

College of Saint Benedict and Saint John's University

DigitalCommons@CSB/SJU

---

Honors Theses, 1963-2015

Honors Program

---

1998

## Data Mining in Electronic Media Usage Statistics: A Case Study of Knowledge Discovery in Databases

Peter J. Lindquist

*College of Saint Benedict/Saint John's University*

Follow this and additional works at: [https://digitalcommons.csbsju.edu/honors\\_theses](https://digitalcommons.csbsju.edu/honors_theses)



Part of the [Computer Sciences Commons](#)

---

### Recommended Citation

Lindquist, Peter J., "Data Mining in Electronic Media Usage Statistics: A Case Study of Knowledge Discovery in Databases" (1998). *Honors Theses, 1963-2015*. 660.

[https://digitalcommons.csbsju.edu/honors\\_theses/660](https://digitalcommons.csbsju.edu/honors_theses/660)

Available by permission of the author. Reproduction or retransmission of this material in any form is prohibited without expressed written permission of the author.

# Data Mining on Electronic Media Usage Statistics

*A Case Study of Knowledge Discovery in Databases*

A THESIS

The Honors Program

College of St. Benedict/St. John's University

In Partial Fulfillment

of the Requirements for the Distinction "All College Honors"

and the Degree of Bachelor of Arts

In the Department of Computer Science

by

Peter Lindquist

May 1998

## Table of Contents

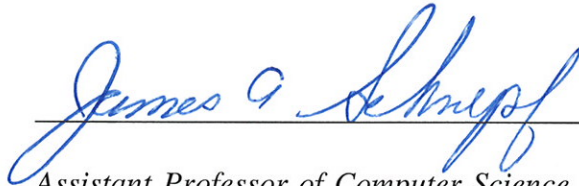
---

Abstract	1
Signatures	2
<i>1.</i> Introduction	3
<i>2.</i> KDD	9
<i>3.</i> Situation	18
<i>4.</i> Steps Analysis	26
<i>5.</i> Analysis	35
Appendix A	38
Appendix B	46
Sources	47

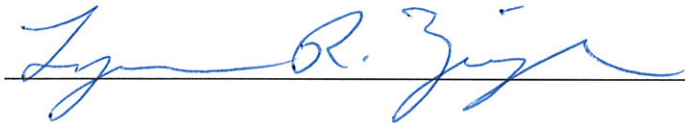
---

## Signatures

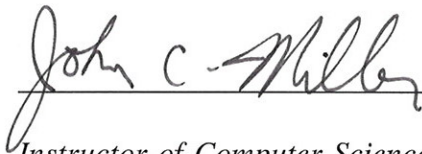
---



*Assistant Professor of Computer Science, Advisor*



*Associate Professor of Computer Science*



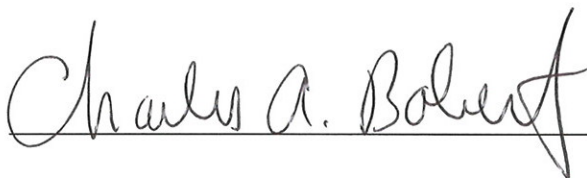
*Instructor of Computer Science*



*Chair, Department of Computer Science*



*Director, Honors Thesis Program*



*Director, Honors Program*

## **Abstract**

---

As databases grow larger, analysts are turning to computers to help them analyze the massive amounts of data their computers have collected. As the difference between having data and having useful information becomes more clear, different methods of using computers to analyze data are becoming available. Knowledge Discovery in Databases (KDD) is a general methodology for preparing the data, using software algorithms to discover new patterns or relationships in the data, and integrating the results back into the system. The KDD methodology is explained and hypothetically applied to usage statistics generated by the CSB/SJU Libraries Internet resources. Examples are drawn from that source and from other industries to clearly illustrate the properties of Knowledge Discovery and decide if KDD is an appropriate methodology for the Libraries to use in this situation.

## **1. Introduction**

---

Computers are in a unique position in our society. We have equipped them to do the tasks we do not wish to do, or that they can perform faster with fewer mistakes than a human makes when performing the same rote tasks. This arrangement usually means computers process vast amounts of information on every conceivable subject because they are installed anywhere people can afford to put them to automate tasks. Thanks to their nature, they can store their data to come back to it whenever they please, and since we associate some intrinsic value with data we have our computers store most of it for us.

New methodologies have grown up out of a need for using computers to analyze this surplus of data stored in computers. Knowledge Discovery in Databases is one process that attempts to effectively learn from raw data stored by our machines. Based around the passive theories of Data Mining, Knowledge Discovery seeks to equip computers to analyze data on their own. The computer and software then report back with their results in order to discover new patterns or relationships within the data. The methodologies' steps will be explored through a case study of usage statistics owned by the CSB/SJU Libraries which possesses its own stack of uninterpreted data just like larger corporations and scientists.

Why is all of this data being stored? First of all, it is easy since the computer already has all of the data in memory. All it has to do is dump the data into files for the long term. Second, long term storage is already cheap while getting cheaper as hard drive prices fall further every day, and long term tape storage costs pennies on a corporation's scale of resources. For example, a modern IDE hard drive is available to the consumer for only about

25 cents per MegaByte. However,

Raw data is rarely of direct benefit. Its true value is predicated on the ability to extract information useful for decision support or exploration and understanding the phenomena governing the data source. (ACM p. 26)

It isn't always obvious that your ability to understand anything from your data is going to be quickly swamped by the sheer amount of data your machines collect.

That's the crucial difference between data and insight. The modern computer is more than capable of storing data about everything it touches: stock reports, stellar analysis, customers, transactions of all kinds... but this data is only useful if it can be researched, added to old knowledge and then used to move us in new directions. In short, we must gain information from the data to understand what it *means*, not just know that it is there, or what it says.

(ACM p. 26)

Our big human brain is absolutely unparalleled at this task of collating new information from data and then understanding how it fits into the big picture. From raw data we can see patterns, analyze pictures, and spot trends that we aren't even consciously aware of until they pop right out in front of us. Unfortunately, we can only do that for limited amounts of data. A human analyst should be able to get familiar with and study a database of a few thousand records with two or three fields each, but the human brain hasn't been invented yet that can take in and usefully correlate a relational database filled with millions of records each containing hundreds of fields.

But the habit of storing data, *which we confuse with having information*, leaves the

owners, whether they be corporations or scientists, with a pile of "knowledge" that they cannot use. It cannot be used for the simple reasons that there is too much of it and it is completely unorganized for useful study. [ACM p. 27] No one can slog through the this pile of data to figure out what all of it *means* because the task is stymied by the amount of data available.

To fill this gap in our understanding, a new field is seeking to combine our human analyzing capabilities with a computer's speed and ability to work with huge stacks of data. Knowledge Discovery in Databases (or just Knowledge Discovery or acronym KDD) seeks to use modern techniques of programming and modeling to analyze data for us. It is currently being applied to many business or scientific applications such as marketing, fraud detection, stellar mapping, and health care trends. This new approach has potential applicability anywhere there is a surplus of data plus a shortage of real information.

These places usually aren't hard to find in the computer age. There are many institutions that are struggling to integrate new technologies into their current practices. Here at CSB/SJU, the Libraries have the same problem other corporations do; they want to understand their data, but they have limited resources with which to accomplish this. In the information age, the role of libraries with their accompanying librarians has changed from the traditional vision of storing and categorizing books so others could find them on demand, to storing information, data, and collected knowledge in many forms. Books have been joined by microfiche, videos, periodicals, and access to more widespread forms of information located on the Internet. However, the shortage of money has never changed, so the library,



along with other organizations, still has to accomplish its goals with limited resources.

Here at CSB/SJU, the Libraries subscribe to Internet database resources that cost money as well generate piles of usage data. Somehow, this data needs to be used to understand how these new electronic resources are being used by the community. Other companies have huge databases containing information on customers and transactions with no reasonable way to figure out what to do with it. For our Libraries though, data on what resources their customers use is what is important. Attempts to gather this data have resulted in a large stack of data that must be analyzed in order to understand it as much as possible.

Can the Knowledge Discovery process be used by the CSB/SJU Libraries? To decide, the circumstances that led to KDD as well as its principles, usefulness, and scope of application will be examined. That information will provide a backdrop to the question of KDD's applicability to the CSB/SJU Libraries' data issues. Each step of the KDD process will be worked through in a theoretical way with examples showing how the data would actually be examined were the process for real.

To be successful in different situations KDD relies on a formal, general set of steps so that it is applicable to situations that are wildly different in design. For example, the same KDD methodology could be applied to learning more about a customer database owned by Target as easily as it could be applied to studying stellar analysis data collected by some research scientists. This generality of theory makes for a useful approach, but it also means the interpretation of individual steps in the process can vary between implementations.

This generality gives KDD a great deal of flexibility, but it does not mean that KDD

is an appropriate tool for every situation. The study done in this paper evaluates whether KDD is a good way to continue to analyze the Libraries' data. As the Libraries' data is examined in this preliminary stage, there may be steps that simply do not apply in more than a trivial way. In order to make it clear why these steps are important, examples will be drawn from other sources and shown alongside the Libraries' implementation to show why a KDD technique exists when the Libraries' particular situation doesn't necessarily warrant approaching the problem in that way.

Is KDD an appropriate approach for the Libraries' situation? That's the sort of question that has to be asked carefully before actually applying KDD to anything. The time, money, and lengthy turn around time involved in properly applying KDD to a problem make impulsiveness a risky proposition at best and a complete, undirected waste of time at worst. If given this unorganized stack of data with accompanying questions about it, is it appropriate for the librarians to turn to a KDD approach to help them understand it? Attempting to answer this question will form analysis done just to evaluate the feasibility of the project as a whole.

Understanding data in order to decide what to do with these limited resources is a major task in this environment. Given all of these methods of getting at information, what are people, or customers, using? For Internet resources, data can be found that "shows" what customers are viewing. However, this data comes in sizable pieces that are unorganized and contain non-intuitive piles of data that show such things as page hits. Our librarians are left in the same position as many others dealing with computers- large amounts of data that

supposedly show what to do with their limited resources. They have no real understanding of how to look at the data, use it, or go about getting more if it isn't enough.

Discovering how to place resources to effectively accomplish goals means answering useful questions such as:

- Should they [the Libraries] keep expanding Internet material availability?
- Should they advertise Internet materials more or less?
- Who are using the materials and for what?
- Are these materials worth the investment?
- Should the investment in Internet materials be increased, decreased, or modified?
- Which, among the multitude of resources, are being used most productively?

It is possible for some or all of these questions to be answered from the data held by the Libraries if proper investigating is carried out. It will be shown in the KDD process that thinking about the process in directed terms like these questions at the beginning ensures efforts stay focused on providing us what we need to know when we are finished.

To map out our route more specifically, we'll first be looking at KDD as a whole, showing what it is, what it is not, and the goals of the process. After showing the situation for the libraries in detail, we'll run through KDD step-by-step to show how the pieces fit together. This is how the library might actually proceed if they applied this process. Finally an attempt will be made to conclude if KDD is a methodology that should actually be used for the Libraries or not.

## 2. KDD

---

This new technique of Knowledge Discovery in Databases, hereafter referred to as Knowledge Discovery or just KDD, grew from Data Mining; the attributes of the DM process need to be understood in order to understand the more refined KDD methodology. In order to introduce it clearly, background will be given on machine analysis and the methodology of Data Mining as it leads into KDD. First, it must be made clear that computers are first and foremost pretty dumb machines. They like to count from 0 to 1 and back again and are really only good at doing just that millions of times in a very short period. To improve this limited functionality, computers use software to coordinate millions of binary data elements to accomplish tasks. This software functions under strict rules, definitions, and expectations in order to interpret its environment properly. Consequently, attempting to use them to do the type of analysis a human brain can do is difficult at best. A human brain can handle new information and situations amazingly well, which computer software cannot do easily at all.

Historically, computers have demonstrated a lot of problems when applied to these sorts of analysis problems. The largest issue was the discovery that if you search long enough it is possible to find patterns in randomly generated data that appear to be statistically relevant, *but are not*. These patterns are not actually significant, hence they have no real use. This means that teaching computers that can only tell the difference between 0 and 1 to discern between what is useful and what is not, is something that humans have had remarkable little success at until recently. This might be largely due to the increased speed of computers with the accompanying increase in the amount of material a computer can analyze.

Improving our interaction with the machines and our understanding of their limitations has allowed us to go further, however. In the present, a couple of different methodologies are available to someone interested in attempting to automate data analysis. Data Mining (DM) is the one that was later incorporated into KDD.

The Data Mining core step of KDD is distinguishable from other data query paradigms by its passive nature. DM in general is the attempt to find patterns in and learn from unorganized data. As it is used here, Data Mining is the process of a computer analyzing data on its own without constant human interaction. In other words, it has computers doing just what they are not inherently good at. Depending on the specific goals of the users, a computer performing Data Mining utilizes different approaches and models, searches for patterns as well as creates and verifies its own hypotheses, and then returns the "information" it finds. Some pre- and postprocessing of data by humans is necessary, but to achieve the benefits of having an unbiased machine looking for patterns, it is desirable to let the machine do as much as possible.

It isn't easy to sort through the multiple claims of Data Mining capabilities. There are other related fields out there that attempt to accomplish the same goal of understanding raw data as DM, but they use different methodologies to solve the problem. The one most often confused with DM is OnLine Analytical Processing (OLAP). OLAP is an interactive, deductive process that requires the user to ask the system a question before any action can be taken. An example of a common question type would be: "Did people in Minnesota buy more sweaters than people in Texas last year?" The system might answer back "People in

Minnesota purchased a half million sweaters as compared to 100,000 purchased in Texas."

One of the primary problems with this operation is that the questions the human can ask are limited by their preconceptions and understanding of the concepts involved in the data.

Another issue is that a human might make connections that don't exist. Just because the computer can compare two numbers doesn't mean the comparison actually has meaning in the real world. While still looking at sweaters, a human analyst might see a connection between a rise in purchases and the cold weather in the previous fall. In reality, those purchases of sweaters might have been caused more by the extra sale offered instead of the weather.

A Data Mining attempt doesn't require the user to ask a direct question in that manner; it is more like telling the system to use a model on the data and report back with results. Using inductive reasoning this way, the system is capable of finding relationships that the humans involved couldn't have had the insight to suspect before starting the process. The more open-ended situation used here might be "Find a model that predicts the sweater buying behavior of customers broken down by geographic location." Instead of getting a simple answer back, the system might say it depends a lot on time of year, but it also might have new insight into how the customer's occupation, income, age, or family status affect their likelihood of buying a sweater. Finding those new relationships can save a company millions in mass mailings and special offerings aimed at different customer segments.

While telling the difference between these processes isn't always easy given the new and relatively undefined nature of data mining, the rule of thumb is that data mining is an automated process. The system searches for information on it's own once started, and works

without human interaction on the problem. Another difference is whether or not the system has a "good statistical basis for what is interesting and what is irrelevant" (Gerber) so as to be able to calculate relevancy on its own.

A classic example is passed down to illustrate how this examination of data by a computer can be problematic, resulting in the more formal approach of KDD. That example is the retail connection between beer and diapers. (Shaku Atre) In it, an unnamed retailer noticed that beer and diapers had a regular tendency to end up in the same shopping cart on Friday nights. However, knowing that doesn't tell us why or what action to take that would result in more sales. Possibly Dad is being sent out for diapers and is getting a six-pack as well, or maybe Mom is in the store for a six-pack when junior starts to cry. Either way, there are many wildly different scenarios as to how the shelves and items should be arranged to encourage this behavior and the accompanying impulse buying, but without more information there is no way to direct the efforts of the store managers. It is particularly easy to see in this case that while DM possibly returned a potentially very useful connection, there isn't any way to use this knowledge directly.

Knowledge Discovery has evolved out of this "turn the system loose" process. Given this heritage, it should be made clear what Knowledge Discovery *is* and, more importantly, what it *is not*. KDD is an interactive process designed as an evolutionary step past the Data Mining process that helps ensure useful information as a result of the time and money invested. As in standalone Data Mining, KDD is concerned with discovering patterns in raw data in order to improve or understand the process that generated the data, but KDD

recognizes the complexities of the issues involved, consequently placing much more emphasis on formal pre- and postprocessing than comparable undirected Data Mining attempts. In effect, KDD attempts to regulate the entire system of working with the data so that efforts remain directed and results are more likely to be understandable and useful to the humans involved. The hope in the beer/diapers example above is that the analysts would be prepared for those types of relationships by having asked careful questions when they started. Perhaps then they would know that the companies standard practice in such situations is to put products next to each other, or far apart with impulse buy items in between. In that case, the way to use the connection is apparent.

For Knowledge Discovery in Databases, optimistic

Proponents envision a rosy workplace where hundreds of employees sit at their computers, feast on data, and then hatch innovative ideas, [while] the reality is much more modest. *Datamation*, Nov 1997 v43 p66

Good results from KDD generally come from highly trained employees who have time to go through the steps of the process properly. They have to understand the business they are working in, as well as be familiar with the form of the data sets they are trying to analyze. Because of these difficulties, the KDD advantages are available only to those willing to invest the time and money into it. In these days of competitive advantage, that means corporations, large businesses, and scientists with deep-pocketed backers. These limitations keep KDD from being a perfect solution to the problem of unexamined data.

KDD, of which Data Mining is just a step, does allow the ability to carefully analyze a surplus of data in a new way. For situations generating gigabytes of data, it may represent



one of the only ways currently existing for analyzing the data at all. It functions well for domains that meet several criteria: they need to have varied data without existing models for working with it. In addition, a changing environment that needs good information to make decisions is more flexible and able to use the information gained. Most importantly, there needs to be a rich payoff for making the right decisions so the expense of KDD can be justified. (ACM p. 48) If carefully done in the business world, KDD is capable of generating revenue by saving money or spotting new trends in the market, but the technology is still very young and hard to apply successfully. The high payoff is needed to recover the amount of the investment, which may be anywhere from \$10,000 for a small business's workhorse computer to \$500,000 or more for software, storage space, and employee time for a large corporation. It must be emphasized that it is most definitely not a quick or easy solution to the general problem. However, it can provide a framework within which to work on the solution. This framework is designed to circumvent the problems that can arise when Data Mining is attempted in an unfocused manner.

Now that a general idea for the organization of KDD has been laid out, a useful definition can be stated.

*KDD is the nontrivial, interactive, and iterative process of identifying valid, novel, potentially useful, and ultimately understandable patterns, models, or structures in data.*

As usual in modern computing theory, the buzzwords are thick in this definition. In practical terms, Knowledge Discovery means that someone is applying the steps of the process again and again to the data in order to gain reasonable, useful, and hopefully

profitable insight from it.

**Buzzwords:**

- *Nontrivial:* The search isn't just for limited explanations or data sets, patterns or parameters are the goal.
- *Interactive:* A partnership between the user and machine during different stages.
- *Iterative:* The entire process doesn't happen just once, it is more useful when new insight gained is used to look at the data again and expand your knowledge of the system.
- *Valid:* With some degree of certainty it should be possible to apply the new pattern to the data set reliably.
- *Novel:* The patterns discovered are new to the user and the system.
- *Useful:* Results are either predictive or descriptive so they can be applied.
- *Understandable:* If not immediately, then after post-processing and interpretation a use for the information gained is discovered.
- *Patterns:* An expression, model, or structure that describes patterns in the data.

Like most other uses of a computer, KDD is a partnership between machine and user.

It requires trained humans to work with machines and the data to come up with viable information as a result. Some corporations labor under the misconception that many of their employees will access data stores and come up with revolutionary ways to improve business on a daily basis. Not only is this not very likely to occur, companies must also question the security issues involved with allowing hundreds of employees access to those kinds of data stores, especially when those stores are very likely to have valuable and private customer lists in them. (Datamation, v43, Deborah Asbrand)

Knowledge Discovery seeks to eliminate a lot of the inherent problems of undirected DM by working through a series of steps designed to make sure things are done carefully. As much as possible it also utilizes the human ability to analyze data by using humans to assure the computers get good data to work with and a defined goal for the software to work

towards. Instead of just analyzing data, KDD is interested in the entire process of understanding the data

including how the data is stored and accessed, how algorithms can be scaled to massive datasets and still run efficiently, how results can be interpreted and visualized, and how the overall human-machine interaction can be modeled and supported. (ACM p. 29)

This is an attempt to recognize that all of the variables surrounding the project must be understood and controlled in order for the information gleaned from the data to be as meaningful as possible.

It is also important to note that these attempts aren't taking place only in theory by scientists. Many companies are currently making notable inroads into the realm of practicality, which means implementing partial KDD processes as the need arises. This is because various problems don't require all of the steps, but partial solutions help in specific areas. HyperParallel produces Data Mining templates specifically for the Banking Industry. Chase Manhattan, AT&T, SGI, and IBM are all working on various entries into the software marketplace. American Skiing dealt with legacy systems and switched to data warehouse technology and spent between \$5 and \$6 million dollars on the deal. They own only 6 six ski resorts, so it is obvious that the expense of implementing this warehouse for even a medium sized company can be considerable.

One of the more interesting examples involves finding volcanoes on the surface of the planet Venus. The Magellan spacecraft faithfully took pictures of the surface for more than five years. After that time, it mapped the entire surface to 75 meters per pixel pictures, which means we know more about Venus' surface than we do of Earth's. The size of the

dataset is more than 30,000 images at 1,000 x 1,000 pixels each. The system worked by having the geologists mark the small volcanoes on 30 to 40 images. After that, the system constructed a classifier that distinguishes that terrain feature and went to work on the rest of the images. Interestingly, the system matches a scientist's performance on recognizing high probability volcanoes versus the ones no one is sure about, but the approach doesn't generalize well to more than that.

As we proceed from here, we'll lay out the libraries situation in the form of their data and needs to form a foundation for our analysis.

### **3. Situation**

---

Usually KDD is performed by a corporation that has been collecting data for years. Commonly, this is a set of customer and transaction data that shows buying histories. This sort of information is usually stored in a relational database that may take up Gigabytes of space.

Our libraries have compiled a quite different set of data however. The primary data of interest is a set of usage statistics of web-based resources that the libraries have added to their collection. The two primary resources are the JSTOR and InfoTrack databases. JSTOR and InfoTrack (or IAC, Information Access Company SearchBank) are organizations whose services are located on the World Wide Web. An organization dedicated to helping institutions benefit from new technologies, JSTOR (<http://www.jstor.org/>) provides various journals and the ability to search through them online. With this ability a customer can search reputable academic journals without leaving their chair. InfoTrack (<http://www.searchbank.com/>) is a company whose stated goal is to allow customers to do research wherever they can get connected to the Internet. Together these databases and indexes are powerful research tools that can greatly ease the workload of researching topics.

The subscription purchased by our Libraries allows community members at CSB/SJU to search through and retrieve articles or journal entries from the many magazines and journals contained in these databases. The use of these resources involves someone logging into the database from our campus and searching for some topic among the thousands of articles, periodicals, and journal entries contained in these massive databases. They may also

have the option of retrieving the full text of the items they find, which completely circumvents the step of having to locate a physical copy of the item for their use.

This is typical of how web resources are used. The usage data that is available is a record of what was accessed when. As a result, the data is primarily different ways of tracking the requests made to retrieve different pieces of information held in these databases (a request is called a hit). Broken down in numerous ways, the data describes the requests by users' Internet Browsers to retrieve or perform other operations on files the server is responsible for.

Because the Libraries are dependant upon JSTOR and IAC sharing their data with us, we have little to no input on the format of the data or on its emphasis. The Libraries are just clients that subscribe to a service provided by these organizations. Being a client entitles community members to go to these sites and sort through their databases and retrieve whatever information desired. That relationship means no one in the Libraries has direct access to the provider's machines themselves. The organizations provide our librarians data on how their subscription is being used in the form of showing requests for various documents.

In order for the KDD steps using the Libraries data to make sense later, it is necessary to go through a cursory examination of what that data looks like first. Instead of showing all of it, representative pieces will show the nature of the data.

## JSTOR Data

This usage data is available on the web to anyone using a computer at our schools by entering the URL. Options in the data to be viewed are available through the form in Figure 1. A simple request for a usage summary for the month of March, 1998 yields Figure 2. As shown the information has been automatically

**JSTOR Usage Statistics Request Form**

Welcome! Your IP address: 152.65.167.59  
is registered with JSTOR as part of site: College of Saint Benedict/Saint John's University  
which has institution classification: Small

---

Please see the [Statistics Request Help](#) for an explanation of these options.

For the following time period:

Year 1998    All Months Combined

Show me:

- usage summary
- breakdown by journal title
- breakdown by hour of access
- breakdown by subdomain (where possible)
- breakdown by smaller time-units
- graph of use within this time period

Formatted as: HTML Tables

organized into HTML tables, but it can also be retrieved as plain text columns if the other format is desired.

**JSTOR Usage Statistics Request Form**

Welcome! Your IP address: 152.65.167.59  
is registered with JSTOR as part of site: College of Saint Benedict/Saint John's University  
which has institution classification: Small

---

Please see the [Statistics Report Help](#) for more information.

**Statistics for 1998/03**

**Accesses from Your Site**

	browsing			citations	viewing		printing			searches	total
	title-list	vol/iss	TOCs		pages	jprint	pdf	ps			
accesses	48	33	7	81	423	35	10	4	570	1211	

The 423 pages viewed were from 213 articles.  
an average of 2.0 pages per article.

**Usage Summary**

Including averages of "Small" sites like yours.

	browsing			citations	viewing		printing			searches	total
	title-list	vol/iss	TOCs		pages	(articles)	jprint	pdf	ps		
accesses from your site	48	33	7	81	423	213	35	10	4	570	1211
average of "Small" sites	22	35	23	40	220	108	25	8	0	197	571
totals from all sites	19441	37336	28347	21735	157571	74207	19327	9950	1136	103757	398600

While the format may appear confusing at first, the column headings are uniform from chart to chart to provide consistency. Besides the specific institution's data, a comparison can be made to the average accesses from all "Small" institutions. Broken down into headings, the data available is as follows:

*Browsing:* Accesses which may indicate browsing (searching) behavior-

By looking at the requests visitors have made for journals, volumes and issues, table of contents, and citations these numbers are here to give an idea of how much users are searching for materials. Since much can be gained by viewing the indexes of articles and not actually reading the full text, these numbers show user behavior as they search for materials related to their topic.

*Viewing:* Accesses which reflect that article pages were viewed online-

These numbers represent customers actually requesting article pages and the number of different articles they viewed. These numbers indicate customers that searched the index and found articles interesting enough to request in full format.

*Printing:* Number of articles printed or downloaded from the printing page in a special

printing format- These numbers indicate the number of times customers chose to use the special printing pages of JSTOR in order to retrieve the document in a special format. The postscript (ps) and pdf formats are both listed, as well as the requests that went through the JPRINT helper application in general.

*Searches:* Number of searches performed in all journals-



A more generic term, this shows broad user searching requests.

*Total:* Total number of accesses-

This number summarizes the others for direct comparison to other institutions and the JSTOR total.

Perhaps more usefully, these general statistical categories can be broken down to help provide more insight into user behavior.

Breakdowns Available by:

- ▶ *Journal Title:* The raw number of accesses for each journal in the JSTOR database.
- ▶ *Hour of Access:* The total accesses by hour of the day.
- ▶ *Subdomain:* Accesses from the different subdomains of the csbsju primary domain.
- ▶ *Smaller Time-Units:* The breakdown of large chunks into smaller units of time. For a year, this option breaks the time into months, for a month, this option breaks the month into days.
- ▶ *Visual Graph:* A graph comparing the accesses of the current time period to the previous time period.

A full set of the data available for March of 1998 is located in Appendix A. At <http://stats.jstor.org> the full data sets can be obtained for review.

### InfoTrac Data

The InfoTrac usage data has decidedly less flexibility. It arrives in the form of text files that are emailed to a representative of the library. There are no choices about how it should be formatted; you get the generic statistics that they choose to send out in simple text form. Instead of accesses, IAC is concerned with sessions, however, the accesses per journal are recorded in much the same way. Which will be important later for comparison purposes between JSTOR and IAC data. Here is a section of the data:

---

[ IAC SearchBank Data ]

#### LIBRARY DAILY ACTIVITY REPORT FOR May 1997

DATE	NO. OF LOGINS	TOTAL CONNECT TIME Hours
1-May-1997	2	0.4
2-May-1997	14	1.8
3-May-1997	2	0.2

#### LIBRARY HOURLY ACTIVITY REPORT FOR May 1997

HOUR	NO. OF LOGINS
3-4 PM	15
4-5 PM	14
5-6 PM	8
6-7 PM	11

MONTHLY DATABASE USAGE  
Summarizes usage by database per month  
for the month of March 1998

	Sessions	Views	Retrievals
Expanded Academic ASAP	3718	13698	2179
General BusinessFile ASAP	403	1486	118
General Reference Center	1765	6350	828
Health Reference Center Academic	60	210	15
ISI Current Contents	317	869	112
ISI Current Contents - This Week	34	20	0
PsycINFO	335	1530	260

JOURNAL RETRIEVALS  
Shows usage for each journal  
for the month of March 1998

Source	ISSN	"Views"		"Retrievals"		
		Total	Abs+Ct	Ftxt	PSta	
Psyclit Serial Records		1460	246	246	0	0
The Economist	0013-0613	529	113	91	22	0
Institute for Scientific Information		889	112	112	0	0

---

It is readily apparent that IAC has found a different way to record essentially the same type of data as JSTOR. The number of logins to the system is readily available, and it is easy to see from the 'HOUR' breakdown the most popular times of the day for people in the community. Associated with each journal is a count of citations retrieved for that journal along with its identification code.

Only a small part of the data available has been shown here to represent the broad categories it has. The JSTOR data can be represented on a few pages for each month (as shown in Appendix A), but the InfoTrack data for the same month of March is a 220

Kilobyte text file which translates to about 96 pages of text. This same amount of data arrives faithfully every month.

Now that the data has been laid out and it is clear what is available, the KDD approach to the situation becomes clearer. Is it useful to apply KDD to this data? Well, the data is certainly in a form that could be termed "non-human friendly." It has a lot of numbers and few categories- a situation that tends to cloud relationships in the human mind. On a practical basis, it is nearly impossible for a human to read this data and get anything out of it. Meaningful patterns are lost amid the data that is not sorted or arranged in a useful way. Asking what information can be extracted from the data might help.

As has already been stated, the goal of learning anything here is efficient distribution of resources. A computer, through KDD, might be able to arrive at a model that helps place these Internet resources in context with other library resources. There is a lot of it and most of the information is located in a few fields that are too number oriented to make much sense to a human brain.

From this preliminary look, it is obvious that the question is worth considering.

## 4. Steps Analysis

---

The goal of this section is to work through the steps of KDD with the hypothetical case of working with the Libraries data. To get started, each step and a short definition will be listed. After that, we'll go through each step more thoroughly and apply it to the Libraries data.

### Step Overview-

- *Getting to know the data and the task*: Designed for preliminary study of the scope of the project and data.
- *Acquisition*: Accomplishes bringing the data into a safe environment for analysis.
- *Integration and checking*: Needs to confirm that data's form and contents.
- *Data cleaning*: Inserted to remove obvious flaws in the data.
- *Model and hypothesis development*: Exploration of the data and selection of an appropriate model to use later.
- *Data mining*: "Application of the core discovery procedures to reveal patterns and new knowledge or to verify hypotheses developed prior to this step." (ACM)
- *Testing and verification*: Testing of predictive or descriptive capabilities on reserved data.
- *Interpretation and use*: Integration with existing knowledge and subsequent implementation.

For each step, a longer description will be given that locates the step's position within

KDD. Then the Libraries concern with this task will be presented, along with any special problems that may creep up. If the Libraries step is trivial or of medium difficulty, another example will be provided that shows why that step is important.

### **Studying the Data and the Task**

At this stage of the process it is important to just be able to discover the scope of the project that is being undertaken. Depending on the situation, this may be very daunting. In particular, a consultant coming to work for a company may have to do quite a bit of study before they are prepared to deal with any sort of process the company may use. This preliminary "taking stock" of the situation lets you sketch out the potential resources needed to accomplish the task and gather the general parameters for what you wish to learn from your efforts.

This stage is complicated by several factors. One of the primary issues is that many companies data is spread across internal divisions and systems of different types. Before even attempting to obtain the data, just deciding what data is relevant is a serious issue when it is not readily available. This can be magnified even further by the possibility of having workers that aren't completely familiar with the standard processes of the business.

In our Libraries' situation, this task may be fairly trivial, but still necessary. There are two main Internet Database resources that they subscribe to. Primarily they are interested in learning whether this investment is worth the cost. In order to discover that, they are interested in discovering how the resources are being used and to what extent they are useful

to the community. The data available are the usage statistics provided by the companies that sell the service. A key characteristic of that data is that it comes in a form the libraries have no control over; they must work with what they are given.

It is impossible to start any project before discovering the ground rules involved. Clearly defined parameters at this point in the analysis process can save time and make efforts more productive in the long run.

### **Acquisition**

For any experimentation to be done well, it is necessary to have a controlled environment that keeps your materials clean. This is no different with computers. In this step the data is brought into a so called "data warehouse" that the researchers have easy access to and that provides what they need to do their work. Consisting mostly of management software and hardware designed for high speed access, a data warehouse provides a controllable, regulated way to store data. This warehousing technique is a growing industry in its own right as companies store more and more data.

With the Libraries, this task of data acquisition is easily accomplished due to the readily accessible nature of the data. It is obtainable in a simple text format from both email (from IAC) and the web site of the JSTOR organization. A data warehouse certainly isn't required to hold it, but perhaps a dedicated account on one or more of the campus computer networks would suffice to keep the data separate from other information and secure it from people not familiar with how it should be treated.

This is definitely a step that takes on a whole new meaning for some applications of the KDD methodology. A large corporation may have data for multiple business processes that are all applicable to what they are trying to learn, but are located in different divisions, different states (or even countries) and perhaps most problematic: are located in different operating systems. The task of gathering data to one spot for analysis becomes immediately more complex when the software you wish to use stores data differently than the format you gathered it into. The data then has to be translated between formats prior to working with it.

An issue that may rapidly become a large factor is the support received for research from those different divisions. It is possible that the KDD processes initiated in marketing, for example, simply isn't supported by the data holders in sales. Such a situation may arise from simple differences in the priorities different managers give their employees. A far flung manager's sudden desire to have many employees in someone else's division spend days converting their data is likely to be met with resistance. Or perhaps they are worried about security of their data and maintaining their position in the company. In a very real worst case scenario, efforts to acquire reasonable data could be stymied for weeks.

In this stage, the groundwork for a good, directed KDD effort is laid, which is all important to the later steps. The environment you choose to work in sets the tone for how later stages will have to be handled. For example, are you going to move the data to a Windows based platform or a UNIX platform? Later, when the data mining is performed, that decision is going to have an inestimable impact on how you proceed.



## **Integration and Checking**

For many practical applications this is a "check point" phase that makes sure work is on track for the long term. It is necessary to review the data now that it has been moved to its permanent home and verify that it survived its trip in the expected format. It may also be necessary to convert the files to the standard format for the software you are going to use, and substantial checking is required after that phase to ensure the data itself did not get changed from its previous state. Another issue that needs to be dealt with here is planning for the verification stage of the KDD process. A data set that wasn't used in the analysis is needed to help verify results, and the easiest way to do that is to separate off a section of the data at this stage so that it isn't used by the computer. Later, this section can be pulled out and compared to the computers results to test validity.

In our Libraries' case, this task would almost certainly be a trivial example of why this KDD step is here. The amount of data is such that it could almost be visually verified by an analyst, though simple software could certainly perform the job more accurately and be reusable in the future. A possible difficulty is that the data is in a text format, so pulling the fields out of the files may be an issue, and verifying during this stage that field separators are located in the expected places would be important. Extra checking at this stage can prevent confusion later. It is also easy for the Libraries to separate out some data for use later. They'll need it to see if patterns hold accurate for months not analyzed.

For other applications however, this task could be anything but trivial. Verification of Gigabytes of data that possibly went through multiple transfers to get to its current home

must be done carefully to find subtle bugs. File corruption is a particular problem that must be watched for.

### **Data Cleaning**

This step culminates the preprocessing efforts required by KDD which may be up to 80% of the actual work involved in the whole process. Its goal is to remove obvious errors in the data that could cloud any information gleaned from it later. More specifically, fields that may be damaged or in other ways incomplete must be corrected before software can be turned loose on the data. The other major issue is that there may be duplicate entries in the data that could distort your results. It is possible for databases, especially those with customer information, to contain multiple entries for a single person due to inaccuracies in abbreviations that didn't get matched automatically by the system.

This last issue cannot be emphasized enough. A corporation's database can be very open with many different employees entering information for years. The data may go without being verified for years. Any data obtained from such a situation has to be checked thoroughly for such problems as having Jim Schnepf registered more than once due to two abbreviations being used for the word "Street" in his address. If one employee enters 'St.' and another enters 'Street', the system may record two Jims and permanently cloud his entry until someone discovers it later (if ever). One bank saved up to \$170,000 a month after just this step of KDD thanks to cleaning up their account information. Duplicate mailings were being performed to such an extent that fixing them alone probably paid for the investment in KDD.

In the Libraries' case, the work may be easier, but it is no less important. The transient nature of the Internet and the changing states of the IAC and JSTOR organizations both mean that any data coming from them is in a highly suspect form. As journals are added or removed, or the nature of the data itself changes over time, it must be carefully checked to make sure it is consistent. Decisions must be made on how to handle a journal entry that only appears after a certain time. Should it be discarded since it doesn't appear from the beginning, or should it have a null entry put into the data from earlier than its initial entry? Both options have their merits, but some uniform operation must be done to all of the data to guarantee no surprises later.

Within the KDD method this is the stage that exemplifies the Garbage In -> Garbage Out state of software and analysis with computers. It is extremely important, given the complexity of computers analyzing data, that the data have as few flaws as can be guaranteed by human overseers.

### **Hypothesis and Model Development**

This is the stage where the purpose of the Data Mining segment is decided. It is here where the model is chosen that will decide what the DM algorithm is actually going to search for.

### **Data Mining**

The DM process here is an automated attempt to find either a predictive or descriptive

model of the data. Of course this model is most useful if it answers questions that have already been decided need answering. Using various algorithms, the software will attempt to build a model that does what you want.

In the Libraries' situation, it has to be decided what kind of analysis is needed to obtain the information wanted from the data. The choice is not very hard to make, as a predictive result really wouldn't be useful for a number of reasons. First, JSTOR and InfoTrack both charge lump sums for access to their various databases. Once you have access, you can use that privilege a hundred times an hour or not at all. Either way, the same amount is charged. Therefore, the libraries care little about predicting usage of the resources because it doesn't matter monetarily if they get used less or more in the future.

Secondly, the data doesn't extend far enough into the past to create a reliable prediction pattern. For the two years we have data for, it is easy to assume that usage will remain roughly the same as past levels. Ultimately it comes down to what the Libraries are more interested in knowing. Since it is desirable to maintain these resources if possible, they are most interested in learning about how people have used them. If the databases are not being used to their potential, then advertising could possibly bring the use up. If they are being used sufficiently, for which criteria doesn't yet exist, then the status quo is acceptable and the databases can be left alone.

In the end, what does matter to them is how these resources are being used at all, which means a descriptive result is desirable. The Libraries are interested in the amount these resources are being used and the level to which they are useful to the community.

## **Verification**

In this step, the data carefully separated from the rest back in the Acquisition stage can be compared against the computers conclusions to see if discovered patterns hold true for more data. Most important here is to guard against patterns that are only statistically relevant to the test data. Any patterns found must have greater applicability than that if they are to be trusted for business decisions of any kind.

## **Interpretation and Iteration**

The activities here depend much on how the previous two steps went. It is possible that statistically relevant results were found and those results were verified with the reserved data. If so, then the results need to be interpreted to answer the questions asked in the beginning. If not, then it must be decided what went wrong, or whether the data contains useful information at all.

That last is a more interesting problem because it means iterating through the KDD process again. If there is no fault found with the data, then the acquisition stages don't need to be repeated, but if the reason good results were not found is because there is a flaw in the data or it is incomplete somehow, then better data will have to be arranged for.

## **5. Analysis**

---

As the methodology was gone through step by step, it was apparent how some steps naturally seemed to help answer questions and solve problems. Other steps seemed superfluous to the Libraries situation. But now that the KDD methodology has been applied in this hypothetical format to the Libraries data, we are prepared to look at some of the questions asked at the beginning of the paper.

*Is KDD the methodology a useful one to apply to the Libraries situation?*

The answer to that question, like so many others dealing with computers, is both yes and no. To explain that answer, it must be remembered that KDD is a long process that attempts to solve a number of different tasks within its scope. Performing some steps turned out to be quite useful at organizing data and providing a point of view with which to work with it. Other areas seemed forced, where the data would almost have to be bent unnaturally to fit the methodology. A proper solution will have to take into account what worked and what didn't on a step-by-step basis.

Some steps of the process that definitely could help deal with this data are all of the preprocessing steps that mark the beginning of KDD. This includes Getting to Know the Data, Acquisition, Integration, Data Cleaning, and Model Development. The overriding reason these are useful is that the data is difficult to compare even to itself as long as it remains in separate, unexamined formats that cannot be compared easily. Preprocessing the

data allows for error checking and the chance to learn the data well.

With the Getting to Know the Data step, a librarian (temporarily an analyst) can sit down and become familiar with the definitions of the categories and the general parameters of the data. Proceeding to Integration and Cleaning means the data can be brought to a stable environment to be stored for the long term. Once it has any errors cleaned out of it, it is safe to store for later use- comparing it against future data so trends can be tracked easily. Proceeding to Model Development would allow the Libraries to know how they are interpreting the data and incorporating it into their other business decisions. Long term comparisons are only meaningful if they are made in a regular, defined way and using a model would serve to keep these comparisons stable.

After preprocessing however, it is apparent that the Data Mining stage is problematic. The level of complexity in the data is not what neural net or other machine learning algorithms were designed for. That does not mean that the stage is useless, just that it needs some refinement for this particular case. The data, for example, can definitely use some automated processing- especially the IAC data files with their monthly 90+ page length. But does the computer need to analyze the data on its own? Let us consider what could be learned from models developed by a human analyst without a computer to try and answer that.

Perhaps the most useful thing that could be done is to attempt to categorize users. The journals that come from these databases tend to be subject specific such as the Far Eastern Economic Review and National Catholic Reporter, a potentially useful comparison

would be to see which *subjects* were the most popular for searching. This would allow us to see if there is a marked difference in accesses between subjects, which would indicate classes of community members that were using the Internet differently. To increase usage, the first thing to do would be to increase advertising and Internet awareness among the people interested in the subjects that currently receive the least amount of attention.

It must be acknowledged that this community is based at two liberal arts colleges. Cross disciplinary searches by students are not only possible, they are extremely likely as students not familiar with a field turn to the Internet's searching mechanisms to help them find relevant materials. Despite this sort of behavior, these broad subject categories would be a place to start grouping customer behavior.

Another interesting facet of behavior would be to track the pattern of uses from subdomains on campus. JSTOR includes information on where the accesses occurred from, which of course is mostly the generic pc domain. However, some come from dorms. An increase in searches from the dorms could mean that the Internet is becoming more dominant and readily accessible to students in their homes. Such a sign might indicate the Library should spend more time increasing their Internet resources so the newly aware students have access to a greater amount of material.

The obvious pattern that indicates general usage is the overall trend of accesses. Both JSTOR and IAC provide a sort of general login data that can be used month to month to show whether people are using them at all. However, these numbers are really only useful when compared to themselves. That means they can show, in a relational way, if the usage



goes up or down from time period to time period. They cannot readily be used to compare to any other data because it is too specialized, though it can be used to compare our institution to other schools. If there are significant differences in the data between our school and others, there is likely something different between us. This difference might be traced to the subject matters of the primary majors, level of advertising of the resources, or the staff pushing the Internet more or less.

If all of those things can be learned from this simple appearing data, what cannot be learned? Here, the answer is rather clear that there is no way to know if users are finding materials they can use. Two reasons contribute to this and the first is that the printing stats only indicate how much users actually printed (and that is only if you assume they can use what they print). These stats only indicate trends because articles and other pages can be printed through the browser and the JSTOR/IAC servers would have no way of tracking that to report it. The second reason is that high stats do not equal satisfaction on the part of the customer. Perhaps students were rushed and printed everything they saw, but in the end were not able to use any of the materials they retrieved.

How can satisfaction be tracked? This is a problem that doesn't have an obvious solution. Surveying the users directly to see if they are satisfied has its problems because students have different standards. Some would be depressed that the computer didn't want to do their research for them, leading them to be dissatisfied even if they found fifty articles of excellent material. It would seem that the best temporary solution is for the librarians to objectively evaluate the databases themselves or to work with students regularly and record

how satisfied the students were with the materials located on the Internet.

After considering the data in this way, it has to be determined that the data simply is not complicated enough to warrant true Data Mining, though software that compiled the data into our models would certainly help. A small, custom built package by someone at the colleges could readily compile the data into the categories discussed above so the records are available and compiled for the long term.

As the KDD process continues though, the steps become more useful again. Verification would make sure our models are describing real user behavior, and Integration would ensure the new lessons learned are incorporated into action.

Using these hypothetical models, it is easy to see how some of the questions asked at the very beginning might begin to be answered through this analysis process.

- Should they [the Libraries] keep expanding Internet material availability?
- Should they advertise Internet materials more or less?
- Who are using the materials and for what?
- Are these materials worth the investment?
- Should the investment in Internet materials be increased, decreased, or modified?
- Which, among the multitude of resources, are being used most productively?

After learning a bit about usage patterns, the questions about expanding materials, and advertising could be answered readily.

In the end, the KDD methodology is useful, if not the Data Mining step itself.

## Appendix A

---

### JSTOR Usage Statistics Request Form

Welcome! Your IP address: 152.65.167.59  
 is registered with JSTOR as part of site: College of Saint Benedict/Saint  
 John's University  
 which has institution classification: Small

Please see the Statistics Report Help for more information.

#### Statistics for 1998/03

##### Accesses from Your Site

----- browsing -----		viewing		---- printing ----					
titles	vol/iss	TOCs	citations	pages	jprint	pdf	ps	searches	total
48	33	7	81	423	35	10	4	570	1211

The 423 pages viewed were from 213 articles,  
 an average of 2.0 pages per article.

##### Usage Summary

Including averages of "Small" sites like yours.

----- browsing -----		-- viewing --		---- printing ----					
titles	vol/iss	TOCs	citations	pages (articles)	jprint	pdf	ps	searches	total
your site	48	33	7 81	423 213	35	10	4	570	1211
avg "Small"			22	35	23	40	220		108
25	8	0	197	571					
all sites		19441	37336	28347	21735	157571		74207	
19327	9950	1136	103757	398600					

##### Accesses by Journal

-- viewing --		----- printing -----		---- browsing ----					
pages (articles)	jprint	pdf	ps	total	vol/iss	TOCs	citations		
===== Asian Studies =====									
Journal of Asian Studies					3	0	2		
18	13	0	0	23					
Far Eastern Quarterly					0	0	4		
8	5	0	0	12					
				Total					
				35					
===== Ecology =====									
Annual Review of Ecology and Systematics					0	0	2		
32	3	0	0	34					
Ecological Applications					0	0	1		
5	4	0	0	6					
Ecological Monographs					0	0	0		

Data Mining on Electronic Media Usage Statistics

Ecology	1	1	0	0	0	1	2	0	14
	20	10	2	0	0	38			
						Total			
						79			
===== Economics =====									
The American Economic Review	20	9	2	0	0	28	4	2	0
Econometrica	0	0	0	0	0	2	1	0	1
The Journal of Economic History	16	7	0	0	0	16	0	0	0
The Journal of Economic Perspectives	0	0	0	0	0	4	4	0	0
The Journal of Industrial Economics	5	2	0	0	0	9	3	1	0
The Journal of Political Economy	10	7	1	0	0	16	3	0	2
The Quarterly Journal of Economics	17	7	2	0	0	20	0	0	1
The Review of Economics and Statistics	24	6	1	0	0	27	0	0	2
						Total			
						122			
===== Education =====									
The Journal of Higher Education	10	4	0	0	0	21	5	2	4
						Total			
						21			
===== History =====									
The American Historical Review	32	18	5	1	0	41	2	1	0
The Journal of American History	23	6	0	0	0	24	1	0	0
The Mississippi Valley Historical Review	10	3	0	0	0	11	0	0	1
The Journal of Economic History	16	7	0	0	0	16	0	0	0
The Journal of Military History	1	1	0	0	0	1	0	0	0
Military Affairs	14	6	0	0	0	15	0	0	1
Renaissance Quarterly	4	3	2	1	0	7	0	0	0
Renaissance News	1	1	0	0	0	1	0	0	0
Speculum	21	11	0	0	0	22	0	0	1
The William and Mary Quarterly	3	3	0	0	0	4	0	0	1
						Total			
						142			
===== Mathematics =====									
Journal of the American Mathematical Society	0	0	0	0	0	3	3	0	0
SIAM Review	6	5	0	0	0	7	0	0	1
						Total			
						10			
===== Philosophy =====									

Data Mining on Electronic Media Usage Statistics

The Journal of Philosophy						0	0	6
35	22	0	0	0	41			
Philosophy and Phenomenological Research						0	0	2
25	9	0	0	0	27			
					Total			
					68			
===== Political Science =====								
American Journal of Political Science						0	0	3
15	12	5	0	0	23			
The American Political Science Review						2	1	1
3	3	0	0	0	7			
Journal of Politics						0	0	1
2	1	0	0	0	3			
World Politics						0	0	4
5	3	2	0	0	11			
					Total			
					44			
===== Population Studies =====								
Demography						0	0	3
1	1	0	0	0	4			
Population and Development Review						0	0	0
4	3	0	0	0	4			
Studies in Family Planning						0	0	0
2	2	0	0	0	2			
					Total			
					10			
===== Sociology =====								
Contemporary Sociology						0	0	1
5	4	0	0	0	6			
American Sociological Review						0	0	8
4	2	0	0	0	12			
Annual Review of Sociology						0	0	5
3	3	0	0	0	8			
					Total			
					26			

Accesses by Hour of the Day

Hours are in EST (-0500)

hour	accesses
00:	43
01:	44
02:	6
03:	0
04:	0
05:	3
06:	11
07:	0
08:	0
09:	9
10:	17
11:	34
12:	52
13:	80
14:	64
15:	102
16:	148
17:	128
18:	91

Data Mining on Electronic Media Usage Statistics

```

19:          54
20:         134
21:          85
22:          50
23:          56
    
```

Accesses by Subdomain

```

[Unknown]          2
bdorm.csbsju.edu  27
jdorm.csbsju.edu   3
libr.csbsju.edu    2
pc.csbsju.edu     1172
physics.csbsju.edu 5
    
```

Accesses Over Time

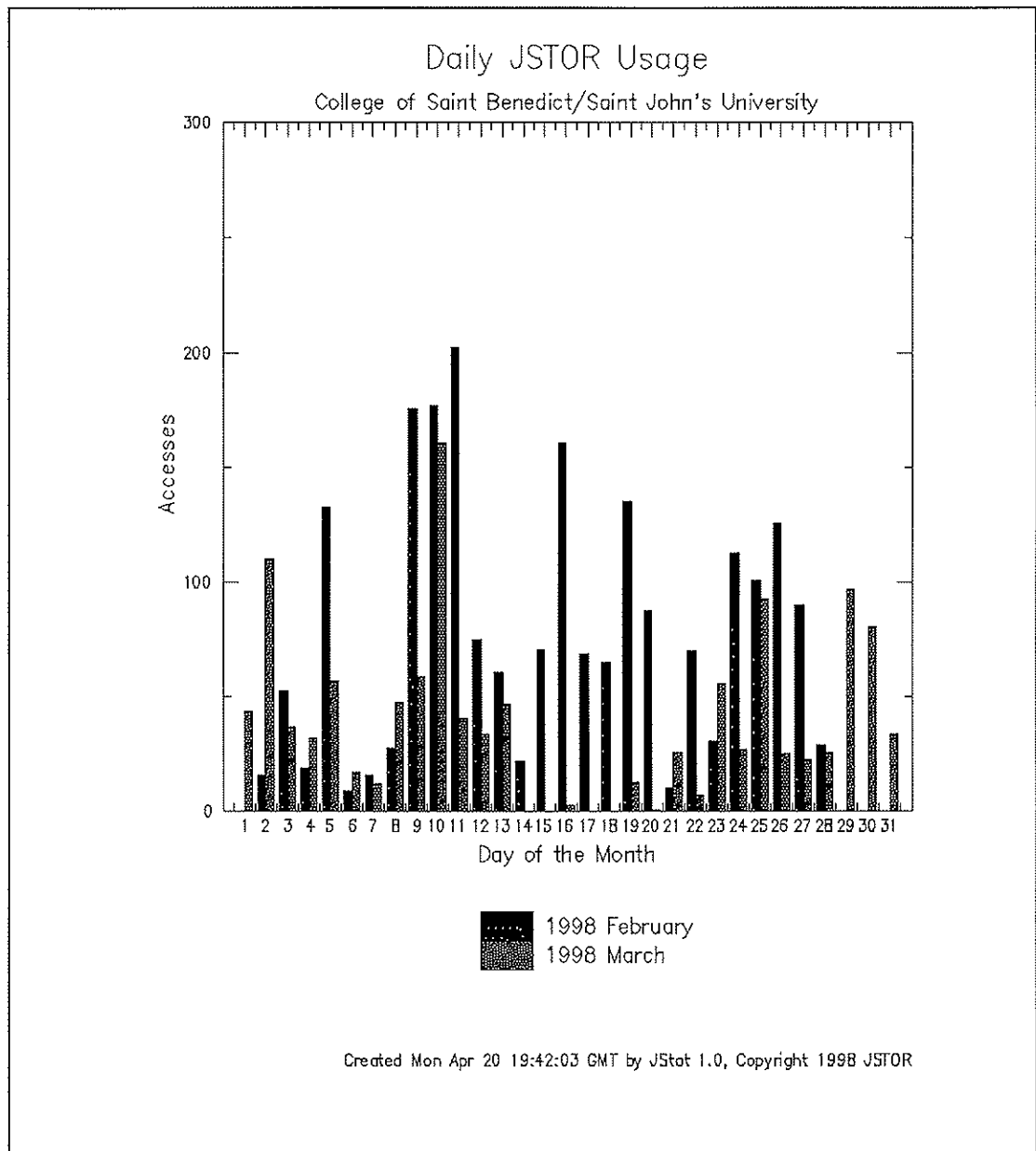
If a time period does not appear, then it contained no accesses from your site.

```

----- browsing -----      -- viewing --
---- printing ----
      titles  vol/iss      TOCs citations  pages (articles)
jprint  pdf      ps searches  total
Day 01      0      0      0      0      1      27      10
      2      0      0      14      44
Day 02     10      6      1      7      24      14
      1      0      0      61     110
Day 03      3      6      2      0      3      2
      3      0      0      20      37
Day 04      0      0      0      0      16      16      4
      0      0      0      16      32
Day 05      2      0      0      2      13      9
      4      1      0      35      57
Day 06      0      0      0      0      1      4      4
      1      0      0      11      17
Day 07      1      1      0      0      0      2      2
      0      0      0      8      12
Day 08      1      0      0      5      9      8
      0      0      0      33      48
Day 09      2      0      0      17      14      12
      0      0      0      26      59
Day 10      1      3      0      25      81      50
      5      0      0      46     161
Day 11      2      0      0      5      6      6
      0      0      0      28      41
Day 12      3      2      1      6      15      5
      0      0      0      7      34
Day 13      2      2      1      1      20      10
      0      0      0      21      47
Day 16      1      0      0      0      0      1      1
      0      0      0      1      3
Day 19      0      0      0      4      1      1
      0      0      0      8      13
Day 20      1      0      0      0      0      0      0
      0      0      0      0      1
Day 21      2      2      1      1      10      8
      0      0      0      10      26
Day 22      0      0      0      5      7      2      2
      0      0      0      5      7
    
```

*Data Mining on Electronic Media Usage Statistics*

Day 23		4	0	0	0	1	22	12		
Day 24	0	2	0	0	0	29	56	8	2	
Day 25	1	0	0	0	0	16	27	55	18	
Day 26	0	1	0	1	0	35	93	4	2	
Day 27	0	1	0	2	0	19	25	3	3	
Day 28	4	1	1	0	0	11	23	7	4	
Day 29	0	1	0	0	0	18	26	19	8	
Day 30	13	5	8	0	4	0	47	97	49	12
Day 31	1	2	0	7	0	0	21	81	8	4
	0	1	0	1	0	0	24	34		





## Appendix B

---

### IAC term definitions

*Location ID* - Unique identifier for a library branch or group of users.

*Total Sessions* - Number of logins to SearchBank.

*Total Connect Time* - Length of SearchBank sessions.

*"Views"* - On-Screen views. Counted when user enters a citation to view more information. An article is counted once per session regardless of the number of times it was viewed.

*"Retrievals"* - Counted for each article printed from an Article (Print) Station or attached printer, downloaded as PDF or PostScript, or Emailed through the SearchBank service. This is not counted when a browser print is performed, unless the reformat option was selected from the Retrieval Screen. An article is counted once per session regardless of the number of times it was retrieved.

*Sessions* - For database usage, counted when a user enters a database within a session.

*"Retrievals"* - For database usage, counted for each article printed from an attached printer or Emailed through the SearchBank service. An article is counted once per session regardless of the number of times it was retrieved.

*Abs+Ct* - Abstract and Extended Citation Retrievals.

*Ftxt* - Full Text and Full Image and Compound Retrievals.

*PSta* - Article (Print) Station Retrievals

## Sources

---

- Asbrand, Deborah. "Is Datamining Ready for the Masses?" Datamation. v43 p66 Nov 1997
- Atre, Shaku. "A Dirty Business?" Computerworld. v31 p73 July 21, 1997
- Brachman, Ronald J. "Mining Business Databases." Communications of the ACM. v39 p42 Nov 1996
- DeJesus, Edmund. "Turn Computers Loose On Your Data . . ." BYTE
- Edelstein, Herb. "Data Mining: Exploiting the Hidden Trends in Your Data." DB2online Magazine. (<http://www.db2mag.com/9701edel.html>) 1997
- Edelstein, Herb. "Predicting Customer Behavior with Neural Nets." DB2online Magazine. (<http://www.db2mag.com/9701edel.html>) 1997
- Etzioni, Oren. "The World Wide Web: Quagmire or Gold Mine?" Communications of the ACM v39 p65 Nov 1996
- Fayyad, Usama. "Data Mining and Knowledge Discovery in Databases." Communications of the ACM. v39 p24 Nov 1996.
- Fayyad, Usama. "Mining Scientific Data." Communications of the ACM. v39 p51 Nov 1996
- Fayyad, Usama. "The KDD Process for Extracting Useful Knowledge from Volumes of Data." Communications of the ACM. v39 p27 Nov 1996
- Gaudin, Sharon. "Data Mining Lifts Ski Marketing." Computerworld. v31 p43 Sep22, 1997
- Gerber, Cheryl. "Excavate Your Data." Datamation. ([http://www.cs.bham.ac.uk/~amp/dm\\_docs/dm\\_intro.html](http://www.cs.bham.ac.uk/~amp/dm_docs/dm_intro.html)) Sep 10, 1997.
- Glymour, Clark. "Statistical Inference and Data Mining." Communications of the ACM v39 p35 Nov 1996
- Hurwicz, Mike. "Dirty Data is Dangerous." BYTE. January 1997
- Imielinski, Tomasz. "A Database Perspective on Knowledge Discovery." Communications of the ACM. v39 p58 Nov 1996

*Data Mining on Electronic Media Usage Statistics*

Inmon, W.H. "The Data Warehouse and Data Mining." Communications of the ACM v39 p49  
Nov 1996

Knowledge Discovery Mine (<http://www.kdnuggets.com>)

Koren, Serg. "Concept Explorer Guides Searches Via Analytical and Intelligent Queries."  
InfoWorld. v19 p72 Sep 15, 1997

Lee, Kerry. "The Darlings of Data." Computerworld. v31 p74 Dec 22, 1997

Levin, Carol. "Searching for meaning: how to craft the perfect query." PC Magazine v15 p36  
Sep 24, 1996

Millman, Howard. "Agents at your service." InfoWorld. v20 p77 Feb 16, 1998

Nash, Kim S. "Electronic Profiling." Computerworld. v32 p1 Feb 9, 1998

Shein, Esther. "Up-to-Date With Fresh-Squeezed Data." PC Week. v15 p83 Mar 9, 1998

Stedman, Craig. "Marketers Turn to Data Mining to Fine-Tune Product Pitches."  
Computerworld. v31 p28 Nov. 24, 1997

The Data Mine (<http://www.cs.bham.ac.uk/~amp/TheDataMine.html>)

Wilson, Lindsay. "Industry Specific Tools Emerging." Computerworld. v31. p67 Dec. 15, 1997