RC 20714 2/4/97
Computer Science/Mathematics 27 pages

Research Report

Data Mining the IBM Official 1996 Olympic Web Site

Sara Elo-Dean

IBM Advanced Internet Technology
Southbury, CT 06488

Marisa Viveros

IBM Research Division
T.J.Watson Research Center
Yorktown Heights, NY 10598

Research Division
Almaden - T.J. Watson - Tokyo - Zurich - Austin

Data Mining the IBM Official 1996 Olympic Web Site

Sara Elo-Dean
IBM Advanced Internet Technology
elo@gemini.ibm.com

Marisa Viveros
IBM T.J. Watson Research Center
viveros@watson.ibm.com

Abstract

Data mining of Web server logs offers the potential for deep analysis of visitor activity at a Web site, activity usually unknown before hand. For instance, the discovery of association rules can show how visitors make up groups with common interests, while a sequential patterns algorithm can uncover the most trodden paths at a site. In fact, the application of a variety of data mining techniques can extract information which no technique applied alone could generate.

This paper describes the application of a range of statistical and data mining techniques to the 1996 Olympic Web Site access logs. The paper also demonstrates and emphasizes the benefits of classifying documents hosted at a Web site. Classification provides a general, high-level description of the document collection, a description independent of individual documents. By augmenting documents with multiple descriptive attributes, the site can be analyzed from several complementary perspectives. The results presented in the paper are accompanied with suggested applications to other Internet services, such as merchandise servers, personalized pages, and on-line databases.

1. Introduction

With the increase of Internet servers, the server logs have become a new source for data mining applications. Data mining offers a deep analysis of the visitors' behavior, often unknown at the start of the analysis. These techniques unveil answers to questions such as: Do the visitors to the web site make up interest groups? How do visitors from various geographic regions differ? What are the most trodden paths through the site?

The Olympic Web Site is ideal for data mining for several reasons:

- The *heterogeneous set of documents* hosted at the site allows for varied patterns of visits
- The documents are *classified* by topic, medium and source, providing multiple views on the same set of documents
- The site hosted one of the *largest web audiences* ever, on the average 100,000 visitors per day
- A *global audience* allows discovering geographic differences in visits
- The *HTTP cookies* issued by the server permit tracking a single browser's actions

1.1 Overview of the Document

This document provides insights into the activity at the Olympic Web Site. It presents the results of a range of statistics and three data mining techniques. Before viewing the results, it is helpful to understand the characteristics of the data on which the analyses were performed. Section 2 describes the classification scheme of the Olympic Web Site documents and the format of the server logs.

Section 3 presents the general statistics. To begin, Section 3.1, *Views Based on Classification*, demonstrates an advantage of classifying documents at a web site: it allows constructing multiple views on the site activity. The Olympic Site activity is shown by source and by sport.

In Section 3.2, *Visibility,* the activity at the site is viewed from the perspective of a content provider. Hit rates to content providers' pages measure their visibility at the Olympic Web Site. The section also suggests further data mining analyses for content providers.

Resolving the numeric IP addresses to names allows a break down of the activity by region, shown in Section 3.3, *Analysis by Geographic Region*. The experiment shows that web spectators at the Olympic Games are a lot like their flag-waiving compatriots in the bleachers.

Section 4 describes the application of data mining techniques to the Olympic Web Site. When readers view pages on Badminton, what other sports are they likely to visit? Section 4.1, *Association Rules*, presents discovered rules and suggests uses for real-time discovery of rules, such as a "Suggested Reading" section on a Personal Page.

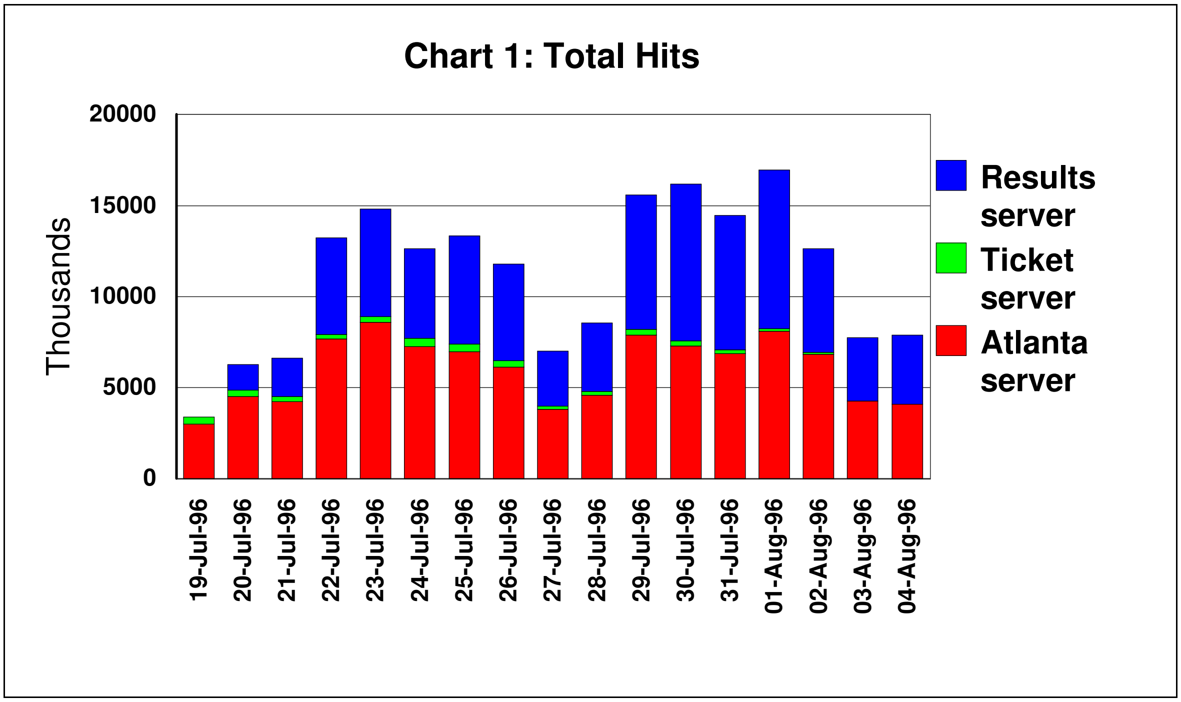
Section 4.2, *Clustering Web Site Visitors*, describes an attempt to group the visitors by sporting interests. The sparse data produces little new information after discovering association rules.

A sample of the most trodden paths at the Olympic Site are uncovered in Section 4.3, *Sequential Patterns.* The sequential patterns algorithm can also help in evaluating web page design and site architecture.

The results of the performed data mining analyses are accompanied by suggested applications to IBM Internet services, such as a merchandise server, personalized pages, and on-line databases. The concluding section encourages the incorporation of data mining into Internet applications in order to add value to users.

1.2. Overview of the Site

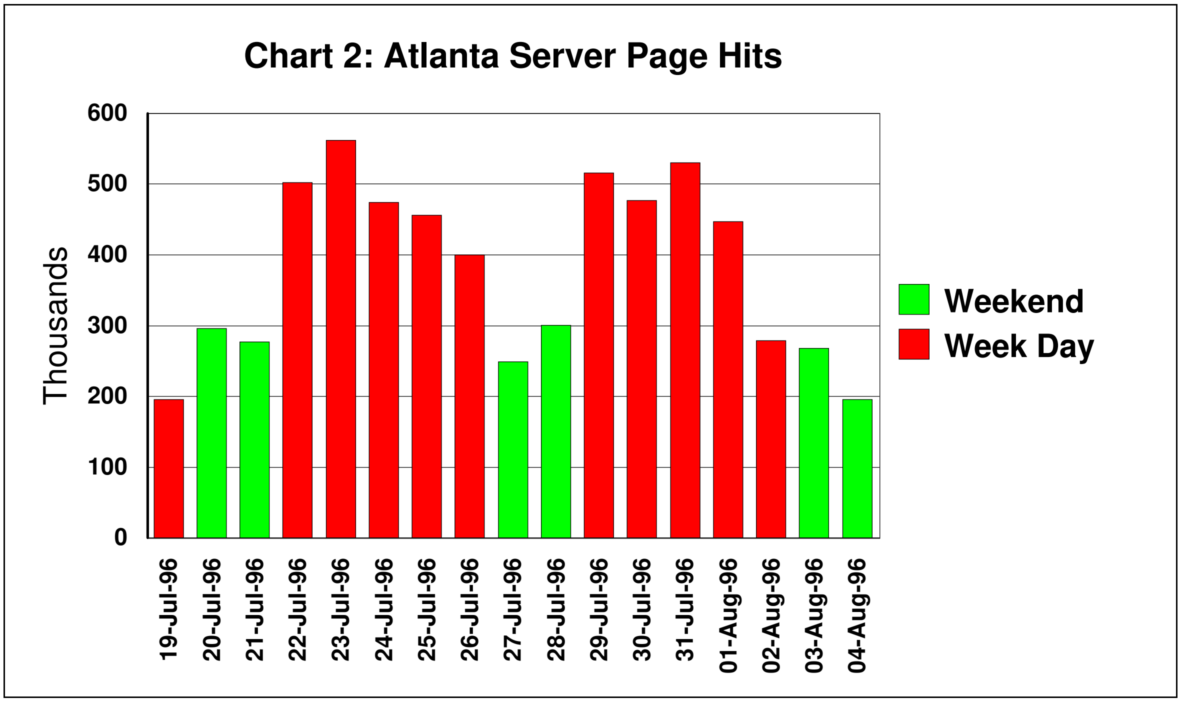
Let's start with the big picture: the overall hit rates, the most common and simplest measure of activity at a site. Chart 1 shows the daily hits, tallied at game time, to the three main Olympic services, the Atlanta News Server, the Results server and the Ticket server.



Generally, a hit is a single request to a server. In the rest of our analysis, however, we prefer to use a more modest and more meaningful definition: *one hit is one request to an html page*. The imbedded graphics and icons are not counted as separate hits. While the number of imbedded elements varies greatly from site to site, pages are usually designed to be conveniently viewed in one go. Thus, counting hits to pages permits more accurate comparisons between sites.

Chart 2 shows the daily hits to pages at the Atlanta Server. Observe the fluctuations between week days and the weekend, with the busiest days midweek during both weeks. The hits to pages represent roughly 6.5% of the total hits to the Atlanta Server. On the average 15 objects, such as scripts, logos, images, video or audio clips, banners and icons were imbedded in an html page. The popular Atlanta Home Page (25% of page hits) and the Sneakpeek pages (30% of page hits), both with many graphics, explain this high ratio.

The Olympic Site was run on IBM's Web Object Manager, or the WOM server. For details on WOM, see http://www.womplex.ibm.com. WOM issued HTTP cookies to requesting browsers. During the two week period, the daily requests came on the average from 100,000 unique browsers. As detailed later, there is not a one-to-one mapping between cookies and individual readers, but a count of the cookies allows us to approximate the number of visitors to the Olympic Site.



The Atlanta Site hosted roughly 1000 html documents one week before the Opening Ceremonies on July 19, 1996. At the closing of the Games, on August 4, 1996, the number of documents had grown to about 9000. Documents from various content providers were added at a continuous flow during the Games.

Given the high activity and the large set of documents, we now begin to “slice” the busy site to uncover the distribution of activity and the patterns in the visitors’ behavior at the site. Because of some missing and corrupt data for the first week of the Olympic Games, we performed the analyses on the Atlanta Server for the week beginning July 28.

The next section outlines the classification scheme of the documents. It also explains the format of the logs and the data preprocessing for the analyses. If you wish to go directly to the experiments and results, please proceed to Section 3 - Multiple Views.

2. Data Description

2.1. Classification of Documents

Classifying documents at a web site means attaching descriptions to the objects that make up the documents. A meaningful classification may include:

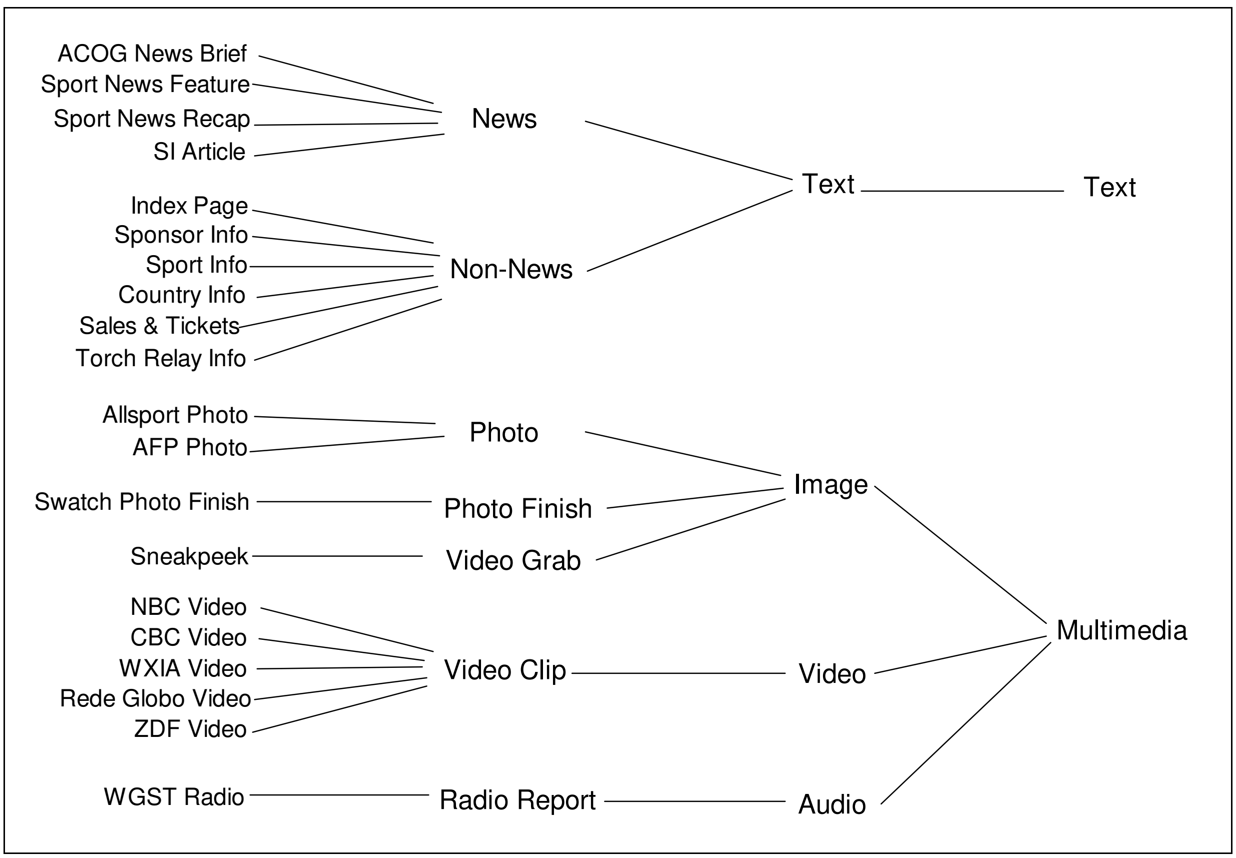
- *syntactic* descriptors, such as the medium of the document, its source, creation date, or importance
- *semantic* descriptors, such as the topics, actors or actions depicted by the object
- *layout* descriptors, such as “sidebar” or “footnote”, to ensure the appropriate layout at time of display

Without a classification, a document is only described by its url. A well thought out url naming scheme can describe a small number of features of a document. However, as the description of the content becomes richer, the definition of each page or object needs to be stored in the server. The WOM server stores an object’s classifications as *meta-data*, a data structure that holds all the characteristics of the object.

For the Olympic Site the WOM meta-data included classification by source, time, medium, and the main topics -- sports, countries and athletes. Classification provides the server knowledge about the stored objects. The richer the classification, the more intelligently processes can act upon the objects. For example, to automatically build the dynamic Country Pages for the Olympic Site, a process retrieved all objects with a given country description, grouped the objects by medium, and presented them in inverse chronological order on the Country Page.

While classification provides a common language to describe all objects at a site, a hierarchic classification, or *taxonomy*, also groups similar objects together. For example, in Figure 1, the taxonomy for the source classification at the Olympic Site allows accessing objects with a specific classification, for instance all NBC Video Clips, or with a general classification, such as all Multimedia documents.

Figure 1: Source Taxonomy



After the Olympic games, the html documents at the site were classified based on their source and the main sport and country to which they relate. For example, the AFP photo page of Romanian Olympic team gymnast Alina Ramona Stoica is classified as “Artistic Gymnastics”, “Romania”, and “AFP Photo”. The classification was done partly automatically by a process that detects the sport codes from the url and the athlete names from the caption. Editors manually classified all photos by sport because the captions did not employ standard names for the sports.

In order to manage the documents at the Olympic Site, a set of editorial tools were developed. Although the tools were not completed in time for these games, such a tool kit would help in the efficient content management of any web site. The tools should allow editors to

- define a taxonomy
- classify new objects using the taxonomy when uploading to the server
- query the server to retrieve objects based on their classification
- modify objects' classification and edit the objects themselves
- create page templates that dynamically generate pages made up of classified objects

2.2. Server Log Format

The statistics and data mining were performed on the logs from the 50 SP2 nodes at the principal Olympic Site located in Southbury (http://www.atlanta.olympic.org) and the relatively low-traffic mirror sites at Keio, Japan, Karlsruhe, Germany, Cornell, USA, and Hursley, UK. For every request from a client browser, the WOM server recorded in the log the following information:

server name:	Atlanta, Keio, Karlsruhe, Cornell, or Hursley
browser:	WWW browser type and version
url:	url requested
timestamp:	the time the server received the request
cookie:	the name of the cookie issued by the Olympic Server to the browser
server ip:	the IP address of the node of the server that handled the request
client ip:	the IP address of the client that submitted the request
bytes served:	the number of bytes served to fill the request
referring url:	the previous url requested

The daily logs were merged and stored on an OS/390 system. To save disk space, the logs were compressed to a fixed length format by replacing variable length strings, such as the browser name and url name with a code to a lookup table. The client and server ips were converted to a fixed length hexadecimal representation with a lookup table of the resolved ip name. Other lookup tables contain the classifications for each html document and the conversion from the client IP address to the resolved domain name. The current compressed format is shown in Table 1.

Table 1: Log Format

server	browser	url	timestamp OS/390	timestamp WOM	cookie	server IP	client IP	bytes
1	5709	2848	AD3DBFAF47A62A07	838702868	MHP2739623629	CC92B00C	CAE656FE	5073
1	1667	9142	AD3DBFAF47BED607	838702867	DXD2739313620	CC92B028	9D50497A	4002
1	5918	8776	AD3DBFAF47B25207	838702868	QQI2729061630	CC92B00C	80EBE4A9	19093
1	2144	10898	AD3DBFAF47CBAE07	838702868	YCS2739725642	CC92B028	89DB4911	2750
1	1936	4580	AD3DBFAF47D86E07	838702868	QKN2737898646	CC92B028	CC76F237	4980

Before applying the data mining algorithms, a considerable amount of work went into preprocessing the log files in MVS. With Perl scripts, the data was transformed to the input format of the three data mining techniques. The Perl scripts generate new files which are then transferred to the AIX platform to perform data mining. The data mining experiments could have been performed on MVS, but the client was not available at the time the experiments were conducted.

Specifically, to generate an input file used by the association rules algorithm, a Perl program accepts as input parameters the type of classification, date, country extension, and a flag indicating sneakpeek documents. For instance, when using the input parameters "sport", "210", "BR", "0", the Perl program scans the log files for day 210, or July 28, 1996. It selects records containing the client's IP address with the Brazil country extension, and non-Sneakpeek identifier. From the selected records are extracted only the unique identifier (or cookie) and the sport classification assigned to the url, if one exists.

3. Multiple Views

The previous section detailed the classification scheme of the documents at the Olympic Site. Before starting the analysis of the Olympic Site using these classifications, we needed to establish the assumptions on which we based the analysis.

Assumptions

How to identify a visitor?

We identify a visitor to the Olympic Site by an HTTP cookie issued by the WOM Server to the client browser. Since the WOM Server logged the cookie of every request, the requests coming from any given browser can be reconstructed. It is impossible to know whether more than one person visits the web site using the same instance of a browser. However, cookies with reasonable hit rates to originate from a single user largely outnumber the "public library cookies" of several hundred hits a day. Thus, during data mining the reasonable cookies outweigh the multi-user cookies. Naturally, we cannot analyze the traces from browsers that do not support cookies at all (about 5%), or those that turn the capability off .

Doesn't the organization of the site bias where people go?

Naturally, the layout of the site influences the sports people follow. A headline on the home page will attract more people than the same headline among other titles in the news section. Four headlines and eight photos got special treatment for 3-4 hours at a time by being displayed on the home page or in the photo album. The rest of the articles and background information were accessible through index pages that lead in an alphabetical order to all sports or countries. The photos were accessible through index pages arranged by the photo editor. In interpreting the activity at the site, we need to keep in mind that the editorial slant is always present.

The Sneakpeek camera was only live during the Atlanta day time hours. How does one account for the site looking different for audiences in different time zones?

Apart from the Sneakpeek live video grabs, the Olympic Site presented the same content to all audiences. While Sneakpeek only presented sports live at the time, the rest of the site presented the 37 sports more evenly. Because Sneakpeek was very popular, it heavily biased the Atlanta day time hours. Thus, in the analyses, we excluded requests to Sneakpeek video grabs in order to observe differences in activity given the same set of offered documents.

Can we assume a person is 'interested' when they click on a link?

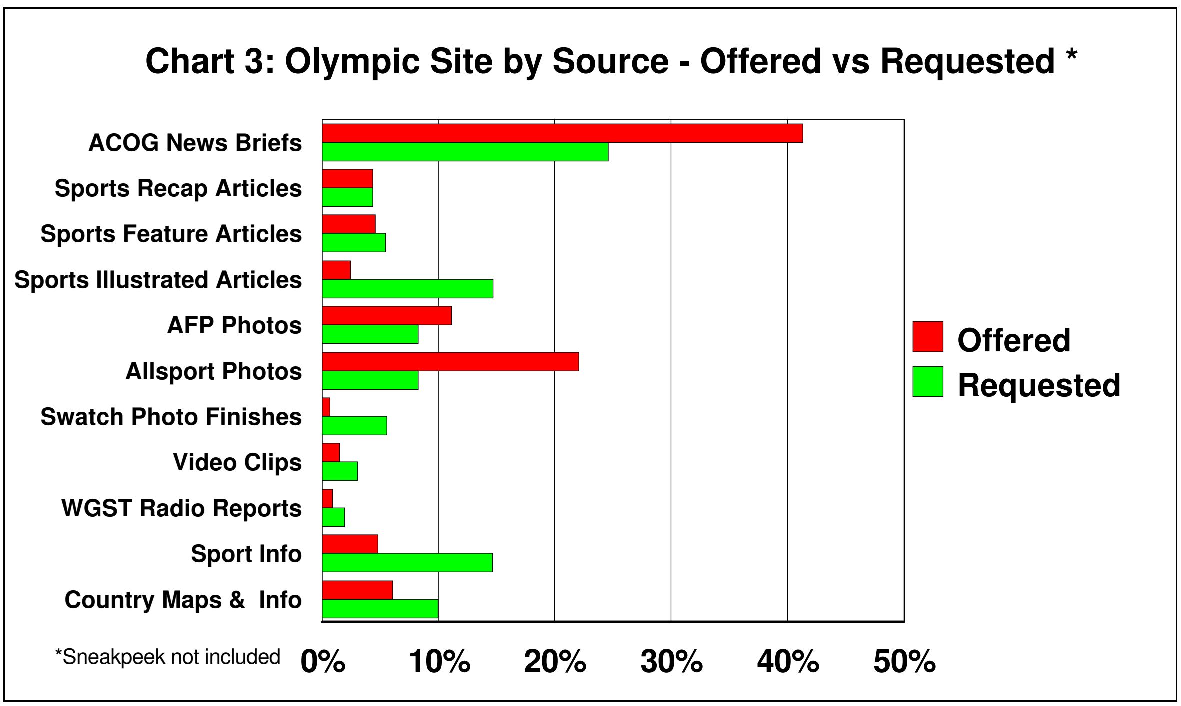
There is always a certain amount of random clicking and jumping, but we can assume that a reader who requested a document did so on purpose. We performed the data mining on the news articles, photos, video clips, sneakpeek images, and background information. None of these documents are behind "surprise links". The index pages leading to the documents indicate the media type as well as the sport, and often the country the document relates to. Readers know what to expect when they click on a title to retrieve the full article, a caption and a thumbnail to retrieve the full picture, or a sport icon to retrieve background information on the sport. Thus, requests of these pages give some indication of the reader's interests.

3.1. Views Based on Classification

Given the above assumptions, we can begin to construct diverse views of the site. As explained in Section 2, multiple classifications are attached to each document. Let's think of the classifications based on source, sport, and country as the axes of a three-dimensional space. Each document occupies a position in this space. If the document collection is projected to any one of the three axes, we can observe the distribution of visitor activity along the selected classification. We call such a projection a *view*.

View by Source

Chart 3 illustrates the activity at the site based on the source classification. The dark grey series represents the distribution of the eleven types of documents from content providers at the site. The light grey series shows the distribution of user requests. By contrasting the two, the chart highlights the sections of the site where supply was higher than demand or vice versa. For example, the small collection of Sports Illustrated articles got in proportion more readers than the large collection of ACOG News Briefs. Quantity does not always guarantee interest. Likewise, the few Swatch Photo Finishes attracted almost as many viewers as the large archives of Allsport and AFP Photos. The collections of background information on sports and countries attracted about 25 % of this activity -- visitors sought general knowledge, not merely the latest reportage from the games. Based on the interest in these sections, one can justify offering background information at future sites.



The popular Sneakpeek, frames grabbed from live video from the venues, attracted huge audiences. Regardless of the sport live at the time, roughly 30% of the daily activity at the site took place on Sneakpeek pages. With a new image grabbed every 10 seconds, Sneakpeek documents were not archived on the server. Thus, it is difficult to compare the number of Sneakpeek documents to static documents offered.

Sneakpeek is an interesting medium, a form of interactive television on the Internet. While the broadcast is live and continuous, the viewers have control over the speed at which they request the images. An illustrative example of Sneakpeek as TV is a viewer connecting from Hong Kong University to watch a judo match nearly live by reloading the video images over a hundred times in a row. But at any time, other viewers watching the same judo match could stop and save the shot of their favorite athlete. Using Sneakpeek as VCR, they can pause the broadcast, and catch up with the live broadcast at their own will. Although the viewers cannot affect the contents of the broadcast in the way interactive TV visionaries hope, they have control over their own Sneakpeek experience.

View by Sport

Not all sports were covered equally at the Olympic Site. One would think that the uneven distribution would have biased readers toward the more widely covered sports. However, contrasting the distribution of offered sports and requested sports in Chart 4 shows that that is not the case. One of the more widely covered sports, Boxing, never received the anticipated audience. On the other hand, sports such as Wrestling and Artistic Gymnastics received more hits in proportion. One can also observe relative differences among disciplines of the same sport. The three types of cycling were rather modestly covered at the Olympic Site, however, visitors preferred the trendy mountain cycling.

This section has demonstrated that classifying documents at a web site allows generating rich statistics. When each object is assigned a number of classifications, the collection of documents and the activity can be viewed in multiple ways. Views of the site activity along different classifications satisfy the curiosity of various parties. For example, a sports analyst may request the sport view while a content provider may find the source view more useful.

The rest of Section 3 uses the classification scheme to compare content providers' coverage and to observe geographic differences in visitor activity at the Olympic Site. The experiments in Section 4 exploit document classification to enrich data mining.

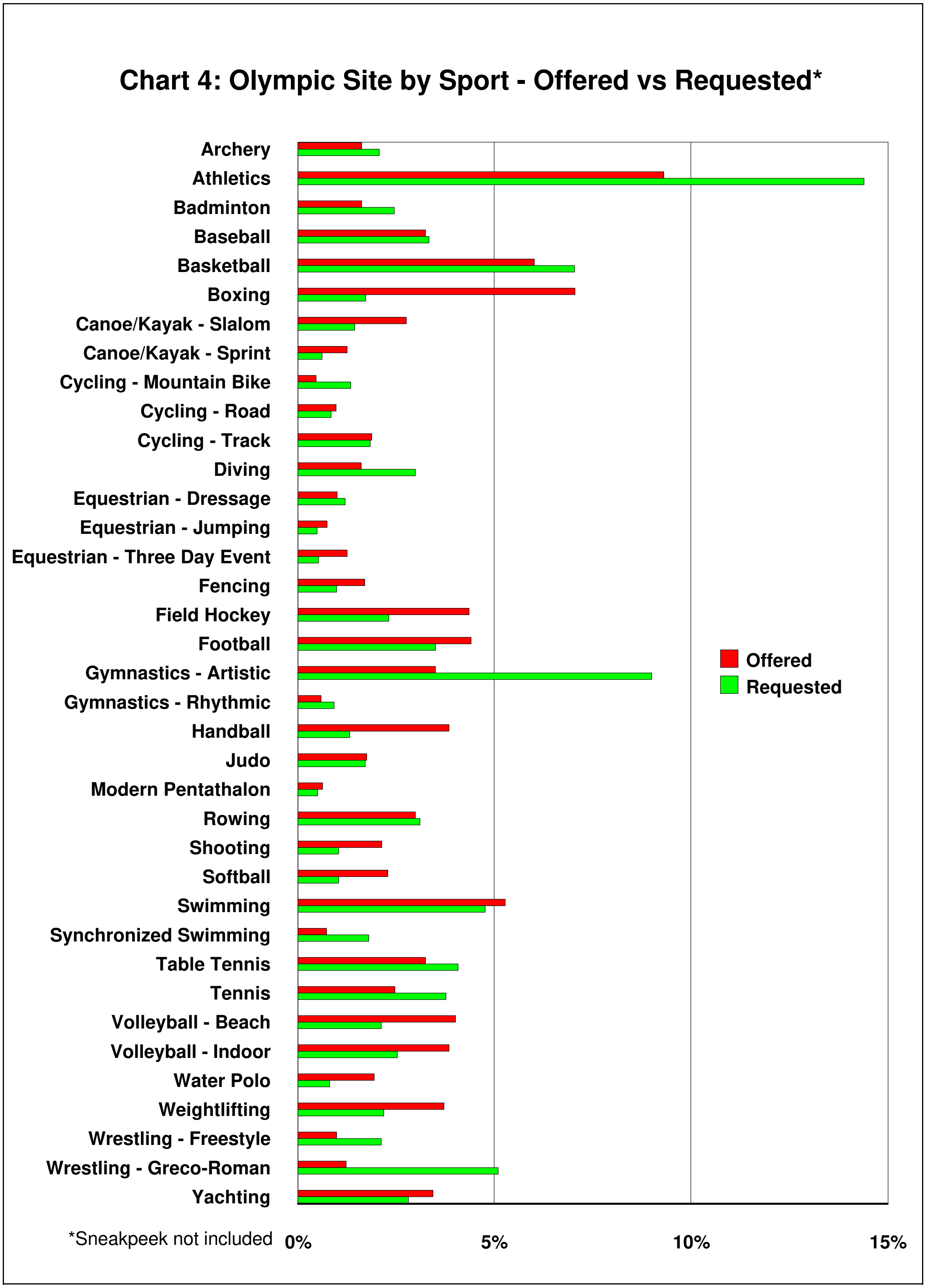


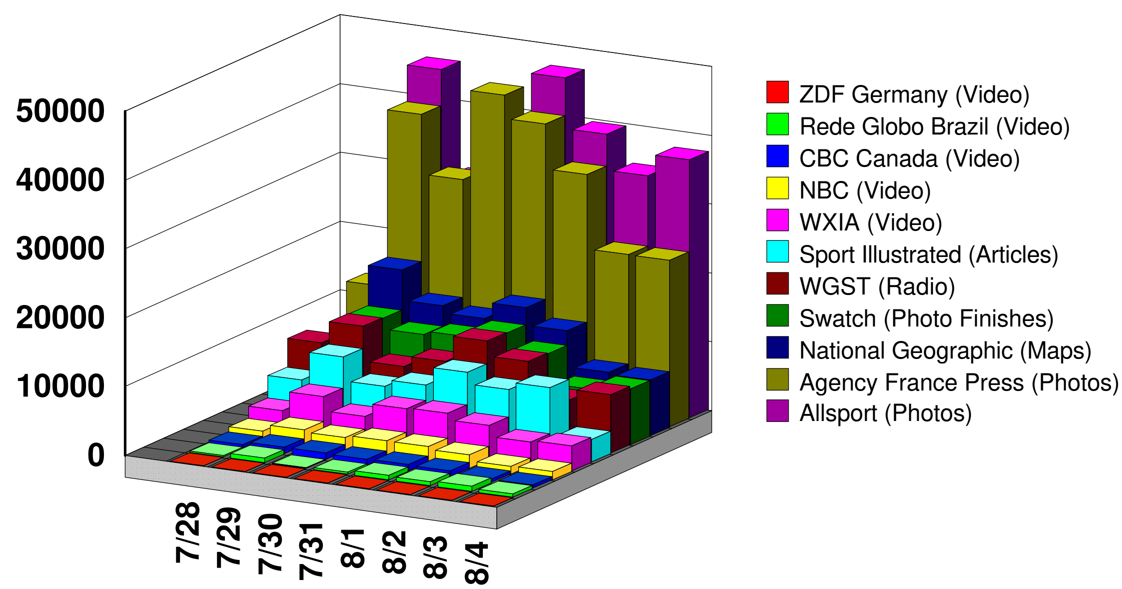
Chart 4: Olympic Site by Sport - Offered vs Requested*
Archery
Athletics
Badminton
Baseball
Basketball
Boxing
Canoe/Kayak - Slalom
Canoe/Kayak - Sprint
Cycling - Mountain Bike
Cycling - Road
Cycling - Track
Diving
Equestrian - Dressage
Equestrian - Jumping
Equestrian - Three Day Event
Fencing
Field Hockey
Football
Gymnastics - Artistic
Gymnastics - Rhythmic
Handball
Judo
Modern Pentathalon
Rowing
Shooting
Softball
Swimming
Synchronized Swimming
Table Tennis
Tennis
Volleyball - Beach
Volleyball - Indoor
Water Polo
Weightlifting
Wrestling - Freestyle
Wrestling - Greco-Roman
Yachting
Offered
Requested
*Sneakpeek not included
0%
5%
10%
15%

3.2. Measuring Visibility for Content Providers

Content providers and sponsors of a web site have great interest in knowing the number of readers who accessed their pages or saw their logo. From a marketing perspective, a tally of hits measures the visibility of a company's content. It may justify the contribution to other sites in the future.

All content providers to the Olympic Site were provided with the daily hit rates to their photo, video, audio or text documents. For example, one hit for Swatch means that a reader either requested a page with a Swatch photo finish, logo and caption, or a photo finish index page, also bearing the Swatch logo.

Chart 5: Hits to Content Provider Pages



Most Popular Items

Another analysis of possible interest to a content provider is comparing the popularity of their documents to one another. At the Olympic Site, all Sports Illustrated articles and Swatch Photo Finishes were offered through index pages without editors highlighting any items over others. Thus, differences in hit rates reflect differences in reader interest. For example, the most popular Sports Illustrated articles in Table 2 indicate that titles suggesting violence or sex seemed to draw the biggest audiences.

Table 2: Most Popular Sports Illustrated Articles

Day	Sports Illustrated Article	Hits
7/28	A Bloody War that Spilled into the Pool	3,395
7/30	A Soviet Strongman's Finest Hour	1,575
7/31	A Soviet Strongman's Finest Hour	764
8/1	Grace Under Pressure	833
8/2	Play Girl	826
8/3	Play Girl	2,565
8/4	School's Out	553

The same analysis for Swatch in Table 3 shows that, although photo finishes covered four sports, the most popular ones portrayed Athletics, Men's 100m and 200m Dashes.

Table 3: Most Popular Swatch Photo Finishes

Day	Swatch Photo Finish	Hits
7/28	Women's 100m Dash Round 1 Heat 6 on Jul 26	127
7/30	Men's 100m Dash Round 1 Heat 12 on Jul 26	127
7/31	Men's 100M Dash Final on Jul 28	570
8/1	Men's 100M Dash Final on Jul 28	115
8/2	Men's 100M Dash Final on Jul 28	129
8/3	Men's 200M Final on Aug 1	162
8/4	Men's 200M Final on Aug 1	138

Comparison with Competitors' Coverage

In addition to measuring the visibility of their own pages, companies may be interested in a comparison with other providers. A company can learn about the strategy of a competitor from the coverage and the activity on their web pages.

Allsport and Agency France Press provided photos to the Olympic Site. The two companies' photos were offered through similar menus. The complete Allsport coverage was uploaded to the server, while AFP photos were selected by the photo editor. As Section 3.1 shows, both photo providers' pages were equally popular. However, the emphasis of their sport and country coverage was different. Neither photo provider's coverage fully correlated with the medal winning nations, as one might have expected.

Allsport, USA
924 photos covering athletes from 59 countries
Top 12 countries covered:

22.0% USA
12.3% UK
11.1% Australia
4.3% Japan
4.2% Ireland
3.4% France
3.0% Italy
2.6% Canada
2.4% Russian Federation
2.0% Argentina
2.0% Spain
2.0% South Korea

AFP, France
461 photos covering athletes from 69 countries
Top 12 countries covered:

13.0% USA
7.2% France
6.1% Cuba
5.2% Germany
4.1% Italy
3.9% Spain
3.1% South Africa
2.8% Canada
2.8% UK
2.6% Netherlands
2.6% Ukraine
2.6% Brazil

Top 16 Medal Winning Nations
with medals won as percentage of total medals at the Games:

12.0% United States
7.5% Russian Federation
7.7% Germany
5.9% China
4.4% France
4.2% Italy
4.9% Australia
3.1% Cuba
2.7% Ukraine
3.2% Korea
2.0% Poland
2.5% Hungary
2.0% Spain
2.4% Romania
2.3% Netherlands
0.9% Greece

Chart 6 shows the differences in the sport coverage of the two photo providers.

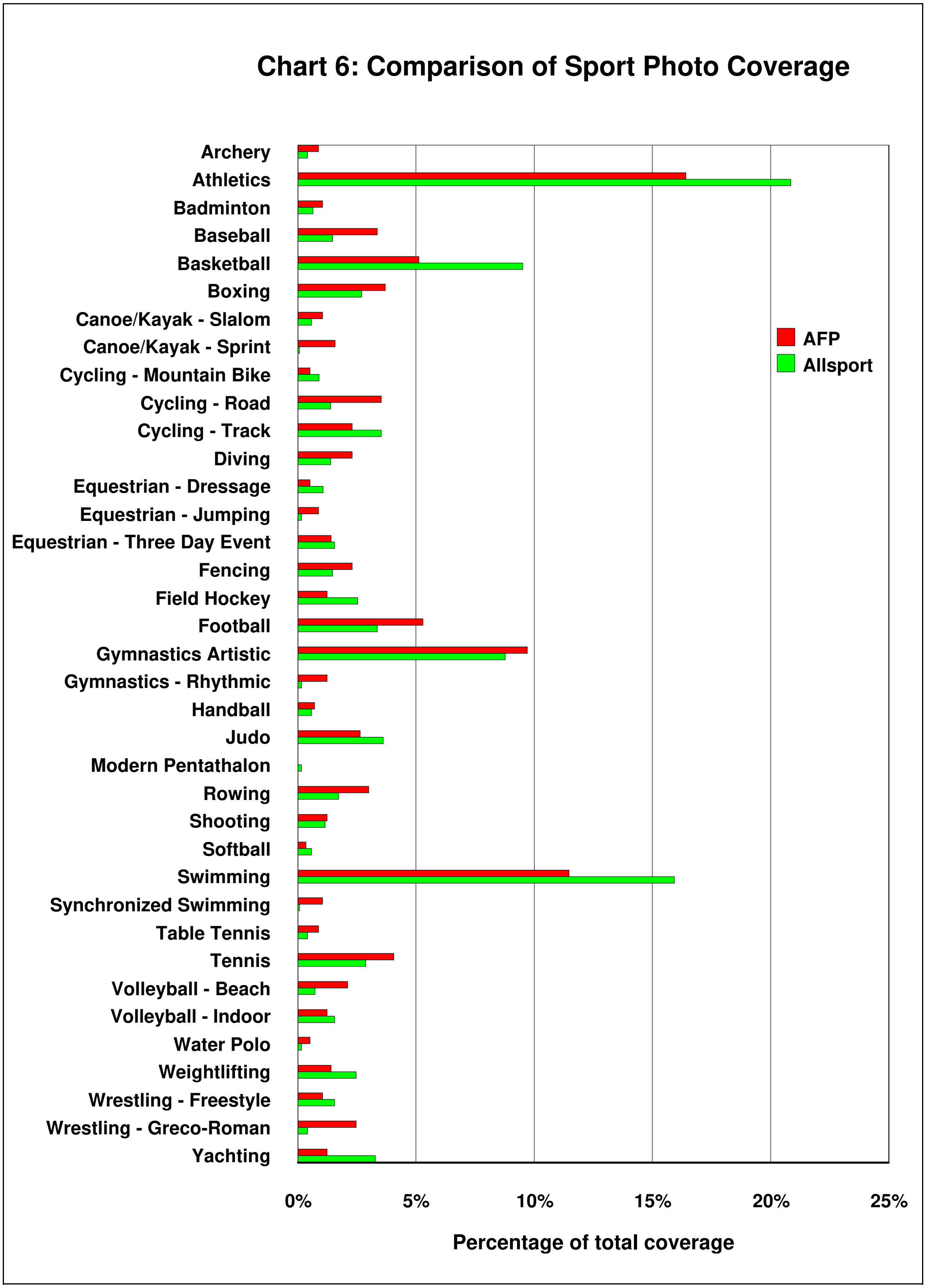


Chart 6: Comparison of Sport Photo Coverage
Archery
Athletics
Badminton
Baseball
Basketball
Boxing
Canoe/Kayak - Slalom
Canoe/Kayak - Sprint
Cycling - Mountain Bike
Cycling - Road
Cycling - Track
Diving
Equestrian - Dressage
Equestrian - Jumping
Equestrian - Three Day Event
Fencing
Field Hockey
Football
Gymnastics Artistic
Gymnastics - Rhythmic
Handball
Judo
Modern Pentathalon
Rowing
Shooting
Softball
Swimming
Synchronized Swimming
Table Tennis
Tennis
Volleyball - Beach
Volleyball - Indoor
Water Polo
Weightlifting
Wrestling - Freestyle
Wrestling - Greco-Roman
Yachting
AFP
Allsport
0%
5%
10%
15%
20%
25%
Percentage of total coverage

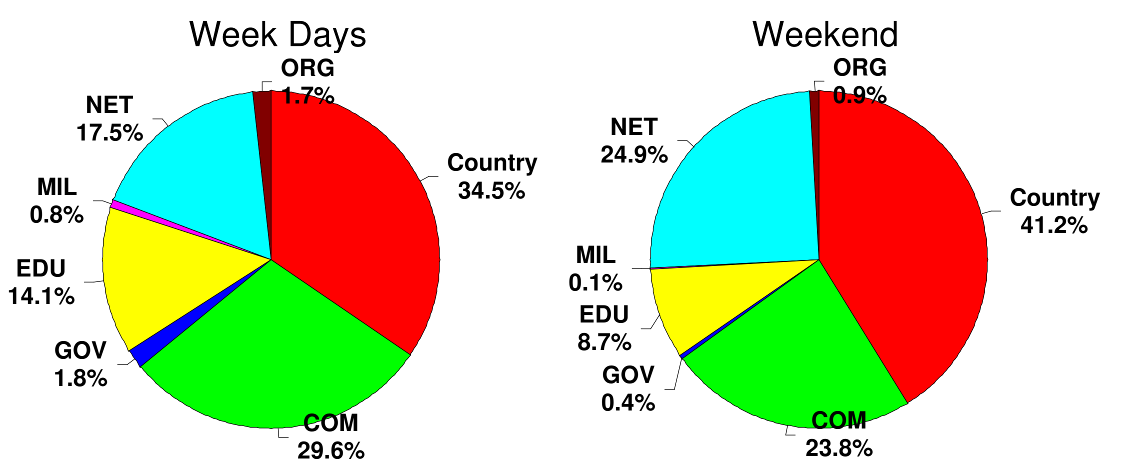
This section has outlined three simple statistics catering to a content provider at a web site. The later sections in this document show that we can easily provide more information, for instance, questions to answers such as: What country did the visitors come from? Did they access the pages during work hours? What else did those visitors see? What kinds of visitors did *not* access the content provider's pages? Knowing the audience is valuable. Internet Sites hosted by IBM could offer such detailed analysis as part of a web service. A content provider to an IBM web site could subscribe to a series of analyses, be it for purposes of targeted advertising or personalized services.

3.2 Client IP Address Analysis

If a site does not require visitors to register, the hosts obtain little demographic information about their audience. Often the only information recorded is the client's numeric IP address. The visiting IP addresses were resolved to a domain name by querying one or more Internet name servers. Some 5% of the IP addresses remained unresolved.

Access providers and companies often issue IP addresses without a country extension. For example, some Canadian companies use the NET extension instead of the CA country extension. Fire walls also hide the individual machines' names. Thus, we cannot make any firm assumption on the true origin of an IP address. Given this uncertainty, we can still observe differences in the distribution of IP domains. Chart 7 shows the distribution for the average of two week days and a weekend. The traffic with university (.EDU) government (.GOV) military (.MIL) commercial (.COM) and non-profit (.ORG) extensions decrease during the weekend, while the country and .NET extensions increase.

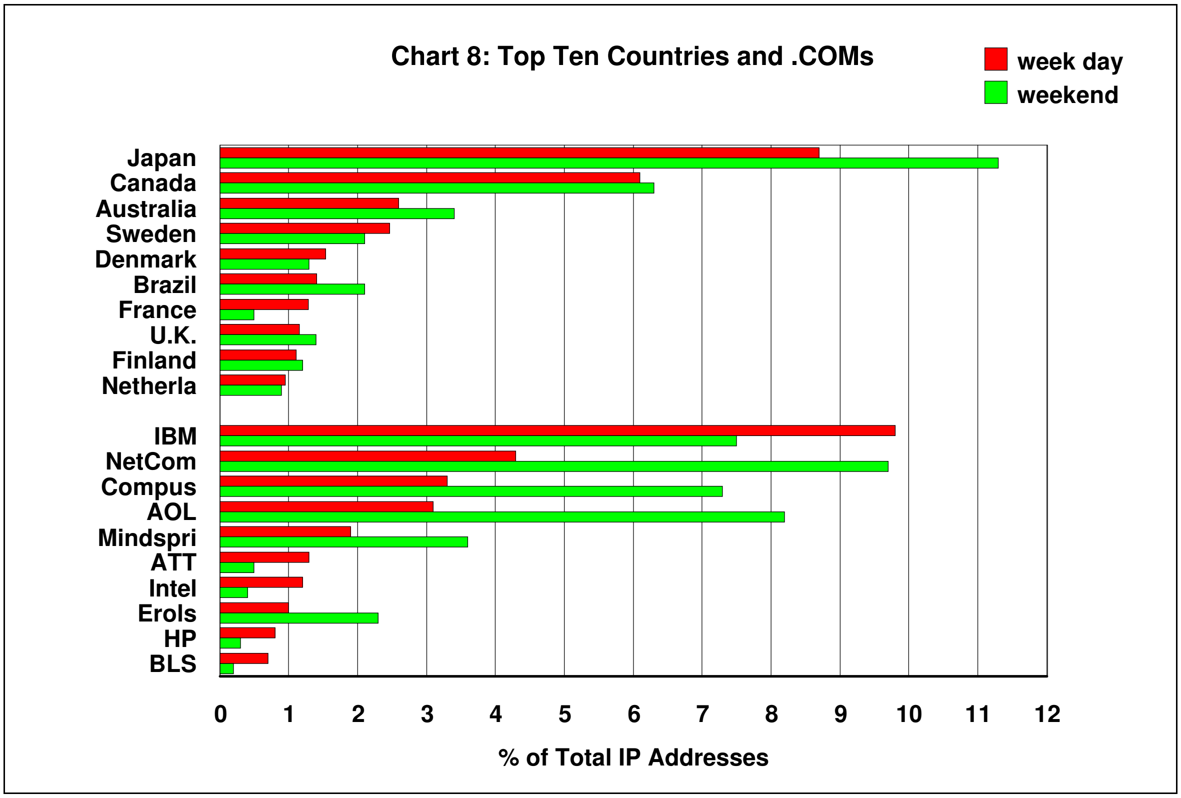
Chart 7: IP Address Distribution



While the site received connections from over 80 countries, including locations such as Nepal, Zimbabwe, and Jordan, the top ten were the most industrialized countries from each continent. Chart 8 shows the most frequent country extensions with the week day and weekend distribution.

Chart 8 also shows the most frequent COM extensions. Addresses ending in IBM.COM dominated during the week with almost 10% of the total traffic at the site. Much of this traffic is due to extensive automated testing performed by the developers of the Olympic site. During the weekend, the percentage of the traffic from the on-line services AOL, Compuserve and Netcom increased significantly.

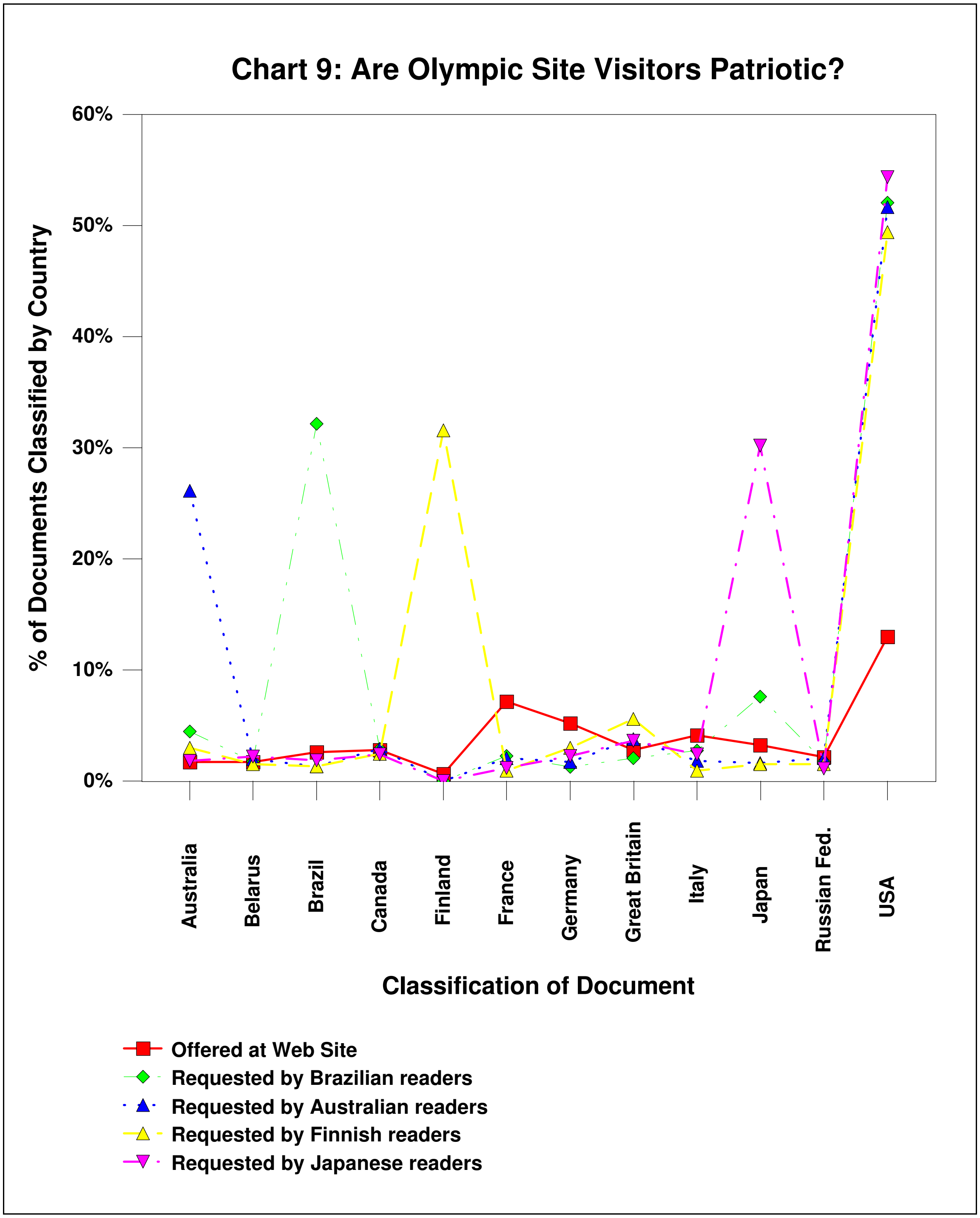
Note, that the overall hit rates were *considerably lower during the weekend*, as illustrated in Chart 2. While the percentage of requests from countries outside of the United States was higher during the weekend, the absolute hits were fewer. Thus, the increase in the percentages in Chart 8 should not be interpreted as an increase in hits.



While access providers may offer connections to clients outside the country borders, most requests with a country extension probably originate from the country. By grouping requests by country extension, we can observe geographic differences in activity. All photos and country information at the site were classified by the country to which they relate. Chart 9 contrasts the distribution of the offered documents by country with the distribution of requests for the documents originating in Japan, Australia, Finland and Brazil. A clear trend emerges: all four countries exhibit peak requests for documents relating to their own country. The mere five documents on Finland at the site attracted without exception all the Finnish visitors. The patriotic behavior of visitors at the Olympic Site justifies offering Country Pages that present all media coverage and results for each country.

13% of the documents related to the host country USA. Visitors from the four sample countries requested documents on USA more than 50% of the time. Because of the Americans' winning streak at the Games, American athletes were frequently exhibited in the Photo Album.

Chart 9 presents only the five most requested countries for each of the four sample countries.



The geographic differences in activity are also evident in the requests for documents relating to a sport. Chart 10 shows the sports favored by the visitors from Japan, Australia, Finland, and Brazil.

In this Section 3, the activity at the Olympic Site has been viewed from various angles: by source, by topic of the document, and by domain of the user. All the techniques have employed simple tallies of hits. However, they have brought out aspects not readily obvious by contrasting several tallies. Often merely juxtaposing two results allows us to see unexpected similarities or differences.

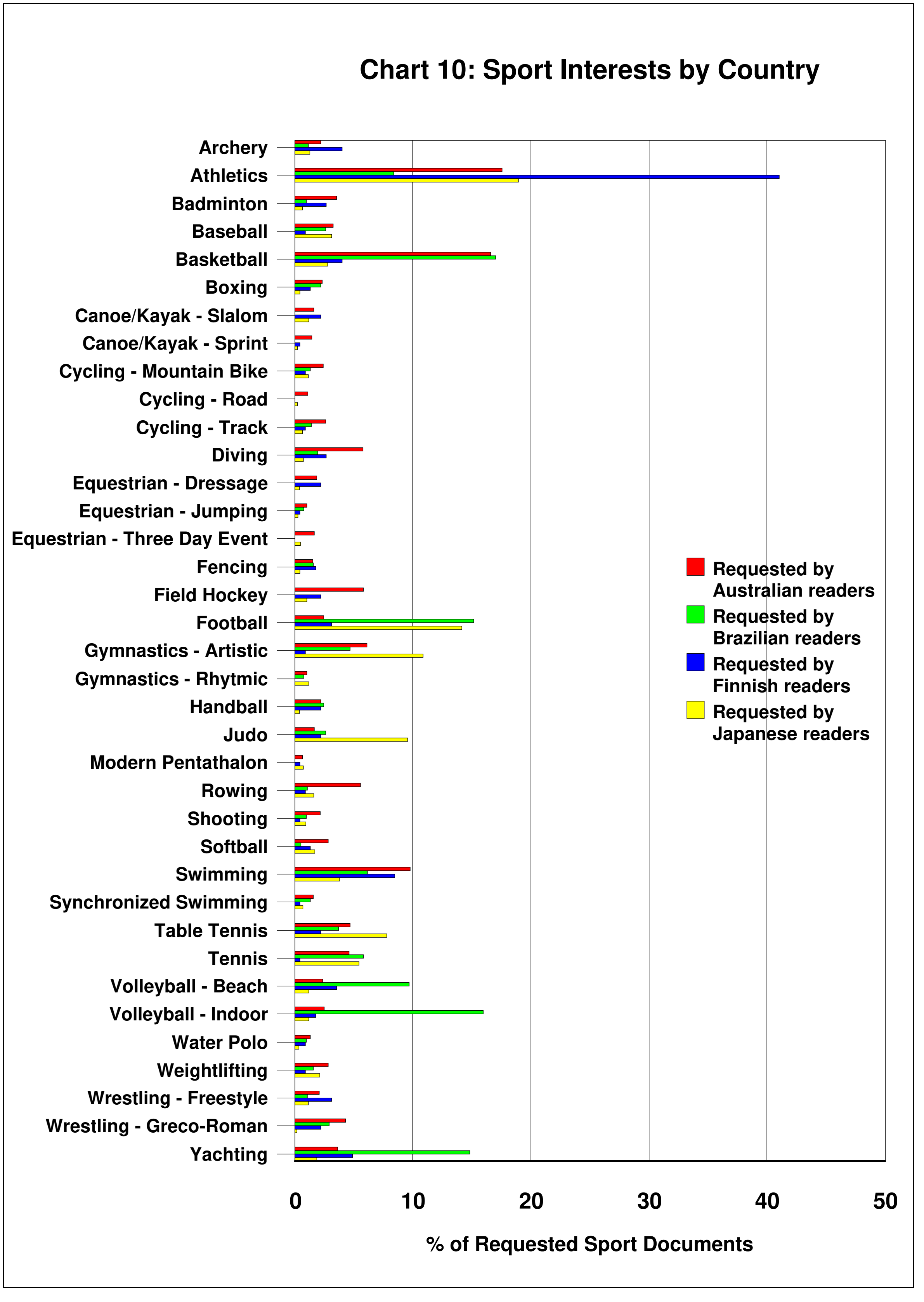


Chart 10: Sport Interests by Country
Archery
Athletics
Badminton
Baseball
Basketball
Boxing
Canoe/Kayak - Slalom
Canoe/Kayak - Sprint
Cycling - Mountain Bike
Cycling - Road
Cycling - Track
Diving
Equestrian - Dressage
Equestrian - Jumping
Equestrian - Three Day Event
Fencing
Field Hockey
Football
Gymnastics - Artistic
Gymnastics - Rhytmic
Handball
Judo
Modern Pentathalon
Rowing
Shooting
Softball
Swimming
Synchronized Swimming
Table Tennis
Tennis
Volleyball - Beach
Volleyball - Indoor
Water Polo
Weightlifting
Wrestling - Freestyle
Wrestling - Greco-Roman
Yachting
Requested by Australian readers
Requested by Brazilian readers
Requested by Finnish readers
Requested by Japanese readers
0
10
20
30
40
50
% of Requested Sport Documents

4. Data Mining Techniques

Given the high number of visitors at the Olympic Games Web Site, on the average 100,000 individual cookies issued per day over a two week period, several questions arise. For instance,

- Can we identify common patterns of interest?
- Can visitors be grouped by their behavior at the site?
- What are the most common paths through the Web Site?

To answer these questions, we applied existing data mining techniques to the Olympic data. We used the IBM Intelligent Miner for the AIX platform. [IBM96] The Intelligent Miner provides six algorithms which can be used alone or in combination. The class of problems to solve and the questions we wanted to answer lead us to apply three of the algorithms, namely the discovery of Associations, Clusters and Sequential Patterns. The following sections give an overview of each technique, present the results and suggest applications to other domains.

4.1. Discovering Associations

This technique discovers affinities among a set of items and generates *production rules*. [AS94] Given a database of transactions where each transaction represents a set of items, the algorithm generates all associations such that the presence of item(s) x in a transaction implies the presence of other item(s) y. The association rule $x => y$ will hold given a *support* and *confidence*, greater than a user-specified minimum support and minimum confidence, where *support* is the number, or fraction, of the transactions that contain a given item set, and *confidence* measures the frequency that items in a multi-item set are found together.

The classical association rule discovered in market basket analysis might be:

> *If a customer buys cheese then the customer also buys bread in 40% of the cases.*
> *This pattern is present in 14% of the transactions.*

In this rule, 40% represents the confidence of the rule and 14% the support of the rule.

Sport Interest Rules

Before running this algorithm, it was necessary to decide what kinds of associations would be meaningful. Since most of the documents are categorized by the main sport they relate to, we decided to look for patterns showing the affinities among sport interests. The purchases in the market basket analysis correspond to requested web pages in the Olympic Web Site analysis. For example,

> *If a visitor accesses a page about Indoor Volleyball then the visitor also accesses a page on Handball in 45% of the cases. This pattern is present in 0.23% of the transactions.*

Input

The association rules algorithm expects as input a file containing two columns, the unique customer identifier and the item, or in the Olympic logs, the cookie and the classification code of the requested page. The file is sorted by cookie to group together all pages accessed by one cookie.

Table 4: Association Rules Input

cookie	sport classification
MHP2739623629	Badminton
MHP2739623629	Handball
MHP2739623629	Archery
DXD2739313620	Badminton
DXD2739313620	Athletics

Results

The following rules resulted from applying associations to a data set containing 57,691 cookies. This represents the activity on any pages relating to a sport during the three-day period, July 28 - July 30, 1996. Requests for Sneakpeek documents were removed because they resulted in already known patterns.

```
12 rules:
Group Support Confidence Antecedent ==>   Consequence
1     0.303%  59.7%      Badminton AND Diving ==> Table Tennis
1     0.303%  45.6%      Table Tennis AND Diving ==> Badminton
1     0.303%  38.5%      Table Tennis AND Badminton ==> Diving

2     0.352%  48.2%      Rowing AND Swimming ==> Gymnastics - Artistic
2     0.335%  46.7%      Basketball AND Swimming ==> Gymnastics -  Artistic
2     0.352%  45.2%      Gymnastics - Artistic AND Rowing ==> Swimming
2     0.335%  29.4%      Gymnastics - Artistic AND Basketball ==> Swimming
2     0.416%  25.2%      Fencing ==> Gymnastics - Artistic
2     0.368%  24.1%      Gymnastics - Rhythmic ==> Gymnastics - Artistic
2     0.352%  23.4%      Gymnastics - Artistic AND Swimming ==> Rowing
2     0.373%  22.6%      Fencing ==> Swimming
2     0.335%  22.2%      Gymnastics - Artistic AND Swimming ==> Basketball
```

The rules separate the visitors into two distinct groups, numbered 1 and 2. The first set of rules shows that the data exhibits affinities among Badminton, Table Tennis and Diving.

> *If a visitor accessed pages on Badminton and Diving, there was a 59.7% chance that the visitor also accessed a page on Table Tennis. 174 visitors, or 0.303% of the 57,691 visitors, confirmed to this rule.*

The confidence of 59.7% means that of five people who accessed pages on Badminton and Diving, three also accessed Table Tennis. The different permutations of the same rule have different measures of confidence. The antecedent of the rule, visitors accessing pages on Table Tennis and pages on Badminton, was a more common combination, thus, the confidence of the complete rule is lower. Because pages on Badminton and those on Diving were accessed less often by the same people, the confidence with which they imply Table Tennis is higher.

Some of the rules in group 2, have confidence values below 30%. For example, one in five people who visited a page on Fencing also went to a page on Swimming. Whether rules are "strong" enough to take into account is somewhat of a subjective decision and depends on the application. By looking at the entire set of produced rules, one can decide on a cutoff point to include a manageable number of the high confidence rules.

While all of the sports included in the above rules are fairly popular, it is also possible to discover rules for the rare, less accessed sports. By lowering the support, we discover that roughly 100 people conform to the rules on equestrian sports. The rules may seem obvious, but the association rules algorithm does not know that the sports are related.

Group	Support	Confidence	Antecedent ==>	Consequence
2	0.220%	27.0%	Equestrian - Jumping ==>	Equestrian - Dressage
2	0.213%	23.8%	Equestrian - Three Day Event ==>	Equestrian - Dressage

Are There “Photo People” and “Text People”?

The separation of readers into groups based on accessed sports lead us to another experiment: did a part of the visitors access only text documents, while others visited multimedia pages? To satisfy our curiosity, we ran the associations algorithm on the previous data set with an alternative classification of the documents, the hierarchic source taxonomy shown in figure 1.

Results

Because Sneakpeek was by far the most popular medium at the Olympic Site, nearly every cookie requested Sneakpeek pages. Thus, the bulk of the production rules generated in the first run were of the form

```
Document ==> Sneakpeek
```

In the second run, we excluded the Sneakpeek documents. However, visitors were still not separated into groups. Only one group of rules emerges, where most of the media types imply visits to text documents.

Group	Support	Confidence	Antecedent ==>	Consequence
2	5.437%	73.8%	Swatch Photo Finish ==>	Text Document
2	5.323%	72.2%	Swatch Photo Finish ==>	Index Page
2	4.326%	71.8%	WGST Radio Report ==>	Text Document
2	4.218%	70.0%	WGST Radio Report ==>	Index Page
2	4.221%	51.1%	Allsport Photo ==>	Text Document
2	3.920%	47.5%	Allsport Photo ==>	Index Page
2	4.473%	47.1%	AFP Photo ==>	Text Document
2	4.163%	43.8%	AFP Photo ==>	Index Page
2	4.960%	41.4%	CountryInfo ==>	Index Page
2	7.567%	30.3%	ACOG News Brief ==>	Index Page
2	4.155%	29.4%	SI Article ==>	Index Page
2	10.042%	28.3%	News ==>	Index Page
2	3.596%	25.5%	SI Article ==>	ACOG News Brief
2	10.042%	21.0%	Index Page ==>	News
2	7.567%	15.1%	Index Page ==>	ACOG News Brief

Employing the hierarchic taxonomy generates rules that contain items from all levels of the hierarchy, not only the “leaves”, or the most specific descriptions in the taxonomy. The results of this experiment seem to indicate that visitors to the Olympic Site browse many sections of the site and view multiple media types.

Web Applications Using Association Rules

Applying association rules allows content providers on the Internet to learn about their audience. As such, it is already a useful analysis. However, integrating data mining results into a web application could improve the service to the customers. The results could also help in the design of new web sites or in modifying existing ones.

Personal Page

Discovery of interest rules could extend a Personal Page application. A Personal Page allows readers to select their favorite topics to follow, as, for example, in the digital newspaper Fishwrap [CMS95]. Upon accessing their Personal Page, readers are presented with the latest stories on these topics. In addition, a Personal Page could offer "Suggested Readings". Association rules would enable a Personal Page application to present documents on topics the reader has *not* visited, but other people with similar browsing patterns have, as presented in [TJGD96].

Such an application should make suggestions when a reader confirms the antecedent of a high-confidence rule. Thus, rules discovered from the actions of the reader community could improve an individual reader's experience. Since rules are based on the classification categories of documents, not individual documents, discovered rules apply to new documents with the same classifications.

Web Commerce

In a web commerce application, association rules play multiple roles. Association rules could group items often purchased together, by placing them at easy access from each other. At the time of a purchase, association rules could also produce a list of further suggested purchases. If a customer confirms the antecedent of a high-confidence association rule, the application could present a sidebar with the items in the consequence of the rule.

In addition, targeted advertising could take advantage of association rules. For example, the sport interest rules discovered earlier in this section could help in the placement of announcements: if a Table Tennis team tries to recruit new members, or a ping-pong ball merchant wants to identify target audiences, advertising on Badminton and Diving pages would be a good start.

4.2. Clustering Olympic Web Site Visitors

A clustering algorithm determines the similarities between a set of items and assigns them to groups of similar items. Neural clustering in the Intelligent Miner employs the Kohonen self-organizing feature map, where a multi-dimensional space is mapped into a two-dimensional space. [Koh89] The self-organizing feature map loosely preserves the topology of the multi-dimensional space, such that similar items are near each other on the two-dimensional map. This paradigm is known as "unsupervised" learning. In this algorithm, the user specifies the number of clusters to produce and the number of passes the algorithm spends fine tuning the result.

Clustering provides the "big picture" of a large audience or customer base. Information about the population can be extracted from the number of members in a cluster and the characteristics the members share. In this experiment, we attempted to cluster the visitors to the Olympic Site into interest groups based on the sports they accessed.

Input

The clustering algorithm expects as input a matrix in which rows represent the items to cluster, or cookies, and columns represent the attributes of interest, or the 37 sports. Thus, a given row contains a zero if the cookie never accessed pages on the sport, or the number of visits to pages on the sport. Values in the sport columns are normalized to fall between 0 and 1. The domain extension of the client IP address is included as a supplemental field. Supplemental fields are useful for interpreting the discovered clusters.

Table 5: Clustering Input

cookie	domain	sport 1	sport 2	sport 3	sport 4	sport 5	sport 6	sport 7	. . .	sport 36
MHP2739623629	JP	0	0	0	0.3	0	0.08	0	...	0
DXD2739313620	EDU	0	0.12	0	0	0	0.01	0	...	0.09

During this experiment, the number of clusters was set to 9 and the number of passes to 30. These parameters were selected because they gave the best results for the clustering.

Results

Due to the sparse nature of the data, the clustering of Olympic visitors by accessed sports did not produce any additional information. Chart 11 illustrates the sparse data -- nearly 80% of the cookies visited only one sport. All of these cookies have an entry in a single column and the other thirty-six columns contain zero. These cookies are therefore very "similar".

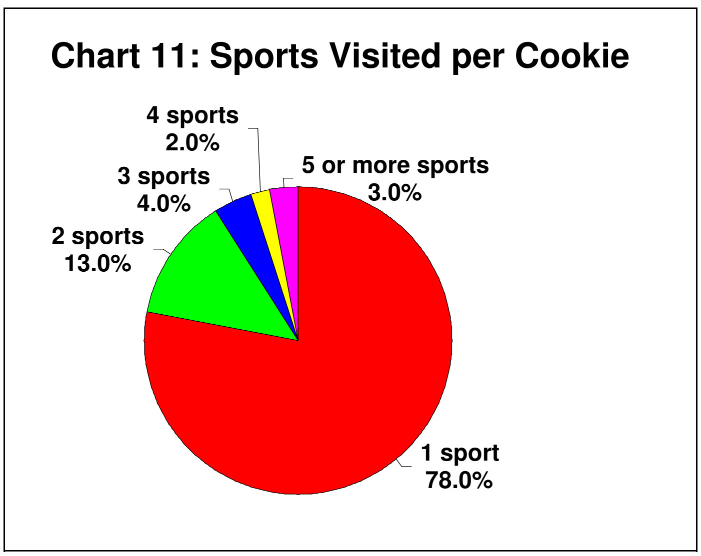


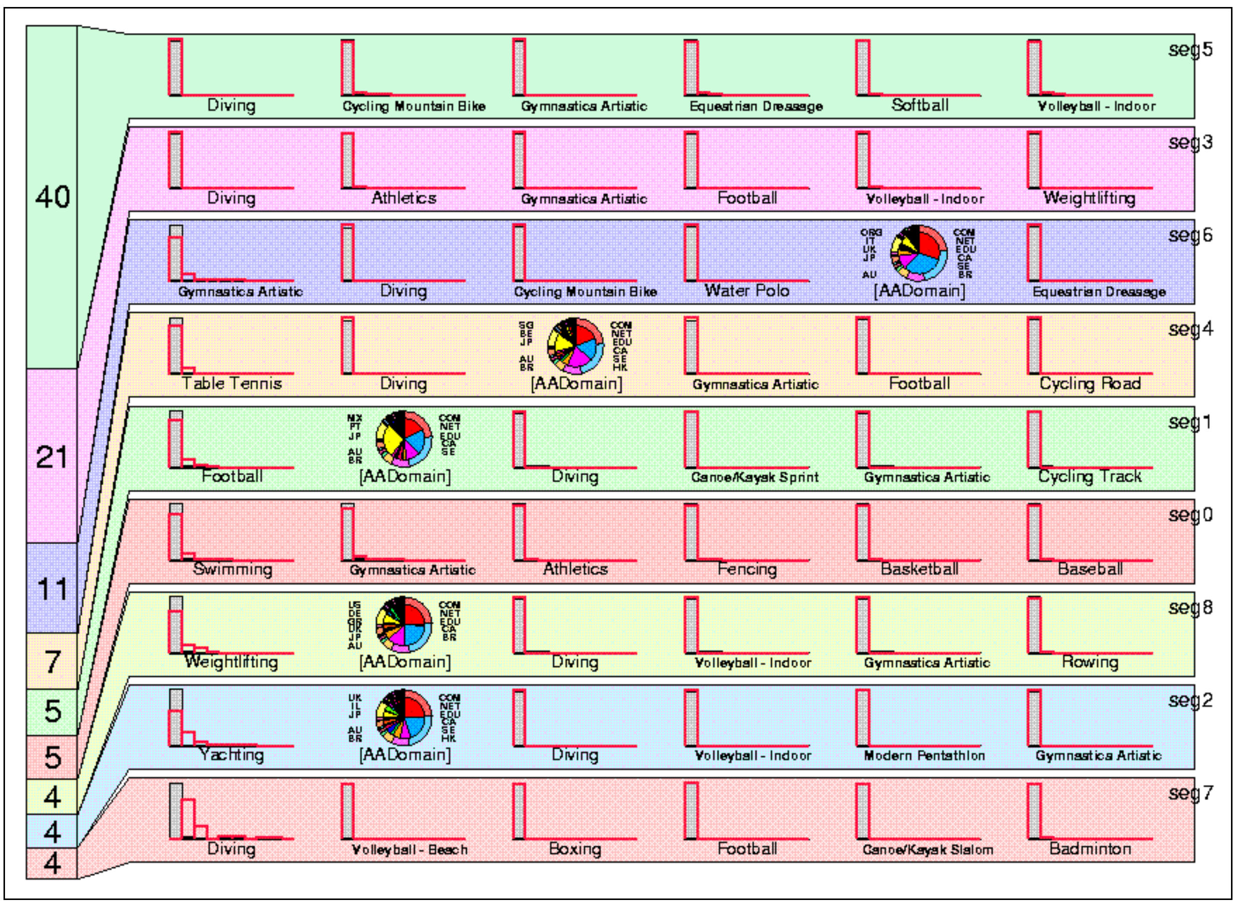
Chart 12 is the graphical output of the clustering algorithm in Intelligent Miner. The leftmost histogram on a given row displays the sport with the highest contribution to the cluster. An interpretation of the clusters uncovers that all clusters, except Seg5 and Seg3, are dominated by a heavy access of a single sport: Artistic Gymnastics, Table Tennis, Football, Swimming, Weightlifting, Yachting, or Diving. Most cookies have a single entry in one of these sport columns. The most popular sport, Athletics, contributes to several clusters, because visitors across many interests visited pages on Athletics.

On the other hand, the dominant feature in the two largest clusters, Seg5 and Seg3, are the sports the visitors did *not* visit. The leftmost histograms for these two clusters represent sports for which visitors had a zero entry.

Comparative analysis

Since the association rules and the clustering algorithm were applied to the same sports data, it is worthwhile to compare the two methods. In clustering, the large number of one-sport cookies dominate. In discovering associations, the one-sport cookies are ignored; the method looks for affinities among sports that occur in the same cookie's transactions. Thus, the associations algorithm focuses only on the portion of the data which exhibits such behavior, regardless of the size or distribution of the data. The associations algorithm also identifies the behavior of rare items, or rare sports in this case. In contrast, clustering is steered by the dominant items, while a rare item can play a significant role if it is the single feature that distinguishes two clusters from one another. Thus, the degree of sparsity of the data may determine which algorithm is more appropriate or provides more meaningful results.

Chart 12: Discovered Clusters of Users



Web Applications Using Clustering

Clustering Users

Clustering is applied in user modeling to construct profiles which represent a group of users. Instead of keeping track of each individual's profile, it may suffice to record the representative profile. A web application could generate customized versions for clusters of users, providing background information relevant to a specific cluster. After constructing initial clusters from a group of users, the application classifies new users to the most appropriate cluster. Occasionally rerunning the clustering algorithm to recreate the clusters assures that the grouping continues to represent a changing community of readers.

Clustering Contents

Clustering the records in a database may help users in composing queries. For example, clustering a large database of text records groups together records with a shared vocabulary. Clustering based on vocabulary can be improved by eliminating common words shared by all records. When a user queries a clustered database, records in the same cluster as the requested document can be offered in addition. If the user is unsure about the exact terms of the query, browsing records similar to the requested documents may help in reiterating a more appropriate query.

4.3. Discovering Sequential Patterns

Discovering sequential patterns means finding patterns where one set of items is followed by another set of items over a period of time. [AS95] Given a database containing sets of transactions ordered over time, applying the sequential patterns algorithm finds the maximal ordered sets of items among existing ones, such that specific item(s) x in a transaction are followed by item(s) y in a subsequent transaction. The sequence $\langle x\ y \rangle$ holds given a user-specified minimum support. The discovered patterns are not necessarily contiguous.

For instance, in the market basket analysis, sequential patterns may indicate the following:

> *Customers who buy a television set buy a VCR within six months.This sequence is supported by 2% of the population.*

Discovering sequential patterns in web server activity means finding common paths through the site. In the case of the Olympic Server, it is interesting to know where visitors chose to go at the main "junctions" in the site.

The key differences between the assumptions of the sequential patterns described in [AS95] and the sequences in web server logs are whether the discovered patterns should be contiguous and the number of items in a transaction. The first assumes that items do not have to follow one another immediately, while the second seeks sequences of consecutive items, or web pages. In addition, in the first algorithm, several items can occur in the same transaction, while the web visitor accesses only one page at a time.

Input

The sequential patterns algorithm expects as input a file containing three columns: a unique customer identifier, a transaction date and time, and item identifiers. In the analysis of the Olympic server logs we used the cookie, the server time stamp and the requested url, as shown in Table 6.

Table 6: Sequential Patterns Input

cookie	time stamp	url code
AHP2739623629	838784680	1005
AHP2739623629	838784802	7112
DXD2739313620	838750818	1005
DXD2739313620	838750847	8805
DXD2739313620	838750925	8809

Results

Examples of sequential patterns obtained by applying the algorithm to the server logs are shown below. Table 7 represents patterns with the highest frequency. For instance, a support of 9.81% indicates that 44,150 cookies conform to this sequence. Table 8 represents sequential patterns with very low support, such as 0.1%, indicating that only 450 cookies conform to the sequence.

Table 7: High Support Sequential Patterns

support	page url	page title
9.81%	/acog/d-index.html /acog/sneakpeek/d-index.html	Altanta Home Page Sneakpeek Main Page
5.77%	/acog/d-index.html /acog/multimedia/d-index.html	Atlanta Home Page Sights & Sounds Main Page
4.74%	/acog/d-index.html /acog/multimedia/d-photoalbum.html	Atlanta Home Page Photo Album Main Page
3.96%	/acog/d-index /acog/news/d-news.html	Atlanta Home Page News Main Page
2.93%	/acog/d-index /acog/schedules/d-index.html	Atlanta Home Page Schedules Main Page
0.63%	/acog/multimedia/afp/d-day-today.html /acog/multimedia/afp/d-day-10-3.html	AFP Photo Index Page AFP Photo Sub-Index Page
0.42%	/acog/sports/d-sports-venues.html /acog/schedules/d-index.html	Sports Main Page Schedules Main Page

Table 8 : Low Support Sequential Patterns

support	page url	page title
0.10%	/acog/d-index.html /acog/multimedia/d-index.html /acog/multimedia/swatch/d-index.html /acog/multimedia/swatch/d-sw072896.html /acog/multimedia/swatch/d-sw-07289602.html	Atlanta Home Page Sights & Sounds Swatch Main Page Swatch Index for 7/28 A Swatch Photo Finish
0.10%	/acog/multimedia/allsport/d-day-04-1.html /acog/multimedia/allsport/d-230796001.html	Allsport Index Page An Allsport Photo

Web Applications Using Sequential Patterns

The output of sequential patterns could benefit several web applications. Knowing where web visitors choose to go from a given page can be useful.

Evaluation of Page Design

Imagine creating alternative versions of the same page and letting visitors' selections indicate the most interesting or inviting options. One could evaluate various alternatives by changing one feature at a time, the color, the font, or the title of a link, and then letting a sufficient audience wander on the pages. Patterns representing the usage of the site provide an evaluation of the page design without conducting explicit evaluations.

Designing a Web Site

Web masters are known to skim server logs to confirm their hypotheses of the most popular paths. An interface to the results of sequential pattern analysis would provide them with a tool to assure that the audience does not have to traverse numerous links to reach the most popular sections. For instance, how does the organization of an index page affect where people opt to go? Intermediary pages may be collapsed as a result of sequential pattern analysis. This technique could be useful in evaluating the organization of a site.

4.4. Data Mining in Internet Applications

When integrating data mining techniques into Internet applications, we believe it should be emphasized to the user that the techniques are imbedded in order to improve the service. Although the content provider's main goal may be well be to sell more items, this should not be the only motivation. Customers are more willing to volunteer information about themselves and their preferences, and to take advantage of the data mining features, if they feel that data mining provides a positive and useful twist to their Internet experience.

Since data mining deduces patterns from a community of customers or readers, the results should be integrated into a web application with caution. The results should be presented to the end-user on the web under the title “Suggested” or “Additional”. Because each visitor may well be the alternative individual and not confirm to the general trend, a presentation should never be locked into a few fixed choices discovered by data mining. The original presentation, same to every visitor, should be accessible at all times. The features created by data mining should serve as sidebars or complementary material.

Conclusion

This report has presented results from mining the 1996 Olympic Web Site access logs. We have highlighted the benefits of classifying documents at a Web site, adding context and meaning to the results obtained. Classification provides a general, high-level description of a document collection, a description which is independent of the individual documents. Attaching multiple classifications to the documents allows viewing the activity at a Web site from many perspectives.

We have presented the results of applying several statistical analyses as well as three data mining techniques, namely association rules, clustering, and sequential patterns. Results from the statistical analyses permit learning basic information, such as the distribution of activity across the topics offered at the site. Similarly, after applying reverse resolution of numeric IP addresses, we have observed differences in visitors' behavior based on their geographic location.

The results presented in the report have been complemented with suggested applications to other domains, such as Web commerce, on-line databases, personalized pages for news services and others. For instance, in a Web commerce application, association rules could identify unrelated items which are often purchased together, helping to place them at easy access from each other.

In conclusion, this paper has demonstrated that applying data mining techniques to Web server logs can lead to important benefits. Depending on the nature of the data, some techniques may reveal more information than others, while together they may unveil valuable characteristics unknown at the beginning of the analysis. We also believe that classifying documents at a Web site using high-level descriptions offers a meaningful interpretation of the activity.

Acknowledgments

We would like to thank the following people. Without their help we could not have conducted these experiments.

- Mike Wright and Sastry Duri at IBM T.J.Watson Center for processing the log data during the Olympic Games and for preprocessing it for our data mining experiments.
- Jeff Milton at IBM Advanced Internet Technology for classifying the Olympic web site documents.
- Josh Knight at IBM T.J. Watson Center for implementing the reverse resolution of the numeric IP addresses.
- Nick Bowen and Dave Grossman for their support in pursuing this project.

Bibliography

[AS94] R. Agrawal and R. Srikant. "Fast Algorithms for Mining Association Rules", *Proceedings of the 20th International Conference on Very Large Data Bases*, pages 487-499, Santiago, Chile, September 1994.

[AS95] R. Agrawal and R. Srikant. "Mining Sequential Patterns", *Proceedings of the International Conference on Data Engineering (IDCE)*, Taipei, Taiwan, March 1995.

[CMS95] P. Chesnais, M. Mucklo, J. Sheena. "The Fishwrap Personalized News System", *Proceedings of the IEEE Second International Workshop on Community Networking*, 1995

[IBM96] IBM Intelligent Miner for AIX, User's Guide, Version 1 Release 1.

[Koh89] T. Kohonen. "Self-Organization and Associative Memory", Springer-Verlag, New York, 1989.

[TJGD96] T. Yan, M. Jacobsen, H. Garcia-Molina, U. Dayal. "From User Access Patterns to Dynamic Hypertext Linking", *Proceedings of the Fifth International World Wide Conference*, Paris, France, May 1996.