techniques for incremental maintenance of the underlying data mining models to achieve incremental monitoring of the composite state. This is essentially an extension of the multi-step mining framework to support incremental model maintenance over data streams.

1.2 Organization of this Paper

We begin by presenting some background material in Section 2. We discuss the increasingly data-intensive nature of mass spectrometry and environmental monitoring and describe the underlying datasets in Section 2.1, and present an overview of data mining techniques in Section 2.2. In Section 3, we discuss some general data mining challenges and approaches that we intend to investigate in the application context of atmospheric aerosol analysis. We then describe a number of specific challenges in analyzing atmospheric aerosol datasets in Sections 4, 5, 6 and 7, including current and proposed approaches. Sections 4 and 5 consider how to interpret a mass spectrum, and the other sections consider how this information can be used to achieve a better understanding of various phenomena associated with atmospheric aerosols.

2. Background

In this section, we provide the necessary background for the rest of this paper.

2.1 Mass Spectrometry and Environmental Monitoring

As scientists, public health officials, and government regulatory agencies strive to better understand the environment, many aspects of this complex system have emerged as both critically important and difficult to understand. One of the primary examples of this is particulate matter. *Aerosol particles*, which are often complex mixtures of organic and inorganic solids and liquid suspended in the air, exist with a variety of sizes, from freshly nucleated particles with nanometer diameters to micrometer sized particles from suspended dust or vehicle emissions (Seinfeld 98). In addition, these particles evolve and "age" during atmospheric transport (Seinfeld 98). The size and composition of the particles is directly related to their origin, evolution and deposition and is intimately related to their environmental and health effects (NRC 98).

The effort to obtain better information about atmospheric particles has recently focused on the development of a variety of instruments that measure physical or chemical properties of aerosols in real time. Among the most complex datasets produced comes from a type of instrument that is becoming ever more popular, the *aerosol time-of-flight mass spectrometer (ATOFMS)* (Prather 94, Noble 96, Gard 97, Suess 99). This instrument samples aerosol particles directly from the ambient air or from an emission source and obtains size (aerodynamic diameter) and chemical composition information on one particle at a time, in real time. Because this instrument represents the most complex aerosol dataset we currently have, we will focus the description here on its data. However, we expect the techniques that we develop for this instrument to be directly applicable to many other complex atmospheric datasets. In the ATOFMS dataset, the chemical composition information is obtained through laser desorption/ionization (LDI) of the individual particles and analysis of the resulting positive and negative ions by *time-of-flight (TOF) mass spectrometry*, which detects ions produced in each ionization event as a function of their arrival time at a detector. This commercial version of this instrument (TSI, Inc., Model 3800 ATOFMS) can currently obtain mass spectra for up to about 120 particles per minute. The actual sampling rate depends on the concentration of particles in the source being sampled.

A *mass spectrum* is a plot of *signal intensity* (often normalized to the largest peak in the spectrum) versus the *mass-to-charge (m/z) ratio* of the detected ions. Thus, the presence of a *peak* indicates the presence of one or more ions containing the *m/z* value indicated, within the ion cloud generated upon the interaction between the particle and the laser beam. In many cases, the ATOFMS generates elemental ions. Thus, the presence of certain peaks indicates that elements such as Na^+ (*m/z* = +23) or Fe^+ (*m/z* = +56) or O^- (*m/z* = -16) ions are present. In other cases, cluster ions are formed, and thus the *m/z* observed represents that of a sum of the atomic weights of various elements; examples include C_3^+ (*m/z* = +36) or $BaOH^+$ (*m/z* = +155) or $C_{16}H_{10}^+$ (*m/z* = +202). In TOF mass spectrometry, the area under the peak is proportional to the concentration of ions that reach the detector. The raw data on the x-axis is in time-of-flight units, which is converted to *m/z* based on the fact that each ion formed is accelerated to the same kinetic energy in the mass spectrometer's source.

The raw mass spectra obtained by ATOFMS have the following parameters *per particle*:

Number of mass spectra	2 (one for positive ions, one for negative ions)
Raw data points in mass spectrum	30,000 points
x-axis distribution	One data point every 2 ns
File format	x values = TOF y values α detector voltage (i.e., signal intensity)

Clearly, this dataset is a time-series of unusual complexity—the per-particle data for each (observed) instant in time is a spectrum containing 30,000 points! Upon processing (i.e., calibration of TOF to *m/z* and picking peaks), information of interest from the mass spectra of each particle includes peak *m/z*, peak area, peak area relative to total area in the spectrum, and peak height.

A growing number of such instruments are being deployed in laboratories and field locations around the world. While aerosol mass spectrometry data is among the most complex environmental particle data that is currently obtained, generating hundreds of unique and data-rich spectra per minute, many other emerging instruments generate two or three-dimensional datasets every few minutes. This contrasts with the traditional idea of a monitoring instrument providing a two-dimensional plot of a single parameter (such as mass concentration) as a function of sampling time, or of filter-based samples providing a time-integrated snapshot of composition over a finite time period. These other data streams are more routinely generated in environmental monitoring. As they are generally simpler in structure, we will describe them as required in the rest of this paper. Instrument development and deployment is progressing at a tremendous pace, far outstripping our current ability to integrate the data from all of these new and complex instruments and conduct studies in a timely manner.

It is important to recognize that extensive work has been done in employing statistics based models for the analysis of filter-based (i.e., off-line) measurements of atmospheric aerosols (Hopke 85). These techniques have only been integrated in a very limited capacity with complex datasets such as ATOFMS data (Bhave 02) and are unlikely to advance further without the development of the tools such as those described in this proposal. Furthermore, these statistical tools are largely limited to linear systems, which cannot address many of the complex processes that occur in the atmosphere. In contrast, the tools that we seek to develop in this project draw from a wide range of disciplines, complementing statistical approaches, are designed to address non-linear processes in the atmosphere as well. Finally, we observe that although the detailed chemistry of aerosols is important if not fundamental, regulatory and analytical constraints will likely require that a more integrative aerosol property (optical or physical) be routinely measured. The tools under development will improve our understanding of the relationships between chemical and physical properties of aerosols.

2.2 Data Mining

Exploratory analysis of large datasets, called *data mining*, draws upon techniques from a range of disciplines, including Database Systems, Machine Learning, Mathematical Programming, and Statistics. Examples of traditional mining techniques include classification and clustering (Berry 99, Fayyad 96, Han 00, Hand 00, Hastie 01, Weiss 97, Witten 99, Cherkassky 98), and time-series analysis (Box 94, Das 98, Faloutsos 94, 97, Weigend 94, Keogh 01, Agrawal 93b, 95c, Chan 99, Chu 99, Loh 00, Popivanov 02, Rafiei 98, 99, Wu 00). In recent years, there has been great emphasis on developing mining algorithms that scale to large datasets (Zhang 97, Gehrke 99, 00, Ganti 99, Bradley 98b, Guha 98, Huang 97), and on mining evolving (Ester 98, Yi 00, Ganti 00, Veloso 02) and continuously generated data (Cortes 00, Lee 98). Market-basket data, or data about customers' transactions, has received much attention, and several pattern detection techniques based on co-occurrence of items within a single transaction (Agrawal 93, 94, 95a, Brin 97, Han 95, Srikant 96b, Imielinski 00, Korn 98, Klemettinen 94, Ozden 98, Tsur 98, Zaki 00) and across a series of transactions (Agrawal 95b, Srikant 96, Joshi 00, Tung 99, Lu 00) have been proposed.

Even though we deal with large, rapidly growing datasets, scalable, fast algorithms are already available for a variety of traditional analysis techniques. e.g., clustering (Bradley 98b, Zhang 97, Ganti 99, Guha 98, Huang 97), classification using decision trees (Gehrke 99, 00) and SVM kernels (Joachims 99, Platt 99a, 99b, Lee 01, Smola 00, DeCoste 99, Fung 01b, Musicant 99, 01b, 01c), discovering associations (Brin 97a, Cheung 96, Fukuda 96, Ganti 99, Han 97, Mannila 94, Miller 97, Park 95, Pasquier 98, Sarawagi 98, Savasere 95, 98, Silverstein 98, Toivonen 96,

Yoda 97, Zaki 97, 99, 01) and detecting sequential patterns (Bayardo 98, Srikant 96, Yang 02, Ayres 02, Shintani 98, Pei 01, Zaki 98, Zaki 01a).

Much of the data we are concerned with is obtained through monitoring activities and is in the form of real-time or semi-continuous streams. Stream database management is being studied in ongoing projects at UC Berkeley, Cornell, Brown/MIT, Stanford, and Wisconsin, among other places (Madden 02, 02a, Chandrasekharan 03, Bonnet 01, Yao 02, Carney 02, Arasu 02, Babcock 02, Babu 01, Manku 02, Manku 98). There is extensive ongoing research in mining data streams (Chakrabarti 02, Cormode 02, Datar 02, Faloutsos 02, 02a, Ge 00, Himberg 01, Hulten 01, Keogh 02, O'Callaghan 02, Wang 02); see (Bradley 02. Garofalakis 02, Smyth 02) for recent surveys. However, existing work does not adequately address fusion of multiple streams using domain knowledge (e.g., underlying chemistry or environmental characteristics).

3. New Mining Paradigms

In this section, we outline some "grand challenge" problems that motivate us. These are problems that go beyond the specific domain of atmospheric aerosols, though they have many concrete applications in this domain.

3.1 Multi-Step Mining

There is no general framework for systematically applying one or more analysis techniques to (parts of) a dataset in a multi-step mining process—there is neither a framework for specification of such multi-step mining strategies, nor a framework for optimizing the computation by taking the interplay of the different mining steps into account. Since much of the time involved in data mining efforts is in user-driven, iterative, exploratory application of data mining algorithms, rather in the execution time of the algorithms themselves, progress on a compositional framework multi-step mining can have a significant payoff by reducing the real bottleneck in most mining efforts, which is the time taken by an analyst to digest the result of each analysis step and to set up the next step.

Indeed, given that database systems have been centrally concerned about optimizing queries composed of multiple operators (Ramakrishnan 02), and given the extensive literature on new database mining algorithms, it is rather surprising that there are no published results on specification and optimization of multi-step mining strategies or even on cost-estimation for data mining algorithms. The closest in spirit to such optimization is an interesting series of papers on how to constrain the generation of association rules using a range of constraints (Grahne 00, Kamber 97, Lakshmanan 99, Ng 98, Pei 01a, Srikant 97).

Completely automatic discovery of truly useful insights is, in our opinion, a pipe dream in most application scenarios. A more realistic goal, perhaps, is a framework for a user to describe a space of exploratory sequences, together with notions of "interestingness" for the patterns or insights discovered thereby, and for the system to find ways to explore this space intelligently and efficiently. Domain knowledge can be exploited in one of two ways—by using it to focus the patterns found by a given mining technique, and by using it to select appropriate techniques for different tasks and to combine the results. While much remains to be done, there has already been some work showing how mining algorithms can be adapted to incorporate prior knowledge (Towell 94, Fung 01, Clark 94, Zhou 01, Padmanabhan 98, 00, Cook 96, Clair 98). There is also some work on post-processing association rules (Lent 97).

We seek to generalize and adapt existing algorithms, or develop new ones when necessary, to create a suite of algorithms for traditional analysis tasks that can be easily combined and trained with a variety of additional knowledge. One class of multi-step mining strategies that we intend to pursue to this end, called *subset mining*, is described in the next section.

3.2 Subset Mining

Database queries and data mining algorithms allow us to discover various properties of a given dataset, ranging from simple aggregates such as average temperature by location or more complex patterns such as clusters and classes. Typically, data mining tasks consist of finding "interesting" patterns in a dataset. Often, however, the questions of interest have the form "Is there some *subset* of the data that is interesting?" The interestingness of a

data subset can be measured in a number of ways, e.g., through a database query that computes an aggregate or a data mining query that identifies a class of patterns and defines a measure of "interestingness" over patterns.

We define **subset mining** to be the class of analysis tasks that search over subsets of a given dataset to identify interesting subsets. The distinguishing characteristic is that some computation is (in principle, at least) carried out over (all or several, possibly overlapping) subsets of a dataset. Clearly, the complexity of the base computation is amplified manifold because of the number of potential subsets over which it must be iterated. Often, however, domain knowledge can be brought to bear on which subsets to consider potentially interesting, and there may be structural relationships between these subsets (e.g., disjointness or a predictable form of overlap). Exploiting these characteristics to arrive at an efficient evaluation plan can make the difference between whether or not a given subset mining task is computable, given reasonable computing resources.

- *We propose to develop a framework for specifying subset mining tasks using other data mining algorithms as building blocks, and automatically arriving at an efficient execution strategy.*

We introduce the subset mining paradigm through an example. Consider a table (*HgReadings*) of hourly reactive mercury-level readings (Landis 02), with one row for each reading, and another table of particulate sulfate ion-concentration readings (*IonReadings*), also measured hourly. If we want to find all times at which the reactive mercury-level is "high" (above some threshold), this is a simple selection query over the first table. If we want to find the average concentration of, say, particulate sulphate ion, this is a simple aggregate query on the second table. Combining these queries, we can ask for the average concentration of sulphate ion when the reactive mercury-level is high. All three queries are readily expressed in SQL, the standard database query language.

In contrast, consider the following query, which is of great interest in atmospheric studies seeking to understand the sources of reactive gaseous mercury: *Are certain ranges of reactive mercury levels strongly correlated to unusually high concentrations of particulate sulphate ion?* This is an example of a subset mining query, and it is not expressible in SQL without significant restrictions. As another example, if we have identified *clusters* based on the location of each reading, we can readily refine the previous query to ask whether there are such correlations (for some ranges of reactive mercury levels) at certain locations. The main challenge is that we must consider all possible reactive mercury ranges, and for each, carry out (at a minimum) a SQL query that calculates corresponding sulphate ion concentrations. In addition, like typical "data mining" questions, this query involves inherently fuzzy criteria ("strong correlation", "unusually high") in whose precise formulation we enjoy some latitude.

To summarize, there are three main parts to a *subset mining* query: (1) A criterion that generates several subsets of a table, (2) A correspondence—typically a relational expression—that generates a subset of a second table for each of these subsets, and (3) A measure of interestingness for the second subset (that indirectly serves as a similar measure for the original subset). To see why subset mining queries are especially useful for integrated analysis of multiple datasets, using a combination of mining techniques, observe that Steps (1) and (3) could both be based on the results of (essentially any) mining techniques, rather than just simple SQL-style selections and aggregation. The computationally challenging aspect arises from the potentially large number of subsets involved, and from the computationally intensive nature of the criterion used for subset generation and interestingness-measurement. (As a special case, we note that Part (2) may be omitted, and we may have to enumerate and identify interesting subsets of a single table.)

We note that subset mining is closely related to the subgroup discovery problem (Klosgen 96, Wrobel 97, Lavrac 02) studied in inductive logic programming (Lavrac 94). Given a description language L and a valuation function d, the subgroup discovery problem consists of finding a set S of sentences such that the valuation of each sentence in S is greater than the valuation of any sentence not in S, and further, each sentence is shorter than some threshold length k. Intuitively, each sentence describes a group of data objects, and we want to find the most concisely described groups with the highest value. Subset mining differs from subgroup discovery in two main ways. First, we have chosen to focus on an important three-step analysis pattern, and want to develop algebraic optimization approaches to exploit the connections between these steps. Second, we have not limited ourselves to a language-centric definition of subgroups; the three steps in our approach can use arbitrary (but well-defined) "black boxes".

In Sections 5, 6 and 7, we will point out numerous concrete instances where the paradigm of subset mining is valuable. A significant technical challenge will be to optimize subset mining, at least for several particular instances (i.e., query classes). This will require research into cost estimation for the data mining techniques used as components of the subset mining instance, as well as ways to "push" abstract constraints derived from the subset-generation component into the interestingness-measure component (and vice-versa). While there are parallels in database query optimization and evaluation of relational algebra operations, we expect that these issues will have to be tackled on a case-by-case basis for different data mining techniques when they are used to instantiate the subset mining framework. These are admittedly difficult challenges, but success is not all-or-nothing. Ultimately, we believe that the ability to identify interesting subsets of a large dataset (and not just specific patterns of interest) will be a significant step forward in the area of data mining, and that if we are able to articulate and efficiently support at least some instances of the paradigm, others in the field will extend the effort by addressing how other mining techniques can be supported in the subset mining context.

3.3 Describing and Monitoring Complex System State

A central research challenge is turning multiple streams of raw mass spectrometer and other sensor readings into meaningful, continuous, dynamic views of environmental characteristics, at the semantic level that environmental scientists think of these characteristics. We see the description of complex environmental states in terms of patterns extracted through a combination of data mining techniques as an (especially important) instance of multi-step mining. The distinguishing characteristic, and the source of optimization challenges, is the stream nature of the underlying data and the desire to maintain state models and measure deviations in real-time. We will build upon the existing literature describing incremental techniques for maintaining a range of data mining models, which we discussed earlier. We will draw upon and extend our prior work in stream processing (Donjerkovic 00, Ramakrishnan 98, Seshadri 95, 96) and stream data mining (Ganti 00, 01). One of the key challenges in monitoring is measuring significant changes, at the level of the model being maintained; we plan to extend the framework that we proposed in (Ganti 02).

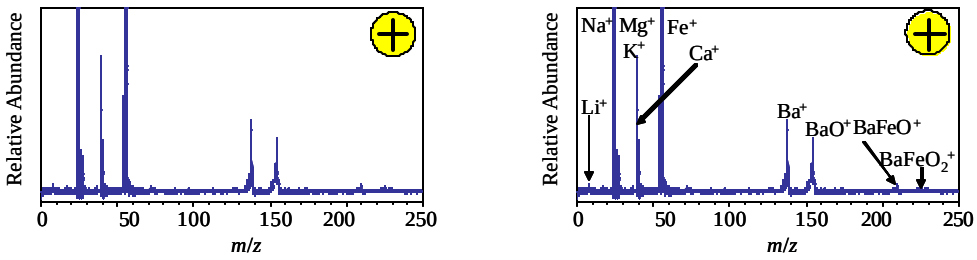
4. Interpreting ATOFMS Data: Labeling Spectra

In this section, we describe a central task in interpreting ATOFMS readings that serves as a first step in further analysis, namely how to label the peaks in a mass spectrum with the ions whose presence they indicate. While we focus on time-of-flight mass spectrometry, the underlying problems (and very likely the solutions) apply to other kinds of mass spectrometry as well. The discussion in this section indicates how the problem of labeling spectra can be formalized rigorously, and is based on (Chen 03, Huang 03), where algorithms for labeling are proposed and evaluated.

The positive mass spectrum of a particle sampled from re-suspended brake dust is shown below. The information content of the labeled spectrum is much greater.

Unlabeled peaks: All interpretation by a user comes from prior knowledge of the translation of *m/z* into chemical composition.

Labeled peaks: Interpretation can be done by someone less familiar with mass spectra, as the spectrum indicates its chemical composition.



For many kinds of analyses, what is significant in each particle's mass spectrum is the composition of the particle, i.e., the ions identified by the peak labels (and, ideally, their proportions in the particle, and our confidence

in having correctly identified them). While this representation is less detailed than the labeled spectrum itself, it allows us to think of the ATOFMS data stream as a time-series of observations, one per observed particle, where each observation is a set of ions (possibly labeled with some additional details). This is precisely the *market-basket* abstraction used in e-commerce: a time-series of *customer transactions*, each recording the items purchased by a customer on a single visit to a store. This analogy opens the door to applying a wide range of association rule and sequential pattern algorithms to the analysis of mass spectrometry data. The quantitative and probabilistic nature of each ion "item", of course, naturally suggests a refinement of the market-basket abstraction that requires a significant rethinking of the corresponding algorithms. While such refinement is a direction for data mining research that we will explore, simply applying standard algorithms holds great potential for many of the problems discussed in Sections 5, 6 and 7.

Viewing it from an optimization perspective, the labeling process is to find an integral linear combination of elements in the database that best approximates the observed spectrum. The search space therefore consists of a variety of spectra created by combining elements in the database, and the objective is to find the spectrum that is as close to the target spectrum as possible. This might be formulated as a mixed integer program or an approximation to one, and solved through known mathematical programming algorithms such as branch-and-bound methods (Wolsley 98, Chvatal 83).

Unfortunately, there are complications in the chemistry. For example, several of the most abundant isotopes of many environmentally relevant metals are subject to severe polyatomic interferences, thereby making it difficult to label the mass spectral peaks unequivocally with the corresponding ions. For example: ^{60}Ni (26.16% abundant) overlaps with $^{44}Ca^{16}O$, $^{23}Na^{37}Cl$, and $^{43}Ca^{16}O^{1}H$. The polyatomic species $^{23}Na^{35}Cl$, $^{42}Ca^{16}O$, $^{29}Si^{29}Si$, as well as Fe, all overlap in mass with the other abundant nickel isotope ^{58}Ni (67.77% abundant). Molecular species interfering with the three abundant Zn isotopes include $^{32}S^{16}O^{16}O$, $^{48}Ca^{16}O$, $^{32}S^{32}S$, $^{48}Ti^{16}O$ (^{64}Zn); $^{56}Fe^{12}C$, $^{34}S^{16}O^{16}O$, $^{32}S^{34}S$ (^{66}Zn); $^{34}S^{34}S$, $^{35}Cl^{16}O^{17}O$ (^{68}Zn).

Significant improvement in analytical accuracy may be achieved through the use of multi-element interference equations. For example, consider ^{58}Ni. The iron interference may be isolated by measuring another Fe isotope (54, 56, 57), and then calculating the expected interference at 58 using natural abundance ratios. The silicon polyatomic interference is addressed by quantifying the ^{28}Si-dimer at mass 56 and then, as before, calculating the ^{29}Si-dimer at mass 58 using abundance ratios. Oxide, carbide, and chloride isobars are addressed in a similar manner by quantifying the respective lighter or heavier identical species. If the most logical or practical isotope for interference correction is itself subject to isobars, then one must first isolate the interference on this isotope before attempting to correct another with it. Clearly, these "rules" can get complicated very quickly, and, importantly, must be constrained to avoid over-correction. To date, no attempts have been made at improving the quantification of zinc and nickel by accounting for these interferences; to do so requires us to find ways to incorporate sophisticated domain knowledge (such as multi-element interference equations) into the mining algorithms used to label peaks. The development of such tools will greatly enhance the utility of ATOFMS data. More generally, such developments would also directly benefit other mass spectrometry techniques including inductively coupled plasma mass spectrometry (ICPMS) analysis.

Domain knowledge can also be integrated with an approach rooted in semi-supervised learning (Fung 99, Bennett 98, Basu 02, Letouzey 00, Blum 01, Goldman 00, Mitchell 99). Semi-supervised learning is typically used when the goal is to learn a relationship when some of the data is labeled, though a large quantity is unlabeled. While it is expensive to have human experts label the peaks for a large number of spectra, we could instead use clustering techniques to identify a small number of prototypical spectra. Human experts would label these prototypes, which we would then use as part of a training set in order to build a classification scheme. Semi-supervised learning would then train on both the labeled and unlabeled spectra. We would then store as artifacts which spectra had been labeled by humans and which had been automatically classified, possibly for use in confidence measurements when this labeling is used in conjunction with other algorithms.

4.1 A Simplistic Formalization of the Labeling Problem

In this section, we present one formalization of the labeling problem. A **spectrum** (mass spectrum) is a plot of signal intensity against the mass-to-charge (m/z) ratio of the detected ions, which can be represented as a vector

$\vec{b} = [b_1, b_2 \cdots b_r]$. $b_i \in R$ is the signal intensity at m/z value *i*. The **signature** of an ion is a vector $\vec{s} = [I_1, I_2 \cdots I_r]$, $I_i \in R$ and $\sum_i I_i = 1$, representing the distribution of its isotopes, i.e. I_i is the abundance of its isotope with m/z value *i*. A **signature database** is a set of signatures $S = \{\vec{s}_1, \vec{s}_2 \cdots \vec{s}_n\}$, in which $\vec{s}_j$ is the signature of ion *j*. For simplicity, we assume all the spectra and signatures have the same 'range' and 'granularity' over m/z axis, so that they have the same dimension r and the i[th] element of a spectrum or signature vector always corresponds to the same m/z value i.

The task of '*mass spectrum labeling*' is to find the ions identified by the peaks in the spectrum and, ideally, their quantities in the particle. Formally, a **label** of a spectrum is a set of <chemical element, quantity> tuples, which summarizes the composition of the particle described by the mass spectrum. The labeling process can also be interpreted as searching for an integral linear combination of elements that best approximates the observed spectrum. In the most ideal case, the spectrum should be the weighted sum of the chemical element signatures. That is, $\vec{b} = \sum_j w_j \vec{s}_j$, where w_j is the quantity of chemical element *j* in the particle represented by mass spectrum $\vec{b}$. This formalization of the problem is summarized in Table 1.

The formula given in Table 1 is a theoretical abstraction which gives us a tool to understand the nature of the problem, but reality is never perfect. The claim that a spectrum is the exact linear combination of element signatures is a bit unrealistic. In real applications, the observed spectrum usually contains a certain amount of noise and calibration discrepancies. What we really want is the linear combination of element signatures that approximately matches our observation.

Input:	1. An $m \times n$ matrix A=$[S_1, S_2, ..., S_n]$. The kth column S_k is the signature vector of chemical element k. It is a normalized n-dimensional vector, s.t. $\sum_i S_k[i] = 1$ 2. A normalized n-dimensional vector $\vec{b}$, s.t. $\sum_i \vec{b}[i] = 1$. It is the mass spectrum of the particle being analyzed.
Output:	1. An m-dimensional vector $\vec{x}$. The ith element of $\vec{x}$ is the proportion of chemical element i in the particle represented by spectrum $\vec{b}$
Constraint:	$A\vec{x} = \vec{b}$, $\vec{x} \geq 0$

Table 1. Labeling as a Linear Program

4.2 Distance Function for Signatures

In order to formalize labeling with approximate matches, we need a notion of closeness between signatures. Therefore, we introduce an error bound E with respect to a certain distance function *D*. The linear equation model of Table 1 then becomes an optimization task:

$$\begin{aligned} &Seek\ \vec{a},\ s.t. \\ &D(A\vec{a}, \vec{b}) < E, \quad a \geq 0 \end{aligned} \qquad (0.1)$$

Given a signature database *A* which contains *n* element signatures, and an input spectrum $\vec{b}$, the search space for the optimization task defined in (0.1) is an *n*-dimensional continuous space. The solution space for input spectrum $\vec{b}$ is a subspace within this search space. It is defined as follows:

Definition: Given a signature database *A*, an input spectrum $\vec{b}$, and an error bound *E* with respect to a distance function *D*, the **solution space** of spectrum $\vec{b}$,

$$L_{\vec{b}} = \{\vec{a} \mid D(A\vec{a}, \vec{b}) < E \text{ and } \vec{a} \geq 0\} \quad (0.2)$$

It is worth noting that the choice of the distance function *D* will significantly change the complexity of the problem. Some commonly used distance functions include Manhattan distance and Euclidean distance. The Manhattan distance between two vectors is defined as the sum of absolute difference in each dimension of the two vectors. Formally, it is defined as: $d(\vec{V}_1, \vec{V}_2) = \sum_i |\vec{V}_1[i] - \vec{V}_2[i]|$. Table 2 formalizes the labeling problem as an optimization task, using Manhattan distance.

Seek $\vec{a}, \vec{s}$ *s.t.*
$A\vec{a} - \vec{b} \leq s$
$A\vec{a} - \vec{b} \geq -s$
$\|s\| = \sum_{i=1}^{m} s_i \leq E$
$a_i \geq 0, s_i \geq 0$, for $i = 1,2,3,...,m$

Table **2**. Labeling as Optimization using Manhattan Distance

4.3 Discretization

In practice, we only care about those solutions, or labels, that are significantly different. A natural approach to deal with a continuous space is to discretize it into grids, so that the number of possible solutions is finite. We use a **threshold vector** $\vec{t} = [t_1, t_2, ..., t_{d+1}]$ to divide each dimension of the search space into *d* ranges, where t_i and t_{i+1} are the lower bound and upper bound of range *i*. Given a threshold vector, we introduce the notion of **index vector** to represent a continuous subspace.

Definition: Given a threshold vector $\vec{t} = [t_1, t_2, ..., t_{d+1}]$, an **index vector** $I = [(l_1, h_1), (l_2, h_2), ..., (l_n, h_n)], l_i < h_i, l_i, h_i \in \square$ represents a continuous subspace,

$$S_I = \{\vec{a} \mid \forall i, \vec{t}[l_i] < \vec{a}[i] < \vec{t}[h_i], \vec{a}[i] \in R\}$$

Using the *index vector* representation, we in turn define the notion of **cell.**

Definition: A subspace $[(l_1, h_1), (l_2, h_2), ..., (l_n, h_n)]$ is a **cell** if $\forall j, l_j + 1 = h_j$.

A cell is the finest granularity of the discretization, and characterizes the degree of detail users care about. A threshold vector $\vec{t} = [t_1, t_2, ..., t_{d+1}]$ divides the whole search space into d^n *cells*, where *n*, the number of dimensions is the total number of signatures in the signature database.

4.4 An Optimization-Based Formalization of the Labeling Problem

Given an error bound *E* with respect to a distance function *D* and a discretization, we now redefine the task of **spectrum labeling** as:

Find all the cells that intersect the solution space of the input spectrum

A **label** of the spectrum $\vec{b}$ is then simply an integer vector $\vec{x}$ whose corresponding cell intersects $\vec{b}$'s solution space. All the integer vectors whose corresponding cells intersect $\vec{b}$'s solution space form the **label set** of spectrum $\vec{b}$. Formally,

Definition: $\vec{x}$ is a **label** of spectrum $\vec{b}$ if the subspace defined by the *index vector* $X = [(x_1, x_1+1),(x_2, x_2+1),...,(x_n, x_n+1)]$ intersects the *solution space* of spectrum $\vec{b}$. Spectrum $\vec{b}$'s **label set** $L = \{\vec{x} \mid \vec{x} \text{ is a label of } \vec{b}\}$.

5. Interpreting ATOFMS Data: Beyond Labeling

While labeling is an important step in interpreting a mass spectrum, it is not the only step. Calibration and classification of particles require further analysis, as we discuss below.

5.1 Scaling ATOFMS to External Measurements

The ATOFMS data stream is not internally calibrated to represent the mass concentration of each element present in aerosol samples. To this end, limited experiments that seek to calibrate the ATOFMS response with chemical measurements of aerosol samples collected on filters have suggested that when the ATOFMS data is averaged over a moderate number of particles, the measurements can be directly converted to mass concentrations. Previous efforts relied on samples of aerosols collected with filter samplers that were co-located with the ATOFMS and analyzed in a laboratory for their average chemical composition (Hughes 99, Allen 00). These results were then plotted against the average ATOFMS data stream during the same time periods as the filter-based sample.

Such efforts have not been widely pursued due to the enormous resources required for manual data comparison. In many ATOFMS datasets, the ATOFMS is operated continuously for periods in excess of a month. During this extended ATOFMS operation, selected filter-based samples are often collected over periods of 4 to 24 hours and are analyzed by traditional chemical analysis techniques. With appropriate data analysis tools such as *time-series comparisons* (Keogh 01, Agrawal 93b, 95c, Chan 99, Chu 99, Faloutsos 94, 97, Loh 00, Popivanov 02, Rafiei 98, 99, Wu 00) , there is the promise that these filter-based chemical measurements can be used to calibrate the ATOFMS response and then use the calibration to obtain quantitative high-resolution chemical measurements. These calibrations can be used to extrapolate to periods where filter-based samples were not collected (Bhave 01). Data mining tools can be used to calibrate the ATOFMS and check the calibrations for datasets that were not in the training dataset.

The filter-based measurements will be comprised of a data stream that contains average mass concentrations (for specific time-periods) of individual chemical species such as sulfate ion, nitrate ion, ammonium ion, organic carbon, elemental carbon and other chemical species, and will not be available in real time. An interesting challenge in combining the filter-based and ATOFMS datasets is the fact that the collection of filter-based samples is moderately cheap but the chemical analysis of these filters is relatively costly. To this end, an optimization of the number and distribution of filter-based samples is needed as part of the data mining effort. Filter-based samples can be analyzed incrementally using data mining tools to optimize the selection of filters that are analyzed for chemical analysis.

To realize the promise of automated calibration using data mining is a nontrivial task, and involves multi-step mining. First, we need to detect peaks in mass spectra to obtain a stream of observed particle compositions. Next, we need to use time-series comparison techniques to correlate these streams with filter-based observations over time, which are likely of a very different temporal granularity, and which only include data for some of the observed particles in the ATOFMS stream. We also anticipate that frequent itemsets and sequential patterns (Srikant 96, Yang 02, Ayres 02, Shintani 98, Pei 01) extracted from the ATOFMS and filter-observation streams will be strongly correlated, leading to yet another step in the mining of the data: we can create summaries of the ATOFMS and filter-observation streams using these techniques and *then* use time-series comparisons.

Calibration also gives rise to a challenging problem where the subset-mining approach seems applicable, namely to identify those characteristics of the ATOFMS data stream and the off-line data that allow good calibration, and those that preclude good calibration.

5.2 Particle Classification and Clustering

A major goal of particle analysis is to understand the sources of aerosols in the atmosphere. Classifying or clustering particles on the basis of their composition is one approach to identifying likely sources for the particles. Since tools have been developed that use chemical measurements of filter-based aerosol samples to understand their sources, the results of different clustering algorithms can be compared to filter-based source apportionment results to understand the utility of different clustering algorithms. The project team has extensive experience in the application of chemical measurements of aerosols for source apportionment (Schauer 96, Schauer 00, Zheng 02, Schauer 02).

Prior to carrying out similar studies with single-particle mass spectrometry data, the data must be preprocessed and calibrated. Currently, the following steps are required to analyze ATOFMS data, using the software supplied by the instrument manufacturer:

1. *Generate a "Peak List" for a Dataset*: Go from a set of individual spectra to a matrix of particle information, including presence of a peak, its area, relative area, and height, as well as other information about the particle (velocity/size, date, laser power, dataset, etc.). This matrix is stored in a database. This step is currently done using simple thresholding rules for peak and noise levels, and we anticipate it will be improved by the research outlined in Section 4.1.

2. *Generate a Set of Classes*: Using a subset of the data from Step 1, based on domain knowledge, generate classes or "types" that we want to search the database for. These are generated by knowing in advance what we want to look for (e.g., Calcium-containing particles, elemental carbon particles, organic carbon particles, sulfate-containing particles, etc.), and then defining a query (in terms of the thresholding heuristics used for peak detection) that retrieves (mostly) the intended particles. Challenges arise because of isobars such as those described above. This approach is reasonably effective, but is limiting and slow.

3. *Apply the Class Definitions to the Dataset*: Finally, the query is run on the entire dataset from Step 1. This tags each particle as a member of whichever class(es) it matches. Particles often match many classes, given the way we define classes. This is not a problem in most cases (e.g., it is appropriate if a particle that contains calcium sulfate is picked up by a "calcium-containing" class as well as a "sulfate-containing" class).

The obvious data mining challenge here is to apply classification or clustering algorithms to automate the process; Step (2) would be replaced by the training phase, if a learning algorithm (Berry 99, Fayyad 96, Han 00, Hand 00, Hastie 01, Rud 01, Weiss 97, Witten 99) is used. The major challenge, that of incorporating domain knowledge into the appropriate data mining algorithms, has been studied to some degree in the contexts of both supervised and unsupervised learning, (Towell 94, Fung 01, Clark 94, Zhou 01, Padmanabhan 98, 00, Cook 96, Clair 98). While there has been some prior work on clustering particles (Fergenson 01, Phares 01, Hinz 99, Tan 02, Bhave 01, 02, Song 99) it has not been entirely satisfactory, possibly because they largely use ART-2a and similar approaches (Hertz 91, Freeman 91, Carpenter 91), which do not take into account domain knowledge. There are, on the other hand, approaches that allow one to incorporate domain knowledge into clustering algorithms (Clark 94, Béjar 97, Wagstaff 00, Tan 02). Such techniques would allow users to "guide" clustering algorithms towards categories of interest. Finally, variations of scalable clustering algorithms might lend themselves to online clustering in order to provide information on classes found while the experiment is running (Zhang 97, Bradley 98b, Guha 98, Huang 97).

Our emphasis, which will be different from that found in the previous works, will be in enhancing and adapting such techniques to be usable in an interactive fashion. We will therefore also be studying how to *incrementally update* any classifiers that we find as the user changes the knowledge base. The implications of this analysis will be relevant beyond this particular application area, as the types of data mining problems that we will be considering (clustering, identifying prototypes, semi-supervised learning, change point detection, feature selection) all have much wider applicability. Keeping track of who enters new knowledge, and when, will be necessary in determining provenance.

There is a large body of literature on support vector machines (SVMs), which will likely be one of the algorithms used (Vapnik 95, Burges 98, Cristianini 00). The project team has considerable expertise in modifying SVMs to serve particular purposes or to run more quickly (Musicant 99, 00, 01a, 01b, 01c, 02). Many of these new algorithms have been released as software packages that are freely available to the research community (Musicant LSVM, ASVM, ASVR).

6. Using ATOFMS Data to Understand Aerosol Dynamics

Ambient datasets, where the ATOFMS instrument monitored atmospheric aerosol particles continuously over a period of time, have already been collected at several locations by the project team and we anticipate collecting many similar datasets in the near future. Examples include:

- **Northfield, MN**: Ambient particles were monitored continuously from early May through the beginning of July, 2002, to provide a dataset that includes the changes in the particle population as the ground thawed and agricultural activities increased. The thawed ground was tilled and planted and herbicides were applied both in the immediate environment and in the surrounding area. Approximately 33,000 particles are included in this dataset.
- **Atlanta, GA**: Ambient particles were monitored continuously from July 21 through August 30 2002 as part of the Atlanta Aerosol Nucleation and Real-time Characterization Experiment (ANARChE). Approximately 536,000 particles are included in this dataset.

The current capability to analyze this data to detect trends, internal structure, and correlations is significantly limited. In this section, we describe three broad classes of problems for which effective mining techniques are sorely needed.

6.1 Evolution of ATOFMS Streams

An important class of questions concerns trends observed in ATOFMS streams. Changes in the relative population of the multiple "particle types", or the appearance of a new "type" would indicate that there had been a change in the composition of the particles at some time. Such changes in the data stream could arise for different reasons, such as:

- Changes in gas-phase concentration of a species that reacts with chemical components of the particles, thus changing their chemical composition. Note that this will provide particles with a whole range of concentrations of "reactant" and "product", going from all/mostly "reactant" to all/mostly "product" through a mixture of the two; it is not just a binary only "reactant" vs. only "product" situation. Thus, information about "how much" of something is there in addition to "what" is there is required, and the results from Section 4.2 will play an important role, in addition to the (obviously important) results from Sections 4.1. An example of the influence of gas-phase concentrations can be found in Gard, 1998.
- Changes in temperature that influence the partitioning of species onto particles rather than into the gas phase. Again, quantitative particle compositions are required for effective analysis.
- Transient engine emissions, where an engine runs at a constant condition and is then abruptly changed to a new condition, thus changing the emissions.
- Changes in the wind direction, bringing particles to the sampling site from a different source or set of sources, which are geographically distributed.
- Introduction of a new source of particle emissions into the area (e.g., turning something on).
- Removal of a source of emissions from the area (turning something off).

The data mining challenge here is two-fold. First, we must be able to extend the results of Section 4.1, 4.2 and 4.3 on detecting particle compositions to detect changes (and trends) in the mixture of various particle "types" in the ATOFMS data stream. Second, we must be able to correlate these changes with other environmental monitoring data and detect what other monitored parameters change at the same/similar time, and subsequently monitor this correlation.

The current approach is essentially manual, and leaves much room for improvement: For ambient datasets, once the data is classified into "types" the number of each "type" in every 30 minute (or appropriate time) interval is counted and graphed versus time. These graphs are then overlaid with each other to search for correlations.

Detecting when a time series has shifted in some fundamental way is known as the "change point detection" problem, and there is much literature devoted to the topic (Keogh 01). Dynamic Time Warping and other techniques aim to abstract out the main features of a time series (Berndt 96, Keogh 02, Das 98). If each 30 minutes of data is considered to be a point in time, as suggested above, these methods or similar ones would likely identify in an automatic manner the times in which something has changed. Another aspect of this problem is in determining *what aspect* of the time series has changed, which is a subset mining problem. Each time series pattern, or artifact, will need to be archived for further analysis. Finally, we may also use domain knowledge to focus on specific kinds of time-series changes.

6.2 Comparison of Different ATOFMS Streams

There are many cases when we need to compare (subsets of ATOFMS data streams that reflect) "now" and "then", or "here" and "there", or different underlying conditions:

- ATOFMS instruments have sampled in the same location at different times (Atlanta, GA in Summer 1999 and 2002; Caldecott Tunnel, Berkeley, CA in Nov. 1997 and July 2000; Northfield, MN Spring 2002 and Summer 2002, and more to come) for many different studies. A direct comparison of these studies would easily pinpoint the differences and similarities.
- Cross-location comparisons, for example a comparison of Northfield, MN with Minneapolis, MN or Atlanta, GA could help identify rural versus urban sources and other sorts of differences.
- Diesel engine emissions were studied by monitoring particulate emissions from an engine operating at different engine conditions (i.e., speed, load, and timing) and with different diesels fuels (i.e., proposed low emissions diesel fuels). A comparison of the emissions in each case would directly pinpoint differences and similarities between different operating conditions and fuels.

Currently, aerosol scientists analyze each dataset independently and then manually (statistically) compare those sets which we think will elucidate the differences between the two cases of interest. A simple comparison can be made by subtracting a summary of one dataset from a summary of the other, for example subtracting a dataset taken "there" from one taken "here." In the resulting difference, positive-going signal will indicate that a feature is more common in one dataset (e.g. "here") while negative-going signal indicates that a feature is more common in the other (e.g. "there").

It is also possible to think about this question within a single dataset if we want to compare the composition of particles greater than 1.0 µm to those less than 1.0 µm, for example, or those sampled before versus after a particular event. Further, in addition to comparing specific subsets, it would be remarkably useful to have attention drawn to subsets where there is an interesting correlation, which domain scientists may not have hitherto suspected. Clearly, there is a need here to summarize ATOFMS and environmental datasets using a range of techniques and to study correlations, between specific subsets using known time-series techniques, and across a range of potentially interesting subsets, which is an instance of subset mining.

6.3 Associations within the Stream

Looking for correlations and anti-correlations between particle types is one of the most important ways scientists analyze aerosol data. If we see a correlation or an anti-correlation between the presence of certain peaks in the mass spectra and another measured parameter (e.g., relative humidity, NOx concentration, wind direction, another particle class), we can learn a great deal about the underlying chemistry or aerosol sources.

As a specific example, if we use a clustering algorithm that generates a large number of particle clusters, there is potentially a lot of similarity between clusters (e.g., some might even contain the same peaks but differ in ratios of peaks or other subtle properties). Looking for correlations and anti-correlations between clusters could be extremely useful in determining which classes are indeed related or where there is chemistry going on.

The current approach is to make graphs of the population of particles in various “types” (i.e., clusters) as a function of time, and manually look for correlations. Statistical methods are also used to some extent, but there is no way to automatically compare each “type” against every other “type” to test for all possible correlations. This is yet another instance of subset mining, and we expect it to be very useful in finding things that we did not already know to look for.

The problem of finding correlation or anti-correlation between presence of certain peaks in the mass spectra and other measured parameters can be thought of as a traditional feature selection problem. In other words, the goal is to determine which features (measured parameters) have a predictive effect on peaks. Support vector machines and other algorithms designed to handle feature selection might be particularly appropriate here (Bradley 98, Weston 00). Due to the fact that a given spectrum has multiple characteristics, multiple output artificial neural networks might be a strong choice (Fausett 94). Prior knowledge can be integrated if available (Towell 94, Fung 01). In neural networks (Towell 94), prior knowledge is typically expressed as "if-then" implications which are appropriately integrated as part of the structure of the network. Support vector machines (Fung 01) allow the user to specify a specific subset of the input space where the classification is believed to be known. We may need to build on these ideas and develop new versions of knowledge-based algorithms to handle the kind of domain knowledge that we have.

This stream correlation problem becomes an instance of subset pattern mining, as noted above, when patterns to be found differ among different subsets of particles.

7. Fusion of ATOFMS Stream with Other Data Streams

Some tasks require ATOFMS data to be analyzed in light of external data. We discuss two examples of such tasks.

7.1 Understanding the Sources of Aerosols

The fusion of ATOFMS data with meteorological data and data on the concentration of gas-phase pollutants (e.g., ozone, carbon monoxide, sulfur dioxide, and nitrogen oxides) can be used to locate the sources of different aerosols. Clustering techniques that presort, or partition, ATOFMS measurements by meteorological or gas-phase pollutant data can improve our understanding of aerosol sources. The presorting is clear-cut in only very limited cases. Such cases include the wind blowing directly from a chemical manufacturing facility, or high sulfur dioxide levels that can be shown to directly result from a single power plant in some locations. In contrast, a broad range of source emissions and meteorological conditions can effect ozone concentrations in such a way that there is no well-defined categorization of the ozone concentrations in the context of aerosol sources. Likewise, high wind speeds will lead to high dispersion of pollutants and will tend to decrease concentrations, which can be counteracted by increase in aerosol concentrations that can result from wind blown dust. To this end, wind speed data can be used as an important parameter to understand the sources of particulate matter in the atmosphere.

This task can be viewed as a supervised learning problem where the goal is to use the features in one stream (meteorological data and gas-phase pollutants) to predict the features in the other stream (aerosol data). The data points of interest are aggregates over a variety of parameters for an appropriate amount of time. Determining which features are the important ones is then a feature selection problem. Algorithms such as support vector machines, neural networks, decision trees, and others would all seem to be relevant here, and versions of these algorithms that incorporate domain knowledge can certainly be used. These techniques are somewhat robust to noise, and as such we may be able to use time-bins of shorter length than 30 minutes. This approach will allow us to verify specific correlations that domain scientists suspect to exist. To detect emergent and unsuspected correlations, a complementary approach is to mine each data source to identify interesting phenomena, such as high wind speeds, gas-phase pollutant concentrations, etc., and to formulate the problem of finding interesting correlations as an instance of subset mining.

7.2 Understanding Optical and Chemical Properties of Aerosols

Carbonaceous aerosols absorb light, which impacts global climate change and visibility degradation. Due to the fact that carbonaceous aerosols are comprised of a complex mixture of hundreds of organic compounds and elemental carbon (i.e., soot), the ability to relate the chemical composition of the carbonaceous fraction of aerosols with their light absorption properties has proven to be an intractable problem using traditional techniques to characterize carbonaceous aerosols (Seinfeld 98). Since the ATOFMS provides a data stream that provides a less specific measurement of organic and elemental carbon that relates to the bulk chemical properties of the carbon containing species, there is potential for the ATOFMS to help understand the chemical properties of aerosols that control the light absorbing properties of the carbonaceous aerosols. Through data mining tools, we hope to isolate the ATOFMS characteristics that best correlate with the light absorbing properties of the aerosols.

As another example, the distribution of semi-volatile compounds (compounds that exist simultaneously in the gas and particle phase) is controlled by the atmospheric temperature and the chemical composition of aerosols (Seinfeld 98). To this end, there is a great interest in understanding the properties of aerosols that effect this distribution of contaminants between the gas and particle-phases (Seinfeld 98, Schauer 03). It is important to recognize, however, that the impact of different chemical species in the aerosol phase on the gas/particle partitioning is not a linear sum of the contributing species. To this end, a feedback clustering algorithm could be used to see if different types of ATOFMS clusters would internally have correlations with gas/particle portioning. The clustering approaches would be hypothesis-driven, based on known chemical interactions. As an example, mercury is known to have a strong affinity for elemental carbon and sulfur compounds. Thus, we would be interested to know if the mercury partitioning to particles with high elemental carbon varied with sulfur content.

8. Conclusions

In this paper, we described the EDAM project, which is a collaboration between atmospheric aerosol researchers and computer scientists at Carleton College and University of Wisconsin-Madison. The project is addressing data mining challenges with broad applicability, but with specific instances grounded in the domain of atmospheric aerosol analysis. A novel aspect of the data involved here is complex streams of mass spectra, which must be interpreted and analyzed in conjunction with other data, leading to several challenging problems.

(Agrawal 93) R. Agrawal, T. Imielinski, A. Swami: Mining Associations between Sets of Items in Massive Databases, Proc. of the ACM-SIGMOD 1993 Intl Conference on Management of Data, Washington D.C., May 1993, 207-216.

(Agrawal 93b) Agrawal, R., Faloutsos, C. & Swami, A. (1993). Efficient similarity search in sequence databases. In proceedings of the 4th Int'l Conference on Foundations of Data Organization and Algorithms. Chicago, IL, Oct 13-15. pp 69-84.

(Agrawal 94) R. Agrawal, R. Srikant: "Fast Algorithms for Mining Association Rules", Proc. of the 20th Int'l Conference on Very Large Databases, Santiago, Chile, Sept. 1994. Expanded version available as IBM Research Report RJ9839, June 1994.

(Agrawal 95a) R. Agrawal, H. Mannila, R. Srikant, H. Toivonen and A. I. Verkamo: "Fast Discovery of Association Rules", Advances in Knowledge Discovery and Data Mining, Chapter 12, AAAI/MIT Press, 1995.

(Agrawal 95b) Rakesh Agrawal and Ramakrishnan Srikant. Mining Sequential Patterns. In Proc. of the 11th Int'l Conference on Data Engineering, Taipei, Taiwan, March 1995.

(Agrawal 95c) Agrawal, R., Lin, K. I., Sawhney, H. S. & Shim, K. (1995). Fast similarity search in the presence of noise, scaling, and translation in time-series databases. In proceedings of the 21st Int'l Conference on Very Large Databases. Zurich, Switzerland, Sept. pp 490-50.

(Allen 00) Jonathan O. Allen, David P. Fergenson, Eric A. Gard, Lara S. Hughes, Bradley D. Morrical, Michael J. Kleeman, Deborah S. Gross, Markus E. Gaelli, Kimberly A. Prather, Glen R. Cass,. Particle Detection Efficiencies of Aerosol Time-of-Flight Mass Spectrometers Under Ambient Sampling Conditions, Environ. Sci. Technol. 2000, 34, 211-217.

(Arasu 02) Arvind Arasu, Brian Babcock, Shivnath Babu, Jon McAlister, Jennifer Widom, Characterizing Memory Requirements for Queries over Continuous Data Streams. PODS 2002: 221-232.

(Ayres 02) Jay Ayres, J. E. Gehrke, Tomi Yiu, and Jason Flannick. Sequential PAttern Mining Using Bitmaps. In Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Edmonton, Alberta, Canada, July 2002.

(Babcock 02) Brian Babcock, Shivnath Babu, Mayur Datar, Rajeev Motwani, Jennifer Widom, Models and Issues in Data Stream Systems. PODS 2002: 1-16.

(Babu 01) Shivnath Babu, Jennifer Widom, Continuous Queries over Data Streams. SIGMOD Record 30(3): 109-120 (2001).

(Baker 97) J. E. Baker, Atmospheric Deposition of Contaminants to the Great Lakes and Costal Water, 1997, SETAC Press, Pensacola, FL.

(Basu 02) Sugato Basu, Arindam Banerjee and Raymond J. Mooney: Semi-supervised Clustering by Seeding. Proceedings of the Nineteenth International Conference on Machine Learning (ICML-2002), pp. 19-26, Sydney, Australia, July 2002.

(Bayardo 98) R. J. Bayardo. Efficiently mining long patterns from databases. SIGMOD 98, 85-93, Seattle, Washington.

(Béjar 97) J. Béjar, U. Cortés, R. Sangüesa, M. Poch "Experiments with Domain Knowledge in Knowledge Discovery" Proceedings of the 1st International Conference on The Practical Aplication of Knowledge Discovery and Data Mining (PADD97). London (UK), 65-78, 1997.

(Bennett 98) Kristin P. Bennett and Ayhan Demiriz: Semi-Supervised Support Vector Machines. Advances in Neural Information Processing Systems, 12, M. S. Kearns, S. A. Solla, D. A. Cohn, editors, MIT Press, Cambridge, MA, 1998, pp 368-374.

(Berndt 96) Donald J. Berndt and James Clifford: Finding Patterns in Time Series: A Dynamic Programming Approach. In Advances in Knowledge Discovery and Data Mining, Usama M. Fayyad, Gregory Piatetsky-Shapiro, Padhraic Smyth, Ramasamy Uthurusamy (Eds.), 1996, 229-248.

(Bhave 01) Prakash V. Bhave, Jonathan O. Allen, Bradley D. Morrical, David P. Fergenson, Glen R. Cass, Kimberly A. Prather "A Field-Based Approach for Determining ATOFMS Instrument Sensitivities to Ammonium and Nitrate" Environ. Sci. Technol. 2002, 36, 4868-4879.

(Bhave 02) Prakash V. Bhave, David P. Fergenson, Kimberly A. Prather, Glen R. Cass "Source Apportionment of Fine Particulate Matter by Clustering Single-Particle Data: Tests of Receptor Model Accuracy" Environ. Sci. Technol. 2001, 35, 2060-2072.

(Box 94) G. Box, G. Jenkins, and G. Reinsel. Time Series Analysis: Forecasting and Control. Prentice Hall, Englewood Cliffs, NJ, 1994. 3rd Edition.

(Berry 99) Michael J. A. Berry and Gordon Linoff: Mastering Data Mining. John Wiley & Sons, 1999.

(Blum 01) Avrim Blum and Shuchi Chawla: Learning from Labeled and Unlabeled Data using Graph Mincuts. ICML 2001.

(Bonnet 01) Philippe Bonnet, Johannes Gehrke, Praveen Seshadri, Towards Sensor Database Systems. Mobile Data Management 2001: 3-14.

(Bradley 98) P. S. Bradley & O. L. Mangasarian: Feature Selection via Concave Minimization and Support Vector Machines. Proceedings of the Fifteenth International Conference, J. Shavlik, editor, Morgan Kaufmann, San Francisco, California, 82-90, 1998.

(Bradley 98b) P.S. Bradley, Usama Fayyad, Cory Reina: Scaling Clustering Algorithms to Large Databases. Knowledge Discovery and Data Mining, 9-15 (1998).

(Bradley 02) Paul S. Bradley, Johannes Gehrke, Raghu Ramakrishnan, Ramakrishnan Srikant, Scaling mining algorithms to large databases. CACM 45(8): 38-43 (2002).

(Brin 97) S. Brin, R. Motwani, and C. Silverstein. Beyond market basket: Generalizing association rules to correlations. SIGMOD 97, 265-276, Tucson, Arizona.

(Brin 97a) S. Brin R. Motwani, J. Ullman, and S. Tsur. Dynamic itemset counting and implication rules for market basket data. In SIGMOD 97.

(Burges 98) C.J.C. Burges: A Tutorial on Support Vector Machines for Pattern Recognition. Data Mining and Knowledge Discovery, 2:2, pp. 121-167, 1998.

(Carney 02) D. Carney, U. Çetintemel, M. Cherniack, C. Convey, S. Lee, G. Seidman, M. Stonebraker, N. Tatbul, S. Zdonik, Monitoring Streams - A New Class of Data Management Applications. VLDB 2002.

(Carpenter 91) Gail A. Carpenter, Stephen Grossberg, and D.B. Rosen. Art2-a: An adaptive resonance algorithm for rapid category learning and recognition. Neural Networks, 4:493-504, 1991.

(Chan 99) Chan, K. & Fu, A. W. (1999). Efficient time series matching by wavelets. In proceedings of the 15th IEEE Int'l Conference on Data Engineering. Sydney, Australia, Mar 23-26. pp 126-133.

(Chakrabarti 02) Kaushik Chakrabarti, Eamonn J. Keogh, Sharad Mehrotra, Michael J. Pazzani, Locally adaptive dimensionality reduction for indexing large time series databases. TODS 27(2): 188-228 (2002).

(Chandrasekharan 03) Sirish Chandrasekaran, Owen Cooper, Amol Deshpande, Michael J. Franklin, Joseph M. Hellerstein, Wei Hong, Sailesh Krishnamurthy, Samuel R. Madden, Vijayshankar Raman, Fred Reiss, and Mehul A. Shah, TelegraphCQ: Continuous Dataflow Processing for an Uncertain World. CIDR 2003.

(Chen et al. 2003) Chen, L., Huang, Z., Ramakrishnan, R., et al., Cost-Based Labeling of Groups of Mass Spectra, submitted for publication, University of Wisconsin-Madison.

(Cherkassky 98) V. Cherkassky and F. Mulier: Learning from Data - Concepts, Theory and Methods. Wiley, 1998.

(Cheung 96) D.W. Cheung, J. Han, V. Ng, and C.Y. Wong. Maintenance of discovered association rules in large databases: An incremental updating technique. ICDE 96, New Orleans, LA.

(Chow 95) J. C. Chow, “Critical Review – Measurement Methods to Determine Compliance with Air Quality Standards for Suspended Particle” J. Air & Waste Manage., 1995, 320-382.

(Chu 99) Chu, K. & Wong, M. (1999). Fast time-series searching with scaling and shifting. In proceedings of the 18th ACM Symposium on Principles of Database Systems. Philadelphia, PA, May 31-Jun 2. pp 237-248.

(Chvatal 83) Vasek Chvatal: Linear Programming. W H Freeman & Co., 1983.

(Clair 98) C. Clair, C. Liu and N. Pissinou, "Attribute weighting: a method of applying domain knowledge in the decision tree process," The Seventh International Conference on Information and Knowledge Management, 1998, pp. 259-266.

(Clark 94) M. Clark, L. Hall, C. Li, and D. Goldgof: Knowledge based (re-)clustering. In 12th IAPR International Conference on Pattern Recognition, Jerusalem, Israel, 1994.

(Cook 96) Cook, D. J., Holder, L. B. and Djoko, S. "Scalable Discovery of Informative Structural Concepts Using Domain Knowledge", IEEE Expert, 11(5), 1996.

(Cormode 02) Graham Cormode, Mayur Datar, Piotr Indyk, S. Muthukrishnan, Comparing Data Streams Using Hamming Norms (How to Zero In). VLDB 2002.

(Cortes 00) C. Cortes, K. Fisher, D. Pregibon, and A. Rogers. Hancock: A Language for Extracting Signatures from Data Streams. In Proceedings of the Association for Computing Machinery Sixth International Conference on Knowledge Discovery and Data Mining, pages 9-17, 2000.

(Cristianini 00) N. Cristianini and J. Shawe-Taylor: An Introduction to Support Vector Machines, Cambridge University Press, 2000.

(Das 98) G. Das, K.-I. Lin, H. Manilla, G. Renganathan, and P. Smyth: Rule Discovery from Time Series. In Proc. of KDD '98, Aug 1998.

(Datar 02) Mayur Datar, Aristides Gionis, Piotr Indyk, Rajeev Motwani, Maintaining stream statistics over sliding windows. SODA 2002: 635-644.

(DeCoste 99) Dennis DeCoste.Recent Advances in SMO Speed and Accuracy, NIPS99 Workshop on Learning with Support Vectors, December 1999.

(Donjerkovic 00) , Yannis E. Ioannidis, Raghu Ramakrishnan, Dynamic Histograms: Capturing Evolving Data Sets. ICDE 2000: 86.

(Ester 98) M. Ester, H.-P. Kriegel, J. Sander, M. Wimmer, X.Xu: Incremental Clustering for Mining in a Data Warehousing Environment. Proc. 24th Int. Conf. on Very Large Data Bases, New York, 1998, pp. 323-333.

(Faloutsos 94) C. Faloutsos, M. Ranganthan, and Y. Manolopoulos. Fast subsequence matching in time-series databases. Proc. ACM SIGMOD, 419-429, May 1994.

(Faloutsos 97) Faloutsos, C., Jagadish, H., Mendelzon, A. & Milo, T. (1997). A signature technique for similarity-based queries. In proceedings of the Int'l Conference on Compression and Complexity of Sequences. Positano-Salerno, Italy, Jun 11-13.

(Faloutsos 02) Christos Faloutsos, Sensor Data Mining: Similarity Search and Pattern Analysis. VLDB 2002.

(Faloutsos 02a) Christos Faloutsos, Future directions in data mining: streams, networks, self-similarity and power laws. CIKM 2002: 93.

(Fayyad 96) Usama M. Fayyad, Gregory Piatetsky-Shapiro, Padhraic Smyth, and Ramasamy Uthurusamy: Advances in Knowledge Discovery and Data Mining. AAAI Press, 1996.

(Fausett 94) Laurene V. Fausett. Fundamentals of Neural Networks. Prentice-Hall, 1994.

(Fergenson 01) David P. Fergenson, Xin-Hua Song., Ziad Ramadan, Jonathan O. Allen, Lara S. Hughes, Glen R. Cass, Philip K. Hopke, and Kimberly A. Prather "Quantification of ATOFMS Data by Multivariate Methods" Anal. Chem. 2001, 73, 3535-3541.

(Freeman 91) James A. Freeman and David M. Skapura. Neural Networks: Algorithms, Applications and Programming Techniques. Addison-Wesley, 1991.

(Fukuda 96) T. Fukuda, Y. Morimoto, S. Morishita, and T. Tokuyama. Data mining using two-dimensional optimized association rules: Scheme, algorithms, and visualization. SIGMOD 96, Montreal, Canada.

(Fung 99) Glenn Fung and O.L. Mangasarian: Semi-Supervised Support Vector Machines for Unlabeled Data Classification. Optimization Methods and Software 15, 2001, 29-44.

(Fung 01) Glenn Fung, O.L. Mangasarian, and Jude Shavlik: Knowledge-Based Support Vector Machine Classifiers. Neural Information Processing Systems 2002 (NIPS 2002), Vancouver, BC, December 10-12, 2002.

(Fung 01b) Glenn Fung and O. L. Mangasarian: Proximal Support Vector Machine Classifiers. Proceedings KDD-2001, San Francisco August 26-29, 2001. Association for Computing Machinery, New York, 2001, 77-86.

(Ganti 99) Venkatesh Ganti, Johannes Gehrke, Raghu Ramakrishnan: Mining Very Large Databases. IEEE Computer 32(8): 38-45 (1999).

(Ganti 00) Venkatesh Ganti, Johannes Gehrke, Raghu Ramakrishnan. DEMON: Mining and Monitoring Evolving Data., in ICDE 2000: 439-448, San Diego, CA.

(Ganti 01) Venkatesh Ganti, Johannes Gehrke, Raghu Ramakrishnan, DEMON: Mining and Monitoring Evolving Data. TKDE 13(1): 50-63 (2001).

(Ganti 02) Venkatesh Ganti, Johannes Gehrke, Raghu Ramakrishnan, Wei-Yin Loh, A Framework for Measuring Differences in Data Characteristics. JCSS 64(3): 542-578 (2002).

(Gard 97) E. Gard, J. E. Mayer, B. D. Morrical, T. Dienes, D. P. Fergenson, K. A. Prather. "Real-Time Analysis of Individual Atmospheric Aerosol Particles: Design and Performance of a Portable ATOFMS" Anal. Chem. 1997, 69, 4083-4091.

(Gard 98) E. E. Gard, M. J. Kleeman, D. S. Gross, L. S. Hughes, J. O. Allen, B. D. Morrical, D. P. Fergenson, T. Dienes, M. Gaelli, G. R. Cass, K. A. Prather, "Direct Observation of Heterogeneous Chemistry in the Atmosphere" Science 1998, 279, 1184-1187.

(Garofalakis 02) Minos N. Garofalakis, Johannes Gehrke, Querying and Mining Data Streams: You Only Get One Look. VLDB 2002.

(Ge 00) Xianping Ge, Padhraic Smyth, Deformable Markov model templates for time-series pattern matching. KDD 2000: 81-90.

(Gehrke 99) Johannes Gehrke, Venkatesh Ganti, Raghu Ramakrishnan, Wei-Yin Loh: BOAT-Optimistic Decision Tree Construction. SIGMOD Conference 1999: 169-180

(Gehrke 00) Johannes Gehrke, Raghu Ramakrishnan, Venkatesh Ganti: RainForest - A Framework for Fast Decision Tree Construction of Large Datasets. Data Mining and Knowledge Discovery 4(2/3): 127-162 (2000)

(Goldman 00) Sally Goldman and Yan Zhou: Enhancing Supervised Learning with Unlabeled Data. ICML 2000.

(Grahne 00) G. Grahne, L. Lakshmanan, and X. Wang. Efficient mining of constrained correlated sets. ICDE 00, 512-521, San Diego, CA, Feb. 2000.

(Guha 98) S. Guha, R. Rastogi and K. Shim: CURE: An efficient algorithm for clustering large databases. Proceedings of ACM-SIGMOD 1998 International Conference on Management of Data, Seattle, 1998.

(Han 95) J. Han and Y. Fu. Discovery of multiple-level association rules from large databases. VLDB 95, 420-431, Zurich, Switzerland.

(Han 97) E.-H. Han, G. Karypis, and V. Kumar. Scalable parallel data mining for association rules. SIGMOD 97, Tucson, Arizona.

(Han 00) Jiawei Han, Micheline Kamber: Data Mining : Concepts and Techniques. Morgan Kaufmann, 2000.

(Hand 00) David J. Hand, Heikki Mannila and Padhraic Smyth: Principles of Data Mining. MIT Press, Fall 2000

(Hastie 01) Trevor Hastie, Robert Tibshirani, Jerome Friedman: The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Springer Verlag, 2001.

(Hertz 91) John Hertz, Anders Krogh, and Richard G. Palmer. Introduction to the Theory of Neural Computation, volume 1 of Santa Fe Institute Studies In The Sciences of Complexity Lecture Notes. Addison-Wesley, 1991.

(Hinz 99) K.-P. Hinz, M. Greweling, F. Drews, B. Spengler "Data Processing in On-line Laser Mass Spectrometry of Inorganic, Organic, or Biological Airborne Particles" J. Am. Soc. Mass Spectrom. 1999, 10, 648-660.

(Himberg 01) Johan Himberg, Kalle Korpiaho, Heikki Mannila, Johanna Tikanmäki, Hannu Toivonen, Time Series Segmentation for Context Recognition in Mobile Devices. ICDM 2001: 203-210.

(Hopke 85) P. K. Hopke Receptor Modeling in Environmental Chemistry. 1985, Wiley, New York.

(Huang 97) Zhexue Huang: A Fast Clustering Algorithm to Cluster Very Large Categorical Data Sets in Data Mining. Proc. SIGMOD Workshop on Research Issues on Data Mining and Knowledge Discovery, 1997.

(Huang 03) Huang, Z., Chen, L., Ramakrishnan, R., et al., Algorithms for Labeling a Mass Spectrum, submitted for publication, University of Wisconsin-Madison.

(Hughes 99) L. S. Hughes, J. O. Allen, M .J. Kleeman, R. J. Johnson, G. R. Cass, D. S. Gross, E. E. Gard, M. E. Galli, B. D. Morrical, D. P. Fergenson, T. Dienes, C. A. Noble, D. Y. Liu, P. J. Silva, and K. A. Prather "Size and Composition Distribution of Atmospheric Particles in Southern California" Environ. Sci. Technol., 1999, 3506-3515.

(Hulten 01) Geoff Hulten, Laurie Spencer, Pedro Domingos, Mining time-changing data streams. KDD 2001: 97-106.

(IPCC 96) IPCC, Climate change 1995: The science of climate change, 1996, Cambridge University Press, New York, NY.

(Imielinski 00) T. Imielinski, L. Khachiyan, and A. Abdulghani. Cubegrades: Generalizing association rules. Technical Report, Aug. 2000.

(Joachims 99) T. Joachims, Making Large-Scale SVM Learning Practical. In: Advances in Kernel Methods - Support Vector Learning, B. Schölkopf, C. Burges, and A. Smola (ed.), MIT Press, 1999.

(Joshi 00) A. Joshi and R. Krishnapuram. On Mining Web Access Logs. In Proceedings of the 2000 ACM SIGMOD Workshop on Research Issues in Data Mining and Knowledge Discovery 2000, pp. 63-69, 2000.

(Kamber 97) M. Kamber, J. Han, and J. Y. Chiang. Metarule-guided mining of multi-dimensional association rules using data cubes. KDD 97, 207-210, Newport Beach, California.

(Keogh 01) Keogh, E., Chu, S., Hart, D. and Pazzani, M: An Online Algorithm for Segmenting Time Series. In Proceedings of IEEE International Conference on Data Mining. pp 289-296, 2001.

(Keogh 02) Keogh, E., Lonardi, S and Chiu, W. (2002). Finding Surprising Patterns in a Time Series Database In Linear Time and Space. In the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. July 23 - 26, 2002. Edmonton, Alberta, Canada. pp 550-556.

(Klemettinen 94) M. Klemettinen, H. Mannila, P. Ronkainen, H. Toivonen, and A.I. Verkamo. Finding interesting rules from large sets of discovered association rules. CIKM 94, 401-408, Gaithersburg, Maryland.

(Klosgen 96) W. Klosgen, Explora: A multistrategy and multipattern discovery assistant. In Fayyad et al., Advances in Knowledge Discovery and Data Mining, AAAI/ MIT Press, 1996.

(Korn 98) F. Korn, A. Labrinidis, Y. Kotidis, and C. Faloutsos. Ratio rules: A new paradigm for fast, quantifiable data mining. VLDB 98.

(Lakshmanan 99) L. V. S. Lakshmanan, R. Ng, J. Han and A. Pang, Optimization of Constrained Frequent Set Queries with 2-Variable Constraints, SIGMOD 99.

(Landis 02) M. S. Landis, R. K. Stevens, F. Schaedlich, and E. M. Prestbo "Development and characterization of an annular denuder methodology for the measurement of divalent inorganic reactive gaseous mercury in ambient air" Environmental Science and Technology, 2002, 3000-3009.

(Lavrac 94) N. Lavrac and S. Dzeroski. Inductive Logic Programming: Techniques and Applications. Ellis Horwood, Chichester, 1994.

(Lavrac 02) N. Lavrac, F. Zelezny, P. Flach. RSD: Relational subgroup discovery through first-order feature construction. In 12th Intl. Conf. on Inductive Logic Programming. Springer-Verlag, 2002.

(Lee 98) Wenke Lee and Sal Stolfo. ``Data Mining Approaches for Intrusion Detection'' In Proceedings of the Seventh USENIX Security Symposium (SECURITY '98), San Antonio, TX, January 1998.

(Lee 01) Yuh-Jye Lee and O. L. Mangasarian: RSVM: Reduced Support Vector Machines. Proceedings of the SIAM International Conference on Data Mining, Chicago, April 5-7, 2001.

(Lent 97) B. Lent, A. Swami, and J. Widom. Clustering association rules. ICDE 97, 220-231, Birmingham, England.

(Letouzey 00) F. Letouzey, F. Denis and R. Gilleron: Learning From Positive and Unlabeled Examples. Eleventh International Conference on Algorithmic Learning Theory , Lecture Notes in Artificial Intelligence, 71 – 85, 2000.

(Loh 00) Loh, W., Kim, S. & Whang, K. (2000). Index interpolation: an approach to subsequence matching supporting normalization transform in time-series databases. In proceedings of the 9th ACM CIKM Int'l Conference on Information and Knowledge Management. McLean, VA, Nov 6-11. pp 480-487.

(Lu 00) H. Lu, L. Feng, and J. Han, "Beyond Intra-Transaction Association Analysis:Mining Multi-Dimensional Inter-Transaction Association Rules", ACM Transactions on Information Systems (TOIS'00), 18(4): 423-454, 2000.

(Madden 02) S. Madden, M. Franklin J. Hellerstein, and W. Hong, TAG: a Tiny AGgregation Service for Ad-Hoc Sensor Networks, OSDI 2002

(Madden 02a): S. Madden, J. Hellerstein, and M. Shah, Continuously Adaptive Continuous Queries over Streams, SIGMOD 2002.

(Manku 98) Gurmeet Singh Manku, Sridhar Rajagopalan, Bruce G. Lindsay, Approximate Medians and other Quantiles in One Pass and with Limited Memory. SIGMOD Conference 1998: 426-435.

(Manku 02) Gurmeet Singh Manku, Rajeev Motwani, Approximate Frequency Counts over Data Streams. VLDB 2002.

(Mannila 94) Heikki Mannila, Hannu Toivonen, A. Inkeri Verkamo: Efficient Algorithms for Discovering Association Rules. KDD Workshop 1994: 181-192.

(Miller 97) R.J. Miller and Y. Yang. Association rules over interval data. SIGMOD 97, 452-461, Tucson, Arizona.

(Mitchell 99) T. Mitchell: The Role of Unlabeled Data in Supervised Learning. Proceedings of the Sixth International Colloquium on Cognitive Science, San Sebastian, Spain, 1999

(Musicant 99) O. L. Mangasarian and David R. Musicant: Successive Overrelaxation for Support Vector Machines. IEEE Transactions on Neural Networks, 10, 1999, 1032-1037.

(Musicant 00) O. L. Mangasarian and D. R. Musicant: Robust Linear and Support Vector Regression. IEEE Transactions on Pattern Analysis and Machine Intelligence 22, September 2000, 950-955.

(Musicant 01a) O. L. Mangasarian and David R. Musicant: Data Discrimination via Nonlinear Generalized Support Vector Machines. Complementarity: Applications, Algorithms and Extensions, M. C. Ferris, O. L. Mangasarian and J.-S. Pang, editors, Kluwer Academic Publishers, Dordrecht, The Netherlands, 2001, pages 233-251.

(Musicant 01b) O. L. Mangasarian and David. R. Musicant: Active Set Support Vector Machine Classification. Advances in Neural Information Processing Systems 13, Todd K. Leen, Thomas G. Dietterich, and Volker Tresp, editors. MIT Press, Cambridge, MA, 2001, pages 577-583.

(Musicant 01c) O. L. Mangasarian and David. R. Musicant: Lagrangian Support Vector Machines. Journal of Machine Learning Research 1, March 2001, 161-177.

(Musicant 02) O. L. Mangasarian and David R. Musicant: Large Scale Kernel Regression via Linear Programming. Machine Learning 46, January 2002, 255-269.

(Moosmuller 01) H. Moosmuller, W. P. Arnott, C. F. Rogers, J. L.Bowen, J. A. Gillies, W. R. Pierson, J. F. Collins, T. D. Durbin, J. M. Norbeck "Time Resolved Characterization of Diesel Particulate Emissions. 1. Instruments for Particle Mass Measurements" Environ. Sci. Technol, 2001; 35, 781-787.

(Musicant 03) David R. Musicant, V. Kumar, and A. Ozgur: Optimizing F-Measure with Support Vector Machines. Proceedings of the Sixteenth International Florida Artificial Intelligence Society Conference, to be published May 2003.

(Musicant LSVM) David R. Musicant: Lagrangian Support Vector Machine software, 2000.

(Musicant ASVM) David R. Musicant: Active Support Vector Machine software, 2000.

(Musicant ASVR) David R. Musicant: Active Set Support Vector Regression software, 2002.

(NRC 96) National Research Council, A plan for a research program on aerosol radiative forcing and climate change, 1996, National Academy Press, Washington, DC.

(NRC 98) National Research Council, Research Priorities for Airborne Particulate Matter. I. Immediate Priorities and a Long-Range Research Portfolio, 1998, National Academy Press, Washington, DC.

(Ng 98) R. Ng, L.V.S. Lakshmanan, J. Han & A. Pang. Exploratory mining and pruning optimizations of constrained association rules. SIGMOD 98.

(Noble 96) C. A. Noble, and K. A. Prather: Real-time measurement of correlated size and composition profiles of individual atmospheric aerosol particles. Environ. Sci. Technol, 1996; 30, 2667-2680.

(O'Callaghan 02) Liadan O'Callaghan, Nina Mishra, Adam Meyerson, Sudipto Guha, Rajeev Motwani, Streaming-Data Algorithms For High-Quality Clustering. ICDE 2002.

(Ozden 98) B. Ozden, S. Ramaswamy, and A. Silberschatz. Cyclic association rules. ICDE'98, 412-421, Orlando, FL.S. Ramaswamy, S. Mahajan, and A. Silberschatz. On the discovery of interesting patterns in association rules. VLDB 98, 368-379, New York, NY.

(Padmanabhan 98) B. Padmanabhan, and A. Tuzhilin: A belief-driven method for discovering unexpected patterns. KDD-98, 1998, pp. 94-110.

(Padmanabhan 00) B. Padmanabhan, and A. Tuzhilin: Small is Beautiful: Discovering the Minimal Set of Unexpected Patterns. Procs. ACM SIGKDD International Conference on Knowledge Discovery and Data Mining KDD 2000, pages 54-64.

(Park 95) J. Park, M. Chen, and P. Yu. An effective hash-based algorithm for mining association rules. In SIGMOD 95.

(Pasquier 98) N. Pasquier, Y. Bastide, R. Taouil, and L. Lakhal. Discovering frequent closed itemsets for association rules. ICDT 99, 398-416, Jerusalem, Israel, Jan. 1999.

(Pei 01) J. Pei, J. Han, B. Mortazavi-Asl, H. Pinto, Q. Chen, U. Dayal and M-C. Hsu. PrefixSpan: Mining Sequential Patterns Efficiently by Prefix-Projected Pattern Growth. In. Proc. 2001 Int. Conf. Data Engineering (ICDE'01), pages 215-224, Heidelberg, Germany, April 2001.

(Pei 01a) J. Pei, J. Han, and L. V. S. Lakshmanan, Mining Frequent Itemsets with Convertible Constraints, Proc. 2001 Int. Conf. on Data Engineering (ICDE'01), April 2001.

(Phares 01) Denis J. Phares, Kevin P. Rhoads, Anthony S. Wexler, David B. Kane, Murray V. Johnston "Application of the ART-2a Algorithm to Laser Ablation Aerosol Mass Spectrometry of Particle Standards" Anal. Chem. 2001, 73, 2338-2344.

(Platt 99a) J. Platt, Fast Training of Support Vector Machines using Sequential Minimal Optimization, in Advances in Kernel Methods - Support Vector Learning, B. Schölkopf, C. Burges, and A. Smola, eds., MIT Press, 1999.

(Platt 99b) J. Platt, Using Sparseness and Analytic QP to Speed Training of Support Vector Machines, in Advances in Neural Information Processing Systems 11, M. S. Kearns, S. A. Solla, D. A. Cohn, eds., MIT Press, (1999).

(Popivanov 02) Popivanov, I. & Miller, R. J. Similarity search over time series data using wavelets. In proceedings of the 18th Int'l Conference on Data Engineering. San Jose, CA, Feb 26-Mar 1.

(Prather 94) K. A. Prather, T. Nordmeyer, and K. Salt. Real-time characterization of individual aerosol particles using time-of-flight mass spectrometry. Anal. Chem., 1994; 66, 1403-1407.

(Rafiei 98) Rafiei, D. & Mendelzon, A. O. Efficient retrieval of similar time sequences using dft. In proceedings of the 5th Int'l Conference on Foundations of Data Organization and Algorithms. Kobe, Japan, Nov 12-13.

(Rafiei 99) Rafiei, D. On similarity-based queries for time series data. In proceedings of the 15th IEEE Int'l Conference on Data Engineering. Sydney, Australia, Mar 23-26. pp 410-417.

(Ramakrishnan 98) Raghu Ramakrishnan, Donko Donjerkovic, Arvind Ranganathan, Kevin S. Beyer, Muralidhar Krishnaprasad: SRQL: Sorted Relational Query Language. SSDBM 1998: 84-95.

(Ramakrishnan 02) R. Ramakrishnan and J.G. Gehrke, Database Management Systems, 3ed, McGraw-Hill, 2002.

(Sarawagi 98) S. Sarawagi, S. Thomas, and R. Agrawal. Integrating association rule mining with relational database systems: Alternatives and implications. In SIGMOD 98.

(Savasere 95) Ashok Savasere, Edward Omiecinsky, and Shamkant Navathe. An efficient algorithm for mining association rules in large databases. In 21st Int'l Conf. on Very Large Databases (VLDB), Zurich, Switzerland, Sept. 1995.

(Savasere 98) Savasere A., Omiecinski E., and Navathe S. B. "Mining for Strong Negative Associations in a Large Database of Customer Transactions." Proceedings of the International Conference on Data Engineering, February 1998.

(Schauer 96) J. J. Schauer, W. F. Rogge, L. M. Hildemann, M. A. Mazurek, G. R. Cass, and B. R. T. Simoneit., "Source Apportionment of Airborne Particulate Matter Using Organic Compounds as Tracers." Atmospheric Environment. 1996, 30, 3837-3855.

(Schauer 00) J. J. Schauer and G. R. Cass., "Source Apportionment of Wintertime Gas-Phase and Particle-Phase Air Pollutants Using Organic Compounds as Tracers." Environmental Science and Technology. 2000, 34, 1821-1832.

(Schauer 02) J. J. Schauer, M. P. Fraser, G. R. Cass, and B. R. T. Simoneit, "Source Reconciliation of Atmospheric Gas-Phase and Particle-Phase Pollutants Using Organic Compounds as Tracers" Environmental Science and Technology. 2002, 36, 3806-3814.

(Schauer 03) J. J. Schauer, B. T. Mader, J. T. DeMinter, G. Heidemann, M. S. Bae, J. H. Seinfeld, R. C. Flagan, R. A. Cary, D. Smith, B. J. Huebert, T. Bertram, S. Howell, P. Quinn, T. Bates, B. Turpin, H. J. Lim, J. Yu, and H. Yang., "ACE-Asia Intercomparison of a Thermal Optical Method for the Determination of Particle-Phase Organic and Elemental Carbon." Environmental Science and Technology. 2003, In press.

(Seinfeld 98) J. F. Seinfeld and S. N. Pandis, Atmospheric Chemistry and Physics: From Air Pollution to Climate Change, John Wiley & Sons, 1998, New York.

(Seshadri 95) Praveen Seshadri, Miron Livny, Raghu Ramakrishnan: SEQ: A Model for Sequence Databases. ICDE 1995: 232-239.

(Seshadri 96) Praveen Seshadri, Miron Livny, Raghu Ramakrishnan: The Design and Implementation of a Sequence Database System. VLDB 1996: 99-110.

(Shintani 98) T. Shintani and M. Kitsuregawa. Mining algorithms for sequential patterns in parallel : Hash based approach. Second Pacific--Asia Conference on Knowledge Discovery and Data mining, April 1998.

(Silverstein 98) C. Silverstein, S. Brin, R. Motwani, and J. Ullman. Scalable techniques for mining causal structures. VLDB'98, 594-605, New York, NY.
(Smola 00) A.J. Smola and B. Schölkopf. Sparse greedy matrix approximation for machine learning. In P. Langley, editor, Proc. ICML'00, pages 911-918, San Francisco, 2000. Morgan Kaufmann.

(Smyth 02) Padhraic Smyth, Daryl Pregibon, Christos Faloutsos, Data-driven evolution of data mining algorithms. CACM 45(8): 33-37 (2002).

(Song 99) Xin-Hua Song, Philip K. Hopke, David P. Fergenson, Kimberly A. Prather "Classification of Single Particles Analyzed by ATOFMS Using an Artificial Neural Network, ART-2A" Anal. Chem. 1999, 71, 860-865.

(Srikant 96) Srikant, R., & Agrawal, R.: Mining sequential patterns: Generalizations and performance improvements, Proc. of the Fifth Int'l Conference on Extending Database Technology (EDBT). Avignon, France, 1996.

(Srikant 96b) R. Srikant and R. Agrawal. Mining quantitative association rules in large relational tables. SIGMOD 96, 1-12, Montreal, Canada.

(Srikant 97) R. Srikant, Q. Vu, and R. Agrawal. Mining association rules with item constraints. KDD 97, 67-73, Newport Beach, California.

(Suess 99) D. T. Suess, K. A. Prather "Mass Spectrometry of Aerosols" Chemical Reviews 1999, 99, 3007-3035.

(Tan 02) Phillip V. Tan, Oscar Malpica, Greg J. Evans, Sandy Owega, Michael S. Fila "Chemically-Assigned Classification of Aerosol Mass Spectra" J. Am. Soc. Mass Spectrom. 2002, 13, 826-838.

(Toivonen 96) H. Toivonen. Sampling large databases for association rules. In VLDB 96.

(Towell 94) G.G. Towell and J.W. Shavlik: Knowledge-Based Artificial Neural Networks. Artificial Intelligence, 70, pp. 119-165.

(Tsur 98) D. Tsur, J. D. Ullman, S. Abitboul, C. Clifton, R. Motwani, and S. Nestorov. Query flocks: A generalization of association-rule mining. SIGMOD 98, 1-12, Seattle, Washington.

(Tung 99) A. K. H. Tung, H. Lu, J. Han, and L. Feng, "Breaking the Barrier of Transactions: Mining Inter-Transaction Association Rules", Proc. 1999 Int. Conf. on Knowledge Discovery and Data Mining (KDD 99), San Diego, CA, Aug. 1999, pp. 297-301.

(Turpin 90) B. J. Turpin, R.A. Cary, and J.J. Huntzicker, "An In-Situ, Time-Resolved Analyzer for Aerosol Organic and Elemental Carbon," Aerosol Science and Technology, 1990, 161-171.

(Vapnik 95) V. N. Vapnik: The Nature of Statistical Learning Theory. Springer, 1995.

(Veloso 02) A. Veloso, W. Meira, M. Carvalho, B. Possas, S. Parthasarathy and M.Zaki, Efficiently mining approximate models of associations in Evolving Databases , to appear in ECML/PKDD 2002.

(Wagstaff 00) Kiri Wagstaff and Claire Cardie: Clustering with Instance-level Constraints. ICML-2000.

(Wang 02) Changzhou Wang, Xiaoyang Sean Wang, Supporting Content-Based Searches on Time Series via Approximation. SSDBM 2000: 69-81.

(Weigend 94) A. Weigend and N. Gerschenfeld. Time Series Prediction: Forecasting the Future and Understanding the Past. Addison Wesley, 1994.

(Weiss 97) Sholom M. Weiss and Nitin Indurkhya: Predictive Data Mining: A Practical Guide. Morgan Kaufmann, 1997.

(Weston 00) Jason Weston, Sayan Mukherjee, Olivier Chapelle, Massimiliano Pontil, Tomaso Poggio, Vladimir Vapnik: Feature Selection for SVMs. NIPS 2000: 668-674

(Witten 99) Ian Witten and Eibe Frank: Data Mining, Practical Machine Learning Tools and Techniques with Java Implementations. Morgan Kaufman, 1999.

(Wolsley 98) Wolsley, Laurence: Integer Programming. Wiley-Interscience, 1998.

(Wrobel 97) Wrobel, S. An algorithm for multi-relational discovery of subgroups, Proc. First European Symposion on Principles of Data Mining and Knowledge Discovery.

(Wu 00) Wu, Y., Agrawal, D. & El Abbadi, A. (2000). A comparison of DFT and DWT based similarity search in time-series databases. In proceedings of the 9th ACM CIKM Int'l Conference on Information and Knowledge Management. McLean, VA, Nov 6-11. pp 488-495.

(Yang 02) Mining long sequential patterns in a noisy environment, by Jiong Yang, Wei Wang, Philip Yu, and Jiawei Han, Proceedings of the ACM SIGMOD International Conference on Management of Data (SIGMOD), pp. 406-417, 2002.

(Yi 00) B.-K. Yi, N. D. Sidiropoulos, T. Johnson, A. Biliris, H. V. Jagadish and C. Faloutsos. Online Data Mining for Co-Evolving Time Sequences. In Proceedings of the IEEE Sixteenth International Conference on Data Engineering, pages 13--22, 2000.

(Yoda 97) K. Yoda, T. Fukuda, Y. Morimoto, S. Morishita, and T. Tokuyama. Computing optimized rectilinear regions for association rules. KDD 97, Newport Beach, CA, Aug. 1997.

(Yao 02) Y. Yao and J. Gehrke, The Cougar Approach to In-Network Query Processing in Sensor Networks, SIGMOD Record, September 2002.

(Zaki 97) M. J. Zaki, S. Parthasarathy, M. Ogihara, and W. Li. Parallel algorithm for discovery of association rules. Data Mining and Knowledge Discovery, 1:343-374, 1997.

(Zaki 98) M.J. Zaki. Efficient enumeration of frequent sequences. CIKM 98. Novermber 1998.

(Zaki 99) M. Zaki. CHARM: An Efficient Algorithm for Closed Association Rule Mining, CS-TR99-10, Rensselaer Polytechnic Institute.

(Zaki 00) M. Zaki. Generating Non-Redundant Association Rules. KDD 00. Boston, MA. Aug. 2000.

(Zaki 01) Fast Vertical Mining Using Diffsets, TR01-1, Department of Computer Science, Rensselaer Polytechnic Institute.

(Zaki 01a) M. Zaki. SPADE: An efficient algorithm for mining frequent sequences. Machine Learning, 42(1/2):31-60, 2001.

(Zhang 97) Tian Zhang, Raghu Ramakrishnan, Miron Livny: BIRCH: A New Data Clustering Algorithm and Its Applications. Data Mining and Knowledge Discovery 1(2): 141-182 (1997)

(Zheng 02) M. Zheng, G. R. Cass, J. J. Schauer, E. S. Edgerton., “Source Apportionment of PM2.5 in the Southeastern United States Using Solvent-Extractable Organic Compounds as Tracers” Environmental Science and Technology. 2002, 36, 2361-2371.

(Zhou 01) Z. Zhou, H. Liu, S.Z. Li, and C.S. Chua, "Rule Mining with Prior Knowledge - A Belief Networks Approach", Intelligent Data Analysis, 5(2):95-110, 2001.