



# Determining the "Appropriateness" of Web Site Visits

*Wavecrest Computing's Three-Tiered Approach*

**Wavecrest Computing**  
2006 Vernon Place  
Melbourne, FL 32901  
Toll-free: 877-442-9346  
Voice: 321-953-5351  
Fax: 321-953-5350

**[www.wavecrest.net](http://www.wavecrest.net)**

## Introduction

At Wavecrest Computing, we believe that successful Web-use management can only be achieved through a policy-based approach that can:

- determine the degree to which the work force is using the Web to contribute to the organization's success.
- differentiate "legitimate" from "unacceptable" Web site visits.

In our view, three interrelated processes are needed to simultaneously accomplish these twin objectives. We refer to these as Categorization, Classification and Thresholding. Discussed in some detail below, all three should be reflected in the organization's written Web policy and in the design of the supporting automated monitoring and reporting system. Using these processes, a well-designed policy based system can answer the questions:

1. Which sites were visited, when and by whom?
2. What type of content was sought by the visitors?
3. Were the visits appropriate or abusive?

We should note at this point that, because of the huge size, vast complexity and dynamic nature of the Internet, no system of abuse detection will ever be perfect. The only realistic objective is to obtain sufficient amounts of accurate data to identify trends and trigger deeper investigations of individual situations. The following paragraphs describe how Wavecrest's Web-use monitoring products do this. We'll start our discussion by examining the subject of "categorization".

## Categorization

**General.** In the context of this paper, *categorization* is Wavecrest's approach to: (a) identifying and reporting on the *content* of Web sites visited by monitored users, and (b) grouping those visits in content-labeled *categories*. Categories are major subjects or topics of interest, e.g., sports, finance, pornography, entertainment, shopping, games, etc.

Wavecrest products use a set of categories for reporting and/or filtering purposes. A set can include only standard categories (i.e., generic categories furnished solely by Wavecrest), or it can include both standard and *custom* categories. Custom categories can be created by the organization itself and used to augment the standard categories. (More on custom categories later.) In either case, a single category is comprised of URLs of sites containing similar content.

Whether standard or custom, all categories are incorporated in a built-in "control" or baseline list in the monitoring/reporting system. Then, as discussed below, actual visits are categorized by comparing them during report generation to sites in the list.

**The Wavecrest Categorization (Baseline) List.** As indicated above, Wavecrest products contain and use a built-in baseline "list" of the URLs of thousands upon thousands of heavily trafficked Web sites grouped in 69 standard categories. The list is carefully constructed and revised weekly to include only the most frequently visited sites. We use two primary methods to do this. First, we include in the list the URLs from actual activity (site visit data) as reflected in our major customers' log files. This maximizes the realism, currency and validity of our list. Secondly, we continuously survey reputable journals, magazines and other sources for reliable studies and opinions of what constitutes the most visited sites in the different categories; we then ensure that these sites are incorporated into our list.

**Wavecrest's Primary Categorization Methodology.** During report generation, the URLs of actual site visits (as recorded in log files) are automatically compared to the baseline list and categorized by a matching process. When a match is found, the visited site is 'tagged' with the category of the listed site. This permits large numbers of visits to be segmented (and differentiated) by content categories in various output reports. Without categorization, the

*The best Internet abuse detection systems should be capable of obtaining sufficient amounts of accurate data to identify trends and trigger deeper investigations of individual situations.*

monitoring, reporting and analysis of Internet usage would not be feasible in situations where there is considerable traffic. That is, the number of Web sites is so vast and rapidly-changing, and the amount of potential abuse is so large, that, where traffic volume is high, analysis of activity by individual URL is physically and humanly impossible.

**Note:** One of the Wavecrest categories is designated “Other.” The URLs of visited sites that cannot be identified by the product are placed in this category for reporting purposes. If sufficiently popular, such sites can be included later in the baseline (control) list.

**Wavecrest’s Secondary Categorization Technique.** Before an “unidentifiable” visit is relegated to the “Other” category during report generation, Cyfin subjects it to a second automated categorization procedure. The procedure compares the URL name of the visited site with a carefully researched list of keywords (e.g., “bank”, “casino”, etc.) and URL extensions (.gov, .edu, etc.). In many cases, the result of this “fail-safe” approach is an accurate categorization that can be included in the current report (further reducing the “other” percentage).

**Note:** Tests have shown that use of our categorization approach (control list PLUS fail-safe techniques) will yield an “other” factor of less than 30 percent.

**Custom Categories.** While standard categories are eminently useful, a truly robust Web-use management approach should also include custom categories created by the customer organization itself. Custom categories enable the product user to go well beyond simple abuse identification to more “positive areas” of importance to the organization. Such areas might include sales, volume, product preferences, billing, banking, resource management, purchasing activities, etc. Cyfin’s custom categories provide this capability. They enable organization personnel to easily track very specific categories of core business functions and/or to closely monitor visits to individual Web sites of particular interest – including Intranet sites. Among other advantages, this capability enables organization managers to use Wavecrest reports to correlate Internet activity with other organization data. This in turn enables them to detect business or operational trends and to make strategic as well as tactical decisions in *all* areas, not just personnel administration.

## Classification

**General.** Although categorization – discussed above – is the first and most important process in Web-use management, it is, in most cases, not enough by itself. In an optimized approach, categorization is refined with “classification.”

In this context, classification is the process of rating categories for ‘acceptability.’ The process is conceptually similar to the rating of movies as PG, R, X, etc. Wavecrest products rate and designate each category – and thus the visited URLs in the category – as “acceptable”, “unacceptable”, or “neutral”. Obviously, such classification must reflect the organization’s Web-use policy.

**Wavecrest’s Acceptability Classification Technique.** Wavecrest products contain a “classification” feature that sorts all identifiable visits into acceptable, unacceptable and neutral groupings. To do this they use a process that enables product users to pre-designate *categories* as acceptable, unacceptable or neutral. This can be done for each category except “Other.” This feature is critical to effective, differentiated, and meaningful reporting.

## Thresholding

**General.** At Wavecrest, we refer to the third differentiation process as “thresholding”. *This is the optional process of automatically identifying instances in which the number of visits in a nominally acceptable category become unacceptable – and thus represent “abuse.”* If used, threshold levels are “set” during system installation, again in accordance with organization policy. The term “threshold” is defined as a specified number of visits to a specified category during a 24 hour period of time. Wavecrest products can automatically detect when a threshold has been exceeded and include that information in reports.

*While standard categories are eminently useful, a truly robust differentiation process should also include custom categories.*

**The Thresholding Process.** Wavecrest products enable organizations to easily establish thresholds in one or more categories. This can be done one time, during installation. Changes can be made later if desired. As indicated above, for a given category, a threshold marks the point at which, by policy, provisionally acceptable usage crosses the line into “abuse.” Thresholding enables the organization to permit a reasonable number of personal visits to “acceptable” sites without overloading the system or suffering too much lost productivity.

## Summary

At a minimum, an effective policy-based Web-use management systems needs to be able to answer the questions:

- Which sites were visited?
- When and by whom were they visited?
- Why were they visited? What type of content was sought?
- Were the visits appropriate or abusive?

Using sophisticated but easy-to-interpret category, classification and threshold designations, Wavecrest products answer these questions and more in a clear, efficient and actionable manner.