

# **Expert Data Miner 1.43 User Manual**



Written by Jean-François Beaulieu

Copyright © 2009 by ASCO IT  
All Rights Reserved

Website: <http://www.expertdataminer.com/>



# Index

Index .....	3
Introduction.....	4
Requirements .....	5
1- Expert Data Miner - Getting Started.....	6
2- Fetching the user path from your log file .....	18
3- Website and Page Optimization.....	23
Optimizing your keywords: .....	24
4- Applying Filters to Your Log Files.....	28
5 - Fraud Detection – Pay per Click.....	36
Getting the ROI in a Pay per Click campaign.....	40

# Introduction

[Expert Data Miner](#) is a log analyzer that can parse the two most frequent log files that one can find on an Internet server; IIS log files (W3C extended format) and the NCSA log files that exist on Apache/Linux servers. Together those two formats make almost 95% of the cases that one can find in the industry. EDM will scan quickly the raw data from your log files and reconstruct the user's sessions; with these results you can view which pages were requested the most often and by whom, or get the typical behavior of your visitors by country, landing pages, etc...Because Google Analytics is widely spread today, it is necessary to mention some differences between both systems.

Google Analytics provides you several reports the day after; your log files are upgraded in real time and you can fetch them when you wish during the day.

Google Analytics is a page tagging system that relies on Javascript enabled browsers. The visitors who configured their browsers to disable the Javascript are not recorded. Pages that can't be loaded (when there is an error code) are not recorded. Spiders and robots are not detected. It can be very difficult to obtain statistics for files in which you can't insert the proper tags ( music and video files, pdf files, flash animations, images, etc...). On the other hand a log analyzer like Expert Data Miner can miss some pages if your visitors retrieve them from a cache. This happens more often if the same visitor comes back again and again to ask the same page or if most of your visitors come from the same company. In such a case the request is not recorded in your log file except if you use the tag *nocache* in your pages.

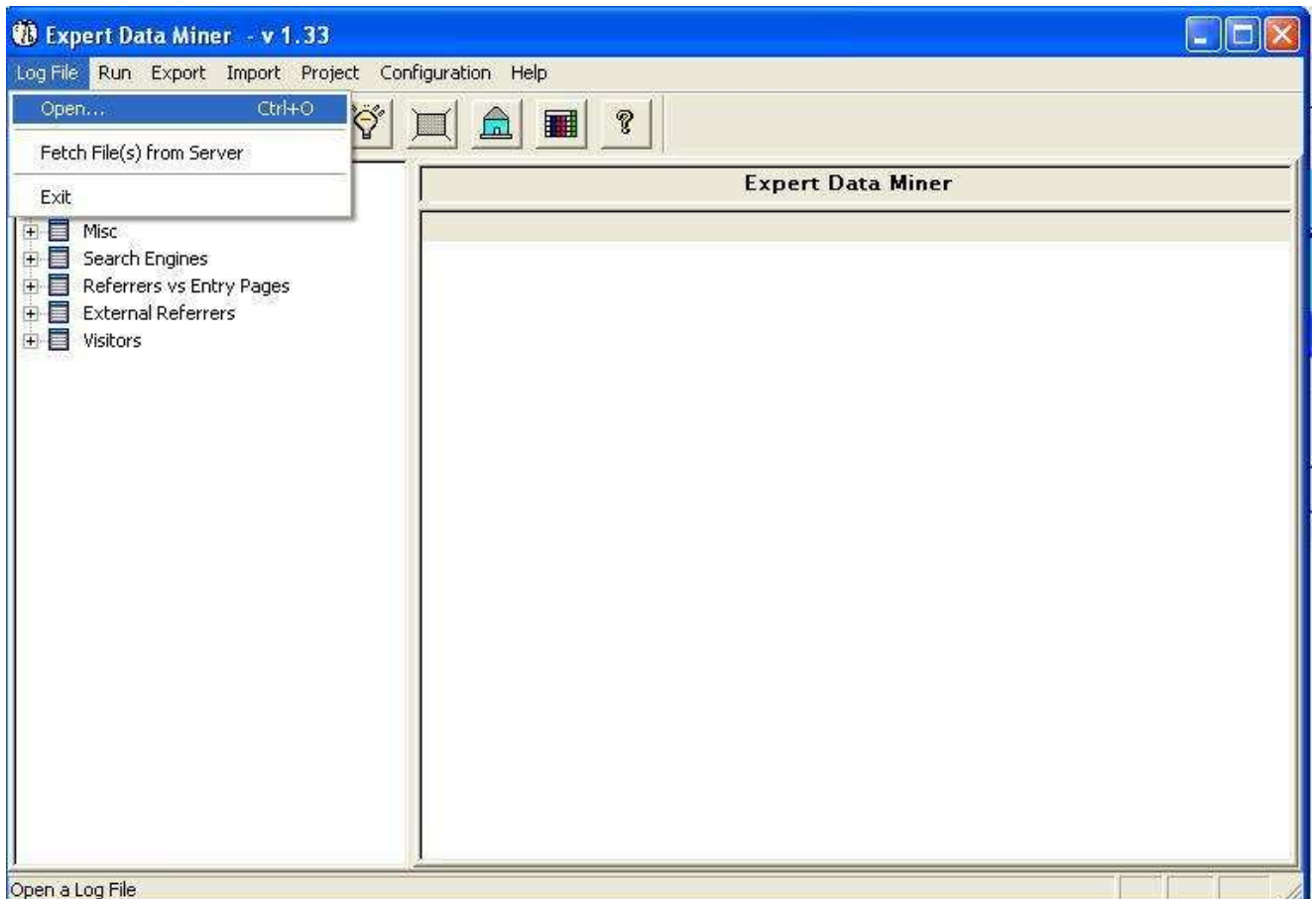
Google Analytics does not allow you to know what happens with each visitor individually; Expert Data Miner does. Google Analytics can provide you relatively accurate statistics regarding the city from which your visitors are coming, EDM cannot for the moment. EDM has a different set of reports that can complete some other reports from GA and help you to detect several cases of [click fraud](#). This is why EDM was designed to support Google Analytics cookies. Finally Google Analytics is free while Expert Data Miner is free only if you have less than 100 visitors per day. However if you have a bigger website you can compare the ratio quality/price with other log analyzers and see by yourself: At 150 or 200\$ EDM is a bargain.

## Requirements

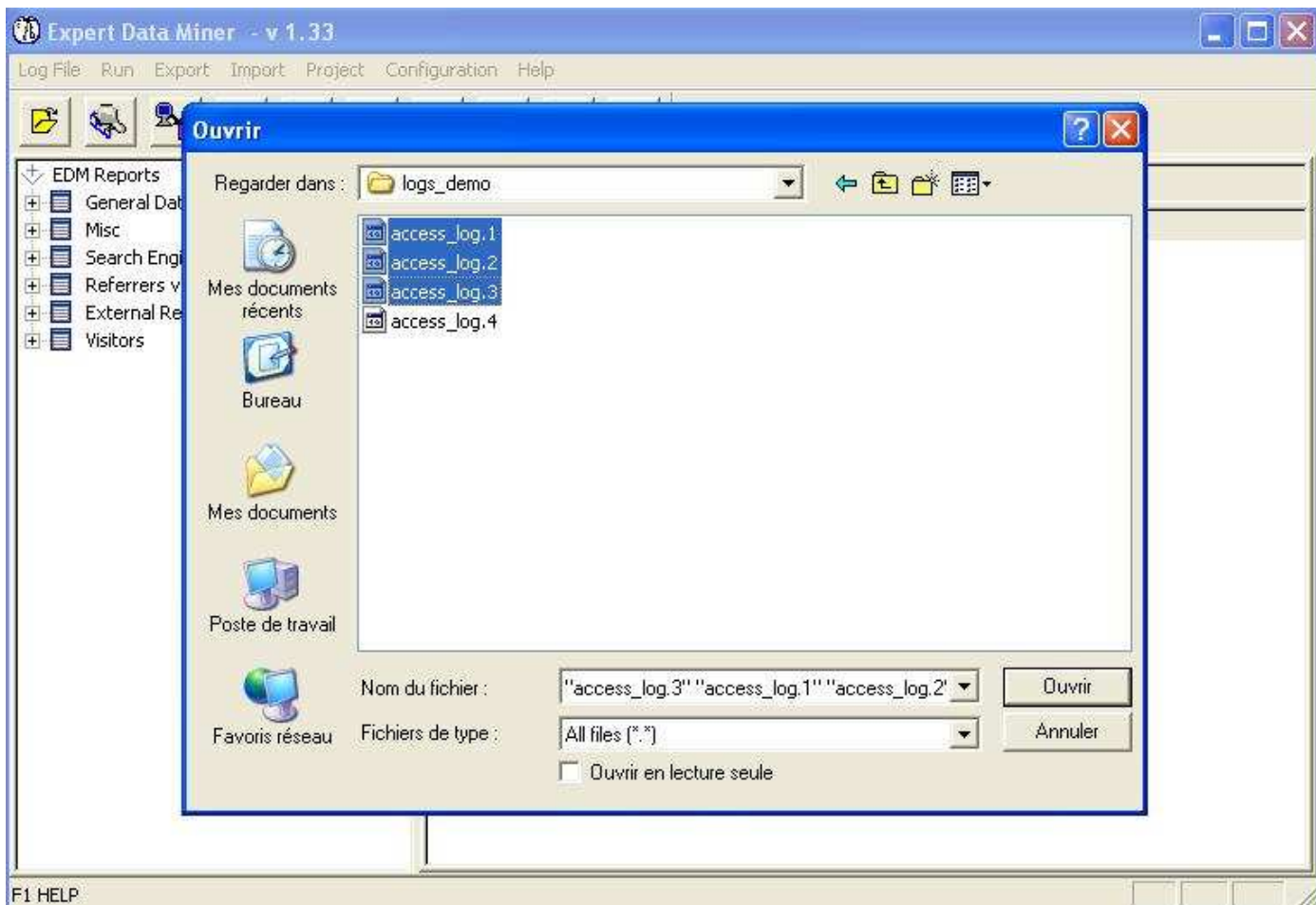
You need to have Windows XP/Vista or Windows 2000/2003. It is assumed that if you run these OS you have enough memory. EDM can run quite well with 512 megs of RAM but the more you have memory the better it is, especially if you wish to scan several hundreds of megs of log files in a non cumulative project. As a thumb rule let say that you need about twice more free memory than the total size of your log files. In a cumulative project (which is using a database) the software is slower but can process several Gigs of log files.

## I- Expert Data Miner - Getting Started

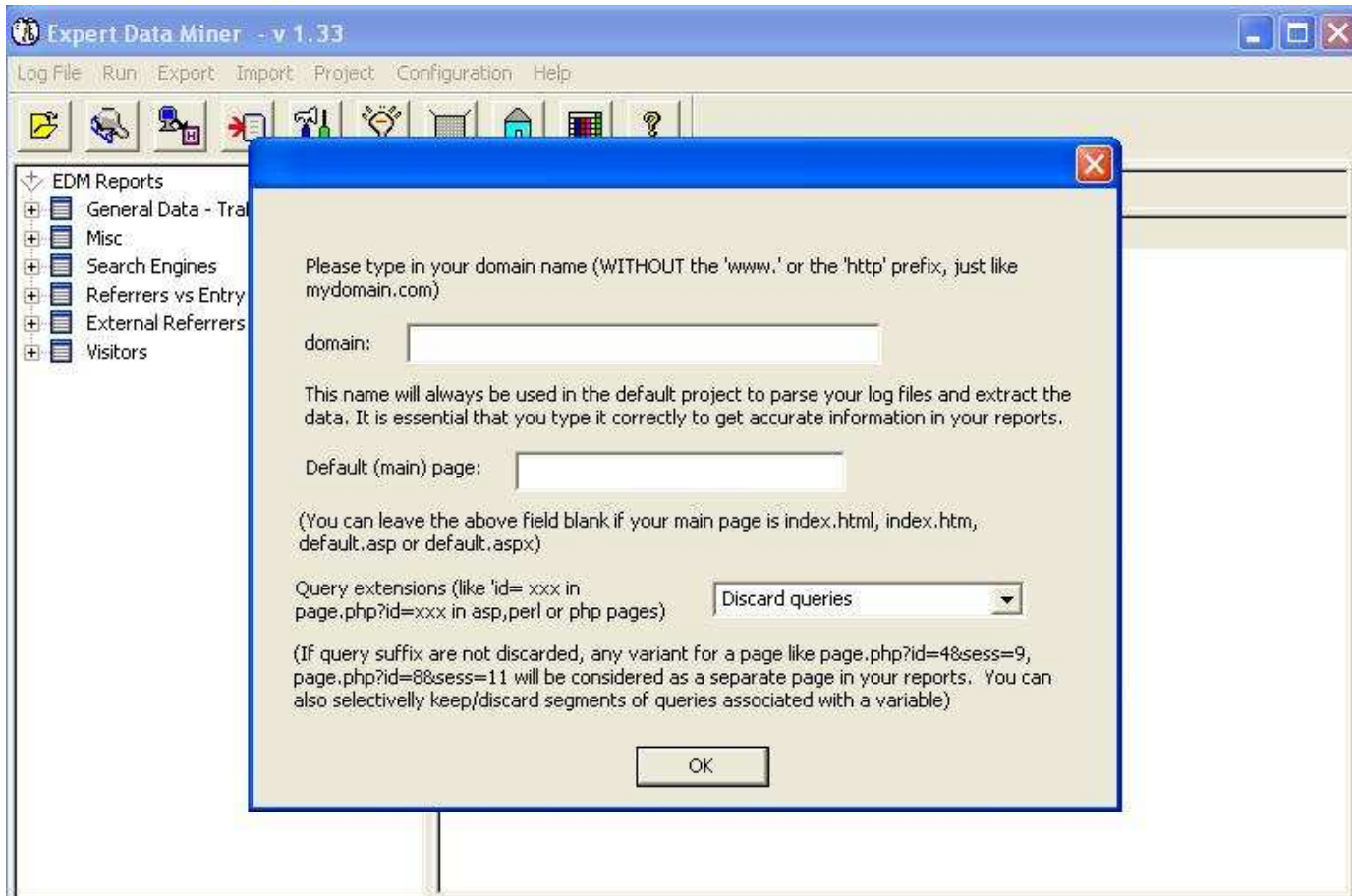
Suppose that you are just using Expert Data Miner for the first time; previously you fetched your log files from your web server. Here we will use Apache log files. The first step is to open your log files from the default project.



In this case, as long that you have enough memory, it is preferable to open several log files at a time. Using the shift key and the mouse both at a time can do the job:



The above dialog appears in French because my Windows XP station is in French, this is a feature controlled by your Windows version, not by EDM. The next step is to parse the log files. But since you are using the system for the first time, you are prompted for the domain name to use. This domain will be used for the default project; if you manage several websites you'll have to create a project for each site rather than to rely on the default project.



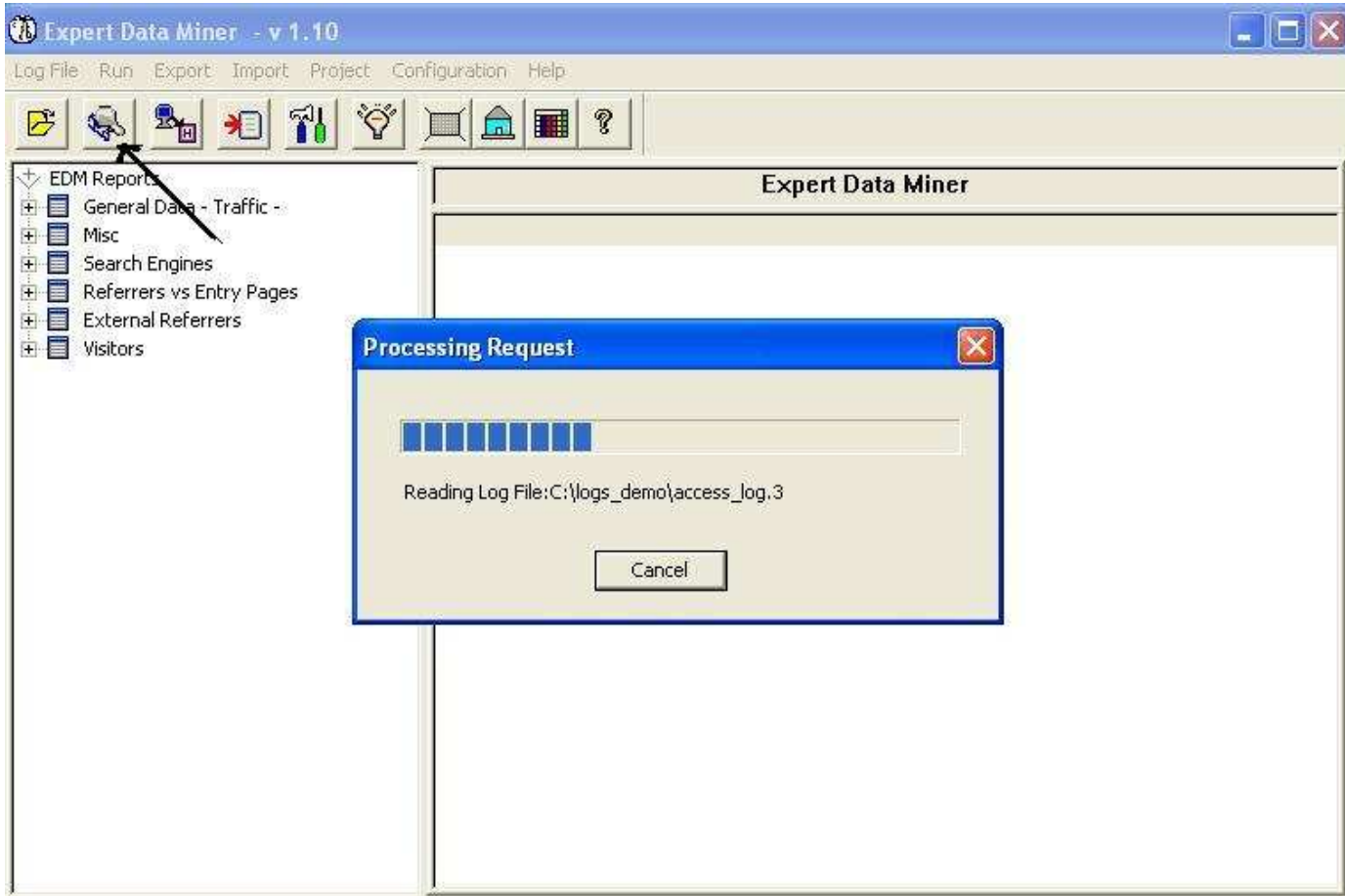
The fictive domain name that is typed here is 'shonxxx.com', **not** 'www.shonxxx.com' or 'http://www.shonxxx.com'. Typing a wrong domain name would affect negatively one report, **External Referrers** but it would also distort heavily the option click path that will be seen later.

**The Default (main page)** can be left to blank in most cases except if your default page is not index.htm, index.html, default.asp or default.aspx (on IIS). When someone visit your domain he can type http://www.yourdomain.com/ or in some cases http://www.yourdomain.com/index.html (or default.asp on IIS) and get the same page. EDM will merge the statistics of all those different requests into an equivalent page "/", your root page, provided that it knows which equivalent pages to use. However equivalences like index.html or default.asp are so frequent that it can manage this by default.

**The query extensions** are the characters that come after the "?" in a request to your pages. Often with php, asp, perl or other kind of pages the output page will be controlled by the query extensions, i.e. http://mypage.asp?id=4 and http://mypage.asp?id=48 will not display the same content. Your log files contain the whole string, but if you wish to regroup the statistics related to http://mypage.asp into a single page you can discard the query extensions. If you have a page like http://mypage.asp?user=A7dhn8562&idPage=4&uid=33&grp=100 you can also decide



to discard selectively some query extensions or to keep some. If any hit on mypage.asp contains a different user number your statistics are useless because each request will get one hit, each will be considered as a request to a different page. Since the variable 'user' plays no logical role in the output content it is better to discard selectively this query extension here. This means that by discarding the variable 'sess' the string user=A7dhn8562 is discarded, the separator '&' also. You can also decide later to discard or keep selectively query variables (and their content) for all your pages or for some specific pages. All you need is to go in the menu *Configuration, Global* and choose the option *Parameters*. Let's parse the log files instead.



The second button under the 'Run' command in the menu is thus pressed. A progress dialog appears. After a few seconds the task is over.

Let's choose the *General Data - Traffic* section. Here the *daily activity* is retrieved:

Expert Data Miner - v 1.10

Log File Run Export Import Project Configuration Help

EDM Reports

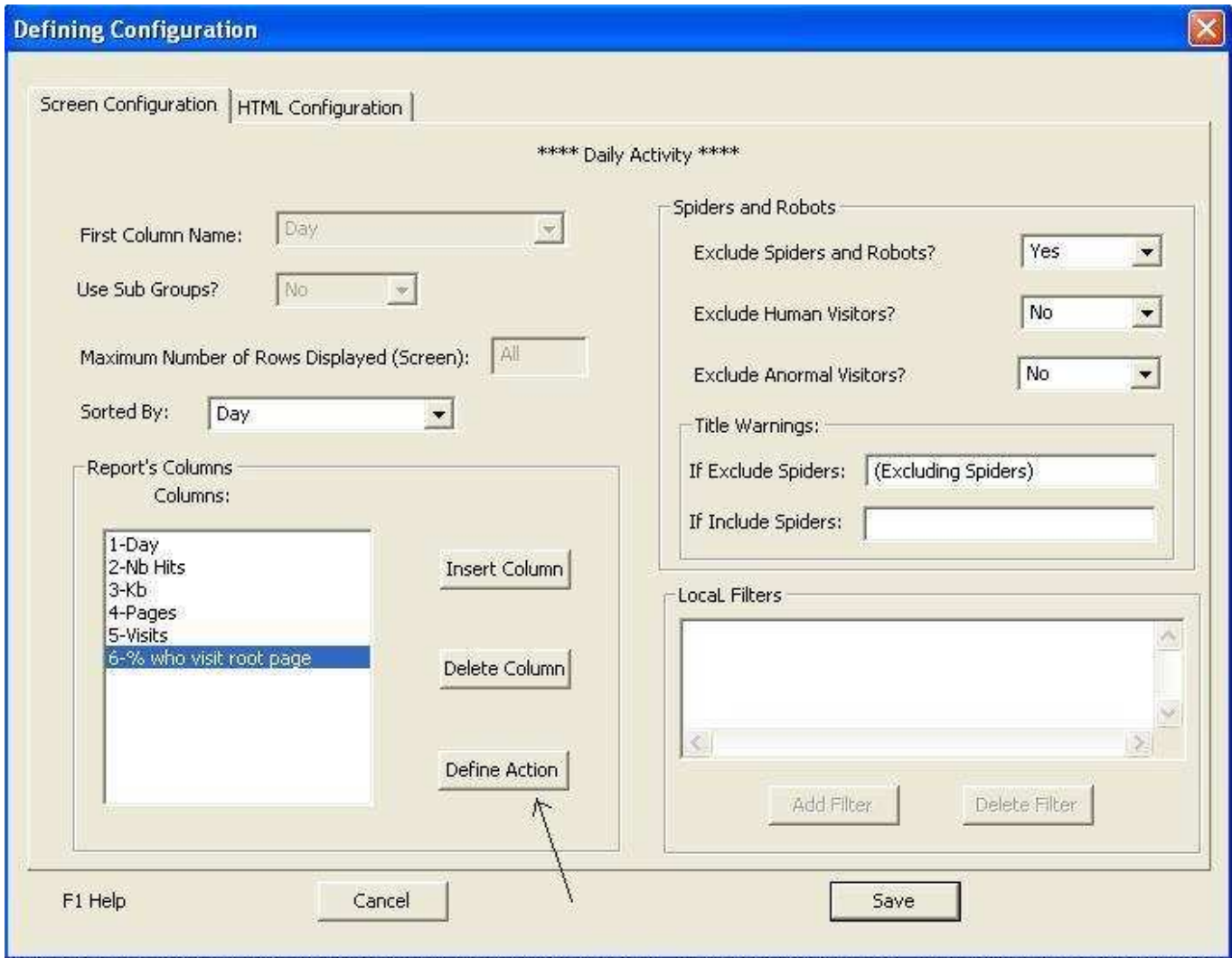
- General Data - Traffic -
  - Summary-Visitors
  - Spiders & Others
  - Daily Activity**
  - Hourly Activity
  - Most Popular Pages
  - Most Popular Downloads
  - Geographical Location
- Misc
- Search Engines
- Referrers vs Entry Pages
- External Referrers
- Visitors

**Daily Activity from 2006/8/18 to 2006/8/25 (Excluding Spiders)**

Day	Nb Hits	Kb	Pages	Visits	% who visit root page
2006/08/18	327	12333.52	55	48	14.29
2006/08/19	47481	8117786.13	11096	2857	9.62
2006/08/20	45170	7409543.90	6538	3179	8.51
2006/08/21	57010	7394233.63	13190	3551	8.29
2006/08/22	61052	8355653.78	8051	4228	9.16
2006/08/23	49966	4420646.52	6873	3907	7.97
2006/08/24	50615	4973543.48	6929	3628	7.31
2006/08/25	70843	6441636.79	9971	3472	7.84
<b>Total</b>	<b>382464</b>	<b>47125377.76</b>	<b>62703</b>	<b>24870</b>	<b>8.37</b>

F1 HELP F8 DISPLAY CHART

The % of people who visited your root page is a user defined column. To see how it is done, I will click on the button with a hammer and the screw driver, just beside the light bulb. The following (configuration) screen appears:



Here you can configure either the layout of your screen or the layout of your HTML report, if you decide to export those results in an HTML report later. There is one such screen available for each report. The combo box *Sorted By* defines the default sort when you open this report or when you output it to an HTML file. You can always change the current sort column when you click on a column header in the main reports.

The column '% who visit root page' can be selected and deleted. One can also add a new column from a pool of predefined columns. But you can also *create* new columns in this pool. This is done by clicking the button **Define Action**. If you click this button, here is what you get:

**Configure Actions**

Edit/Delete Actions

Action Type	Operand	Target	Header Displ.	Tip Info
Match a Page/File	EQUAL	/	% who visit...	This one is provided as an exam
Ask {<=>} than 'X' Pages	LESS THAN	2	Bounce rat...	The percentage of those who g

For the selected Item in the List:

Add New Action/Edit Existing Action

Scope:   Display results as:

Typical Report:  Short Description/Column Header:

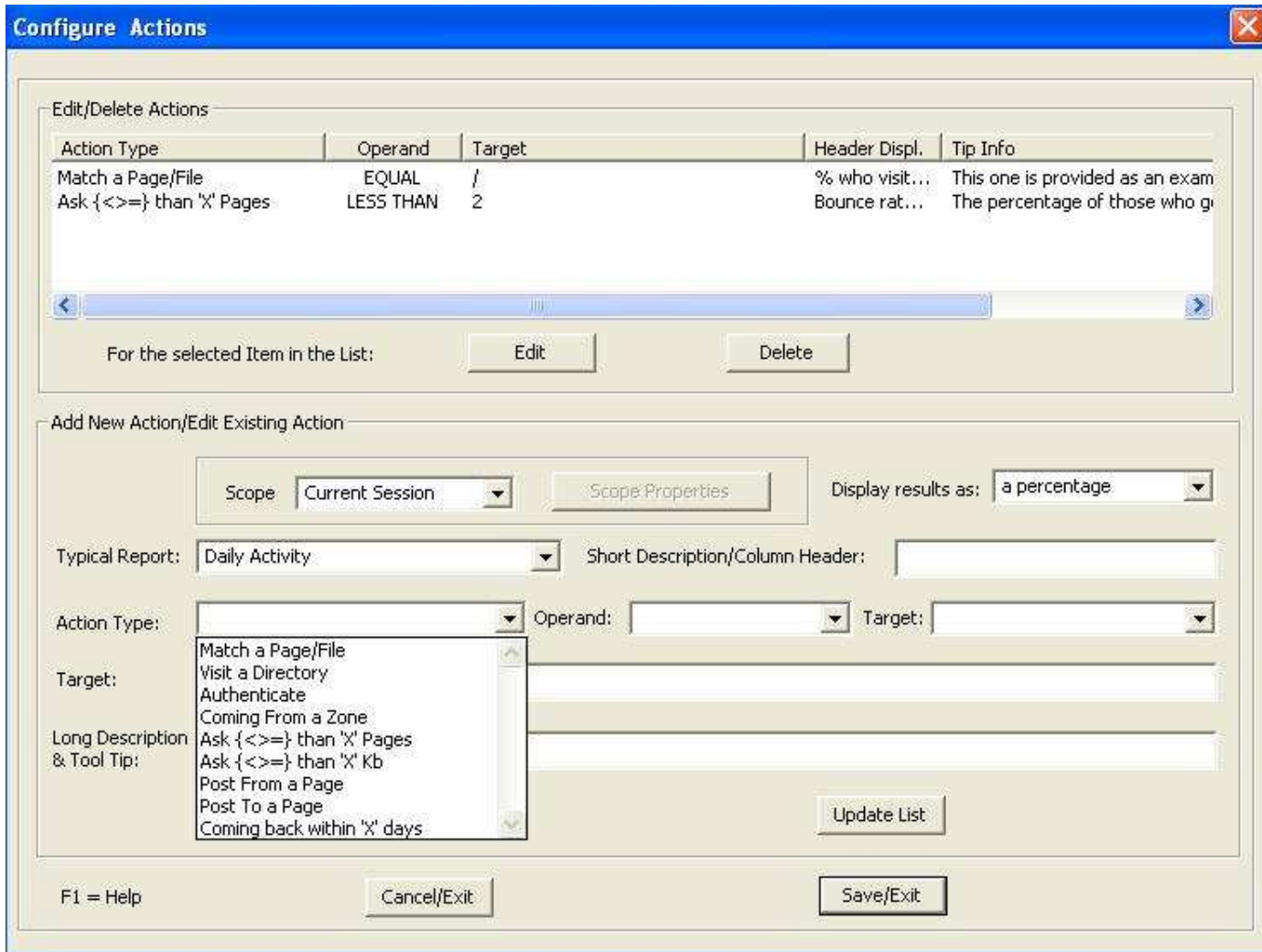
Action Type:  Operand:  Target:

Target:

Long Description & Tool Tip:

F1 = Help

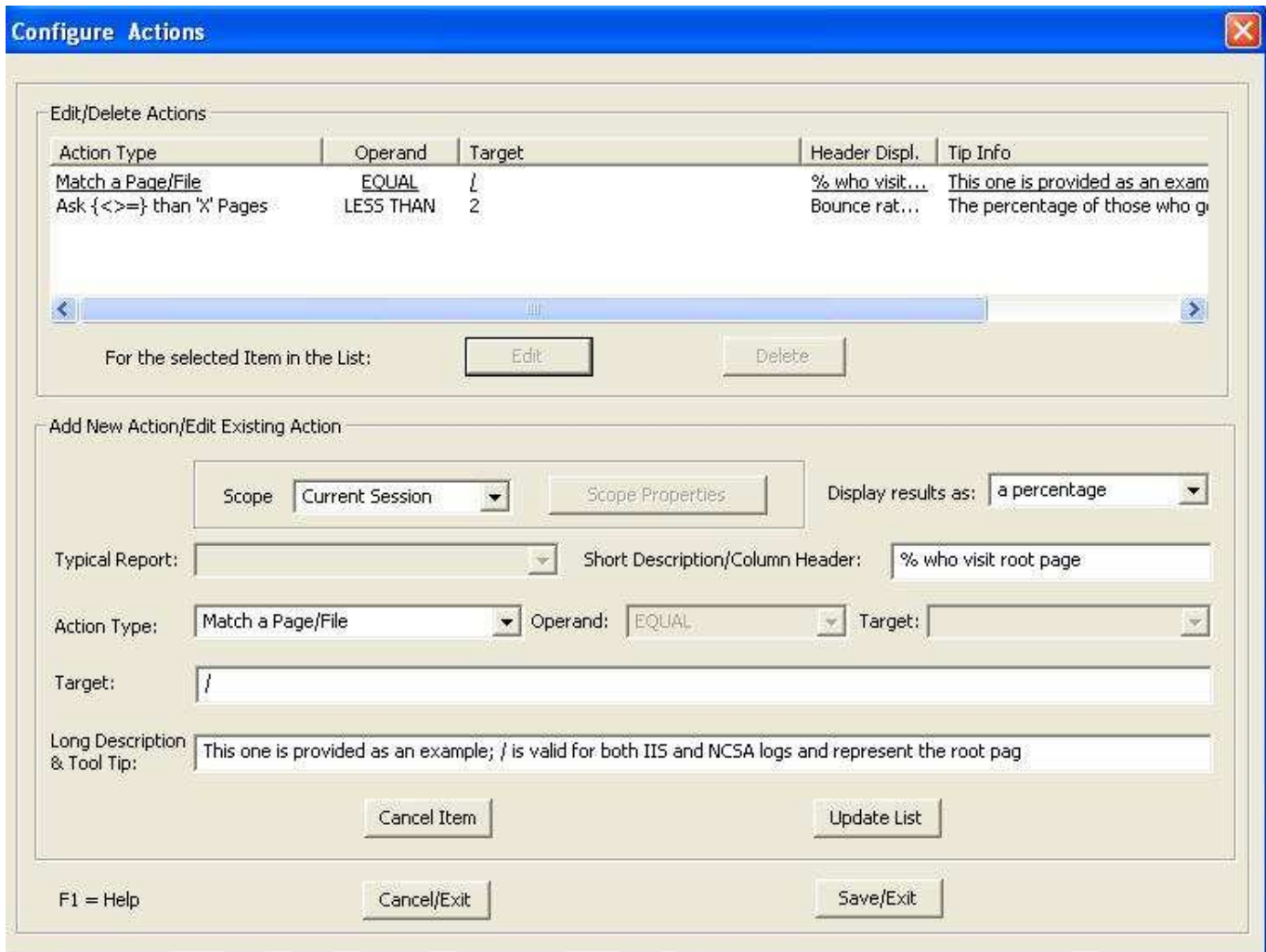
If you want to add a new column in your report, you can select the combo 'Action Type'. The available actions for this report are then shown.



If you want to know what percentage of your users are asking for the page <http://mydomain.com/sub1/mypage.html> during their session, you select 'Match a Page'. The same choice could be done for a downloadable file (zip, mp3, etc..). You then need to type in the target box:

/sub1/mypage.html.

You need also to define the column header and click on the **Update List** button when you have finished before to **Save/Exit**. The Long Description & Tool Tip field is optional; since you are limited to 22 characters for the column header you may prefer to get a longer description when you drag you mouse over the column header in your report later on. But let say that you don't want to add a column right now; just to see the content of a previous action. Select the first line in the list and click the **Edit** button. You will get this:



The target '/' is the last character after your domain's name in <http://www.mydomain.com/>. It is the root page. When EDM scans your log, it will transform URLs like <http://www.mydomain.com/> into <http://www.mydomain.com/> or '/' for the reason that we saw earlier, i.e. because you get the same page when you type one of the above URLs in your browser. For IIS logs, /default.asp and /default.aspx are also transformed into '/' for the same reason.

There is no reason to modify this now so let's click on the *Cancel Item* button. The action that will be created in the pool is *the people who comes from Canada*, so the choice 'Coming from a Zone' will be taken in the combo box **Action Type**.

**Configure Actions**

Edit/Delete Actions

Action Type	Operand	Target	Header Displ.	Tip Info
Match a Page/File	EQUAL	/	% who visit...	This one is provided as an exam
Ask {<=>} than 'X' Pages	LESS THAN	2	Bounce rat...	The percentage of those who g

For the selected Item in the List:

Add New Action/Edit Existing Action

Scope:   Display results as:

Typical Report:  Short Description/Column Header:

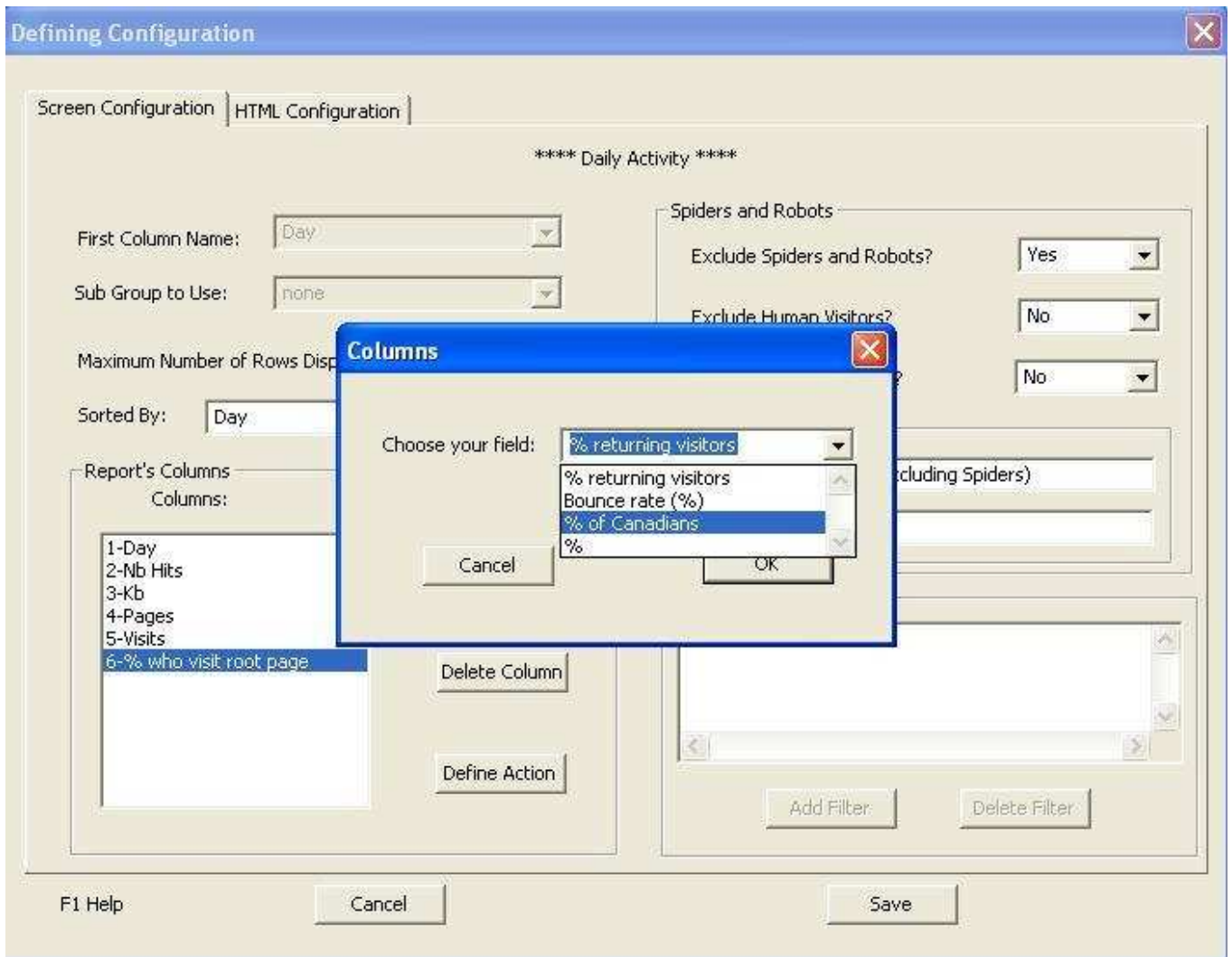
Action Type:  Operand:  Target:

Target:

Long Description & Tool Tip:

F1 = Help

The button **Update List** is then pressed, and finally the button **Save/Exit**. We are back in the previous screen but we need to add this new column somewhere in the report; for the moment it is just in the global pool, but not yet attached to a report. Lets select the column Visits and press the insert button after. The column will thus appear after the column visits.



Once this change is done and saves, click again on the 'Parse Log' button from the main screen and wait that everything is over.



Expert Data Miner - v 1.33

Log File Run Export Import Project Configuration Help

EDM Reports

- General Data - Traffic -
  - Summary-Visitors
  - Spiders & Others
  - Daily Activity**
  - Hourly Activity
  - Most Popular Pages
  - Most Popular Downloads
  - Geographical Location
- Misc
- Search Engines
- Referrers vs Entry Pages
- External Referrers
- Visitors

**Daily Activity from - to - (Excluding Spiders)**

Day	Nb Hits	Kb	Pages	Visits	% of Canadians	% who visit root page
2006/08/18	327	12333.52	55	48	12.24	14.29
2006/08/19	47444	8115535.89	11020	2849	3.87	9.61
2006/08/20	45159	7408192.08	6468	3172	4.31	8.59
2006/08/21	56986	7393191.58	13112	3547	3.82	8.32
2006/08/22	61030	8353574.84	7957	4224	3.43	9.14
2006/08/23	49964	4420593.97	6853	3905	3.73	7.98
2006/08/24	50612	4972908.30	6863	3625	2.99	7.40
2006/08/25	70839	6435523.00	9927	3468	3.79	7.88
<b>Total</b>	<b>382361</b>	<b>47111853.19</b>	<b>62255</b>	<b>24838</b>	<b>3.70</b>	<b>8.41</b>

F1 HELP F8 DISPLAY CHART Right click on a line = Details

Actions can be introduced or removed in nearly the 3/4 of the reports. You can use your imagination and spot situations where cross-linking new columns with a row will give you valuable information. If you work in marketing especially, there is a lot of interesting conclusions that you can draw.

This was for the default project; anytime that you open again your application, you will be able to parse with the domain name that you defined earlier. If you wish to parse log files with other domain's names, you can use the option *Project* from your main menu. You can also create a new project if you wish to use filters in a specific situation, or if you have a large number of log files to process and not enough memory with the default project. If you create a cumulative project, you need to either keep the columns that you define at the outset or either re-scan all your log files when the format of your database is changing. One advantage with the cumulative project is that you can use the option *Fetch Files From Server* and leave it to EDM to discard log lines that were already processed earlier. So it's possible to update your statistics quickly when you press this button.

## 2- Fetching the user path from your log file

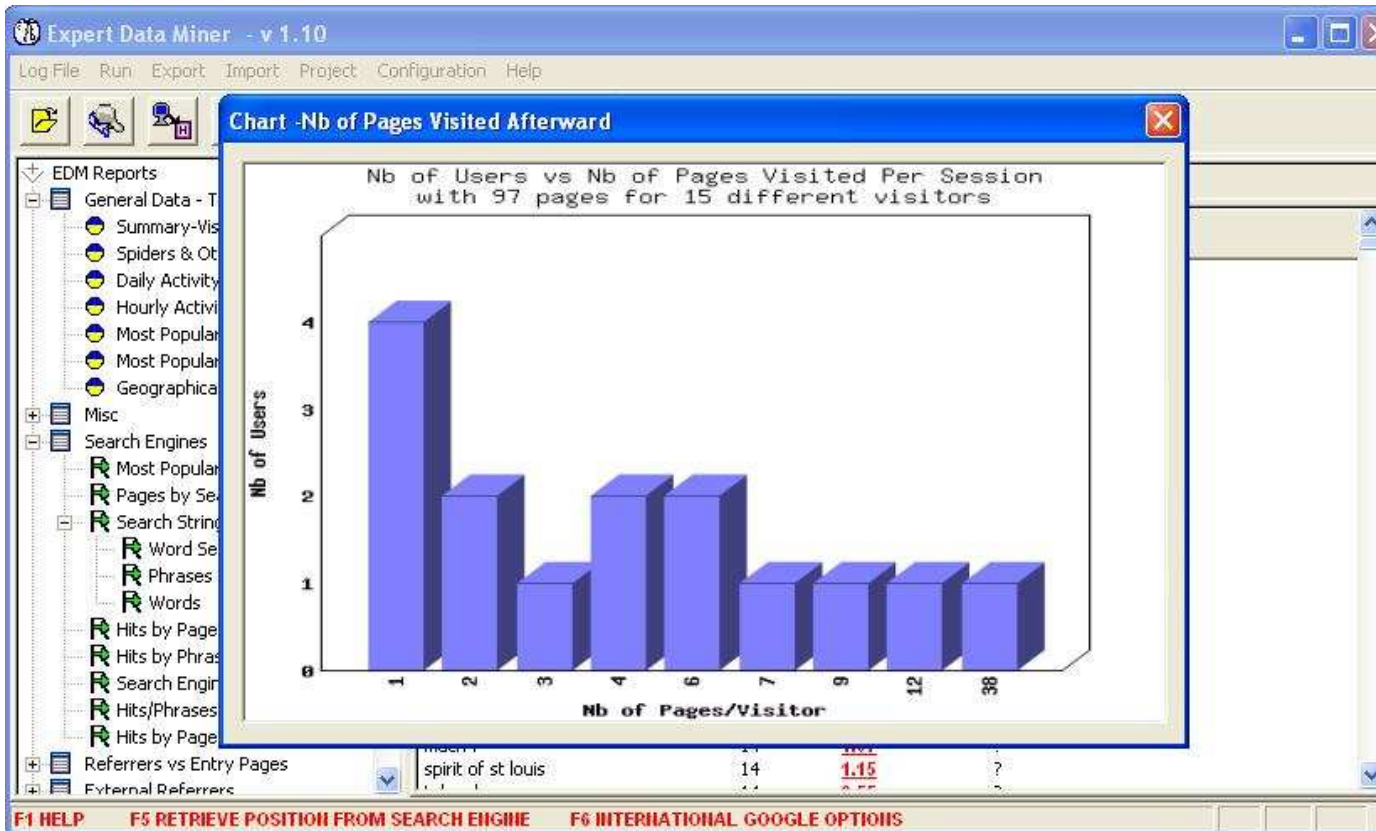
*You can fetch the click trail for any user from 19 reports in Expert Data Miner. The click trail of your visitor will depend heavily of some factors like his landing page, the keywords that he used to find your site, the fact that he is or not a returning user, etc...*

Lets start with the report on search phrases. Behind any search phrase there can be several users who reached your website. The column Avg Pages per Visitor gives you the the average number of pages that your visitors asked during their whole session.

The screenshot shows the Expert Data Miner v 1.10 interface. The main window displays a table titled "Phrases from 2006/8/18 to 2006/8/25". The table has four columns: "Phrase", "Nb Hits", "Avg Pages per Visitor", and "Posit. on google.com". The "Avg Pages per Visitor" column contains several red underlined numbers, with "6.47" highlighted in blue. A context menu is open over the "6.47" value, showing options like "Fetch the Position in Google/Yahoo", "Fetch the click path", "Copy One Row To Clipboard", and "Copy All Rows To Clipboard".

Phrase	Nb Hits	Avg Pages per Visitor	Posit. on google.com
the life of charles lindbergh	83	<u>6.22</u>	?
library	72	<u>8.05</u>	?
aviators	36	<u>1.25</u>	?
ww1 pilots	33	<u>1.14</u>	?
jets	28	<u>9.30</u>	?
helicopters	22	<u>22.00</u>	?
breaking the sound barrier	21	<u>1.00</u>	?
lindbergh 1927	20	<u>1.11</u>	?
b-29	19	<u>1.50</u>	?
boeing	18	<u>1.29</u>	?
hanna reitsch	18	<u>2.27</u>	?
shon.com	18	<u>10.21</u>	?
charles bishop	17	<u>2.71</u>	?
flying	17	<u>3.76</u>	?
forest	16	<u>3.60</u>	?
jet fighters	16	<u>6.47</u>	?
louis		<u>1.36</u>	?
otto		<u>1.07</u>	?
jet pr		<u>1.00</u>	?
avior		<u>1.07</u>	?
mach		<u>1.07</u>	?
spirit of ST LOUIS	14	<u>1.15</u>	?

If you click on one of the underlined numbers (here the 6.47 in red) you get the following distribution.



Here 15 users fetched 16 times the same landing page (one user clicked twice on his Google link) but some users were more interested by your site and asked for more pages. The user for which the click stream will be studied asked for 12 pages during his whole session. One can also fetch the position of this website on Google or Yahoo for the search phrase "jet fighters" or either fetches the click path of the users who reached the site with this phrase. Let's select back the same line, right click and select the **Fetch click path** option (see from picture 1).

Expert Data Miner - v 1.10

Log File Run Export Import Project Configuration Help

Click Path

Select the desired element and double-click to view the user path

IP	Time In	Nb Pages	Country
84.223.139.86	2006/08/22 14:59:09	38	ITALY
85.18.160.134	2006/08/22 06:16:27	12	ITALY
88.36.204.82	2006/08/24 07:52:09	9	ITALY
82.56.65.123	2006/08/25 14:44:19	7	ITALY
201.52.113.124	2006/08/18 23:42:21	6	BRAZIL
88.36.204.82	2006/08/24 02:46:07	6	ITALY
213.140.22.79	2006/08/19 21:17:20	4	ITALY
201.51.62.145	2006/08/24 18:25:19	4	BRAZIL
195.191.195.169	2006/08/19 02:42:33	3	ITALY
213.26.205.138	2006/08/22 04:06:35	2	ITALY
201.50.136.222	2006/08/24 22:33:02	2	BRAZIL
201.29.72.124	2006/08/19 06:31:46	1	BRAZIL
200.206.234.210	2006/08/19 18:21:03	1	BRAZIL
81.118.214.164	2006/08/22 06:14:16	1	ITALY
200.157.34.16	2006/08/23 08:27:41	1	BRAZIL

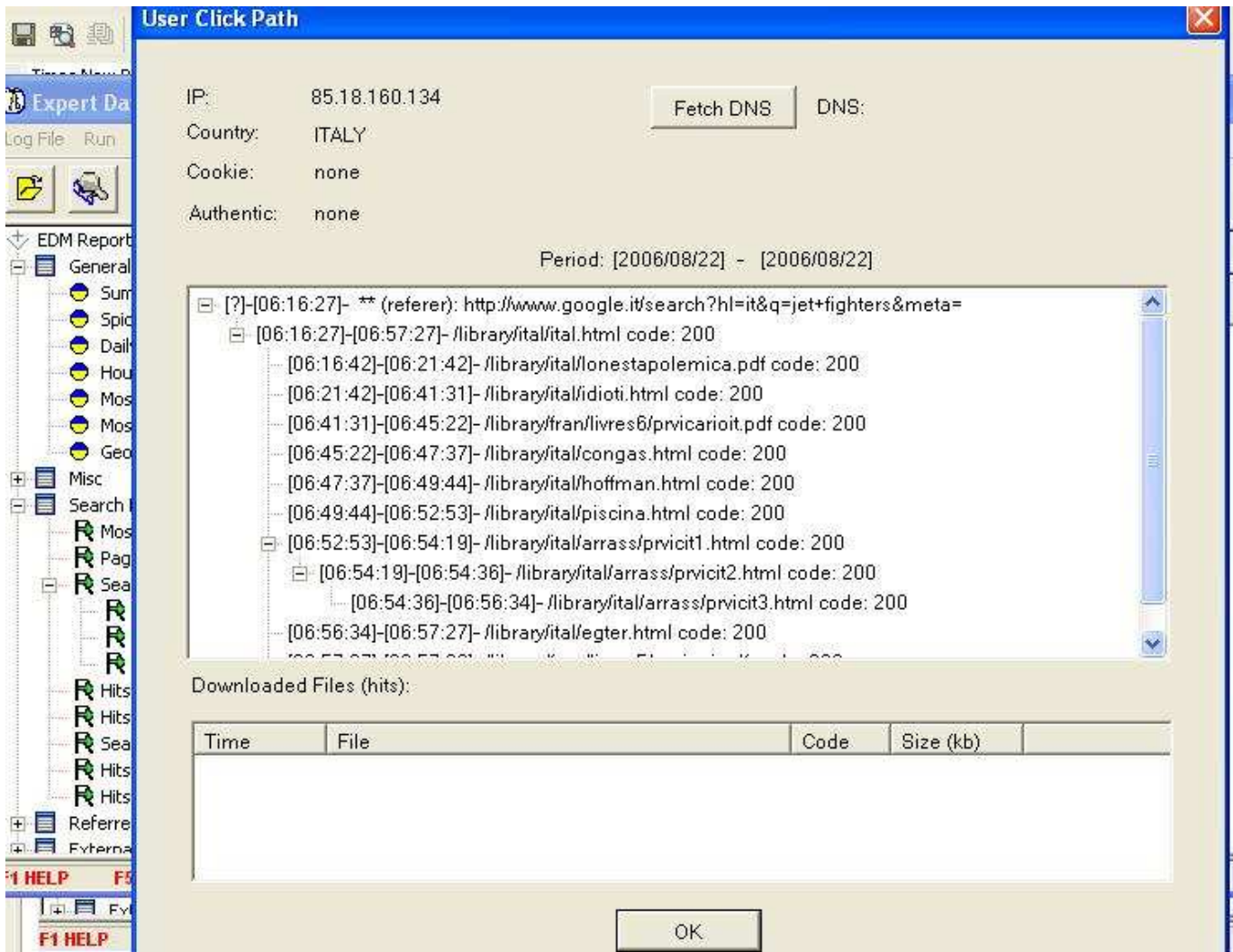
Exit

Referrers vs Entry Pages spirit of st louis 14 1.15 ?

External Referrers

F1 HELP F5 RETRIEVE POSITION FROM SEARCH ENGINE F6 INTERNATIONAL GOOGLE OPTIONS

There are 15 users who got 16 hits, which means that one user clicked twice on his Google link during his session. Let's click on the second line of the above popup now.



If the visitor had downloaded any file, it would be present in the list box below. However the click stream of the user is accessible from this report. Now the click streams of the visitors are accessible not only from this report, but from 19 different reports; all you need is to right click on a line and choose this option when needed. In some cases you can also retrieve an historical chart that corresponds to the line that you view. If you see some anomaly on a curve, you can also click on a peak and the click trail of the visitors for that specific date will appear.

If a user is asking for a page and he wasn't referred by any website or any of your pages, the pages are not nested like the ones that you see above. Often it will be because it's a spider but some browsers will not feed the server with any information in the referrer field also. If a visitor clicks on a link from a page, the second page appears below, shifted to the right. If he presses his back button the next page that appears will be shifted to the left.

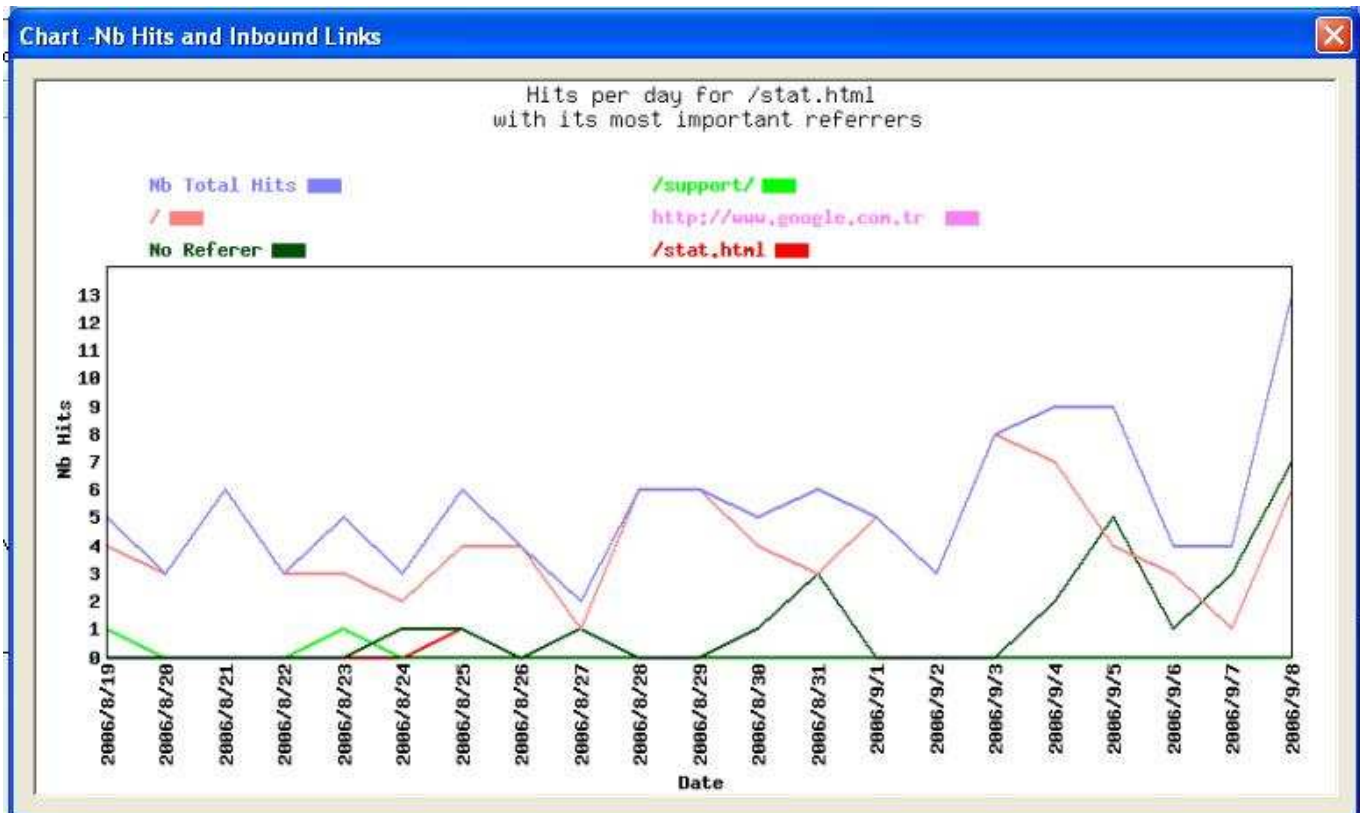
This rule is valid if your visitor asked for less than 2000 pages (except for spiders). Please note that some hits are also discarded for the sake of clarity; For example if one request generates two lines in your log file, one with a code 304 and one with a code 2000, only one line is used to represent the user path.

### 3- Website and Page Optimization

*When one is dealing with website optimization he must also get the proper tools to measure the impact of his changes and test his hypothesis. Expert Data Miner allows you to fetch the history of any page and display the main inbound links behind the fluctuations associated with it. It provides you a group of customizable reports that can be incorporated in your optimization strategy*

Someone can try to boost his position on Google in many ways. The best way is to receive a non reciprocal hyperlink from a 'heavyweight' website, but this is not always easy; sometimes it may take years before to appear in the top ten sites, at least for some keywords. If you are locked with a relatively low page rank, it is still possible to improve your site position on Google, and many will do it with keyword optimization. However this method has its limits, mainly because the page rank is a more important factor. But if someone succeed to double the number of his visitors after a few weeks of effort, it is still a major gain, isn't it?

There is many theories about page optimization, however whatever is your approach you need to assess your hypothesis and see the impact on the pages that matter. In the report Most Popular Pages you have the whole list of your pages and you can sort your data by any field if you click on a header's column. However if you need to focus on a specific page it is possible to retrieve a useful chart when you select the row associated with a page and right click with your mouse.



With this chart you can detect the variations associated with your target page, but also the contribution of the main referrers. In the above example all the referrers except one (the one that starts with http) are internal referrers. So whether you try to dispatch some people from one section of your website to another, or boost your position on Google, you can follow the evolution in one simple graph for *any* page.

### **Optimizing your keywords:**

You can use several reports from Expert Data Miner to improve your strategy of optimization. Many people will only try to optimize their pages in order to be listed among the top 10 sites for the most relevant keywords. However this strategy has little chances to succeed without a decent page rank, especially when too much competitors have a similar goal for the same keywords. The reports on search phrases and word sets will often teach you which other word combinations are performing well. Several of them can be irrelevant, but some others can be interesting prospects. EDM allows you to check quickly your site position on Google or Yahoo for those word sets if you right click on a line and select the *Fetch Position on Search Engine* option from the context menu. This option is available in two of the three reports in the *Search Strings* section. This option is not suitable for a large number of requests, neither for a follow up over weeks (Google Monitor is more appropriate for bulk requests) but it can give you an initial clue. You can quickly see how interesting is such a prospect with the column average number of pages/visitor and the distribution behind it.

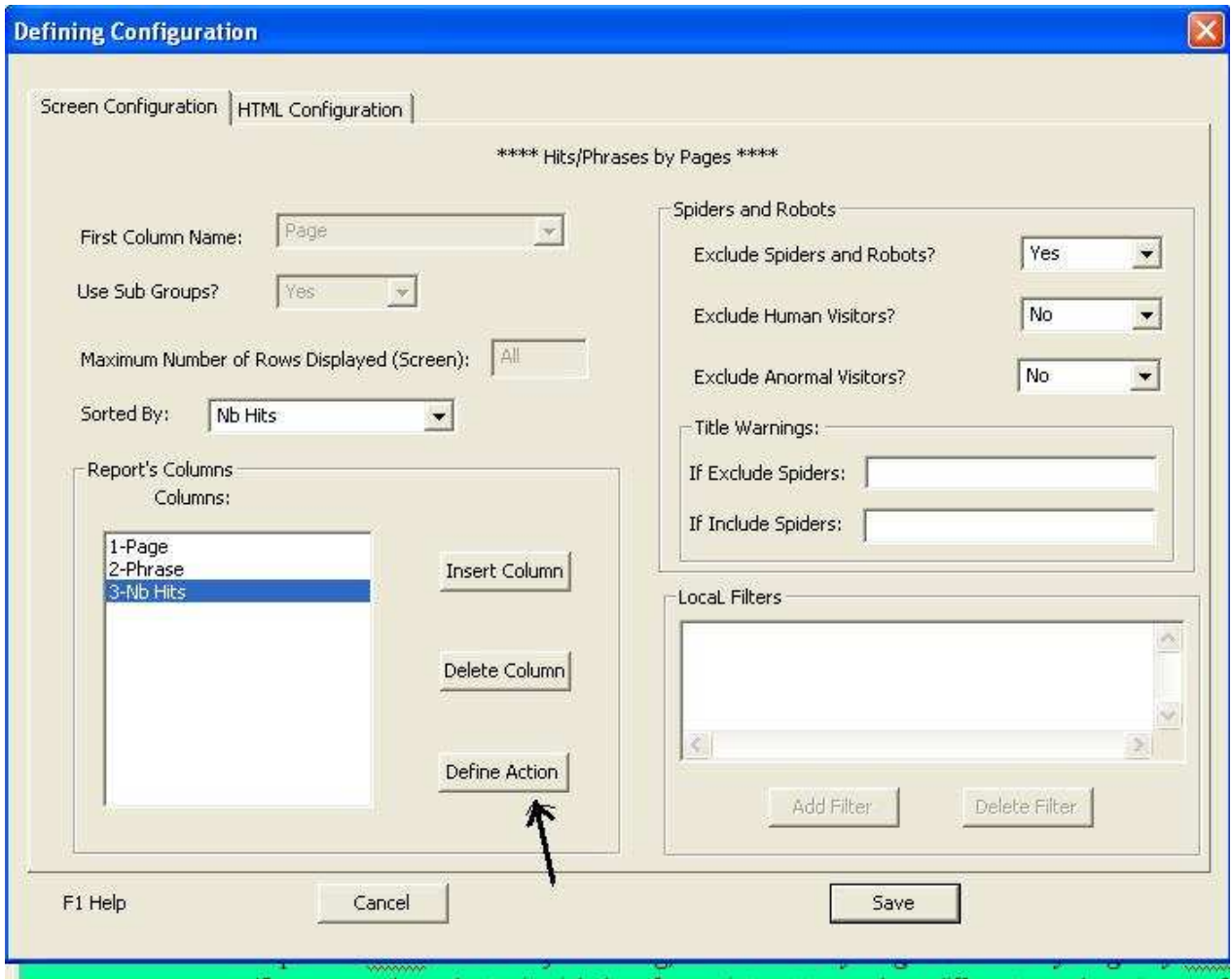
You have then the choice to modify your text slightly and trying to improve your position for this unexpected set of keywords or stick to your initial strategy. You could also wait some days or weeks to see how stable this position around a central point is. Accounting for the fact that many websites who compete for those keywords are changing also (and the Google algorithm as well) it's absolutely possible to see strong variations in a short delay, especially for long phrases. Nevertheless there is groups of search words who will have a relatively stable position. But you need also to see if a sufficient number of users are typing those keywords, or even related word sets. You can use a filter on dates to study the pattern and get the number of hits or use the export option in EDM and store your results of the week in one or several HTML files before to take a decision.

Most people will click on a few links at most when they perform a search. This is why a 6 th position on Yahoo is much better than a 87 th position on Google even if the latest engine is more popular. The same thing could be said about two competing groups of words. If your 87 th position is largely based on a low page rank or a huge number of competitors rather than a poor frequency of the relevant keywords in your text or your paragraph headers/page titles, it is hard to expect a miracle in the short term. One could change slightly his text and try to exploit promising word sets, or split his page in two pages, etc.. Whatever is the strategy, you will often need to test your hypothesis.

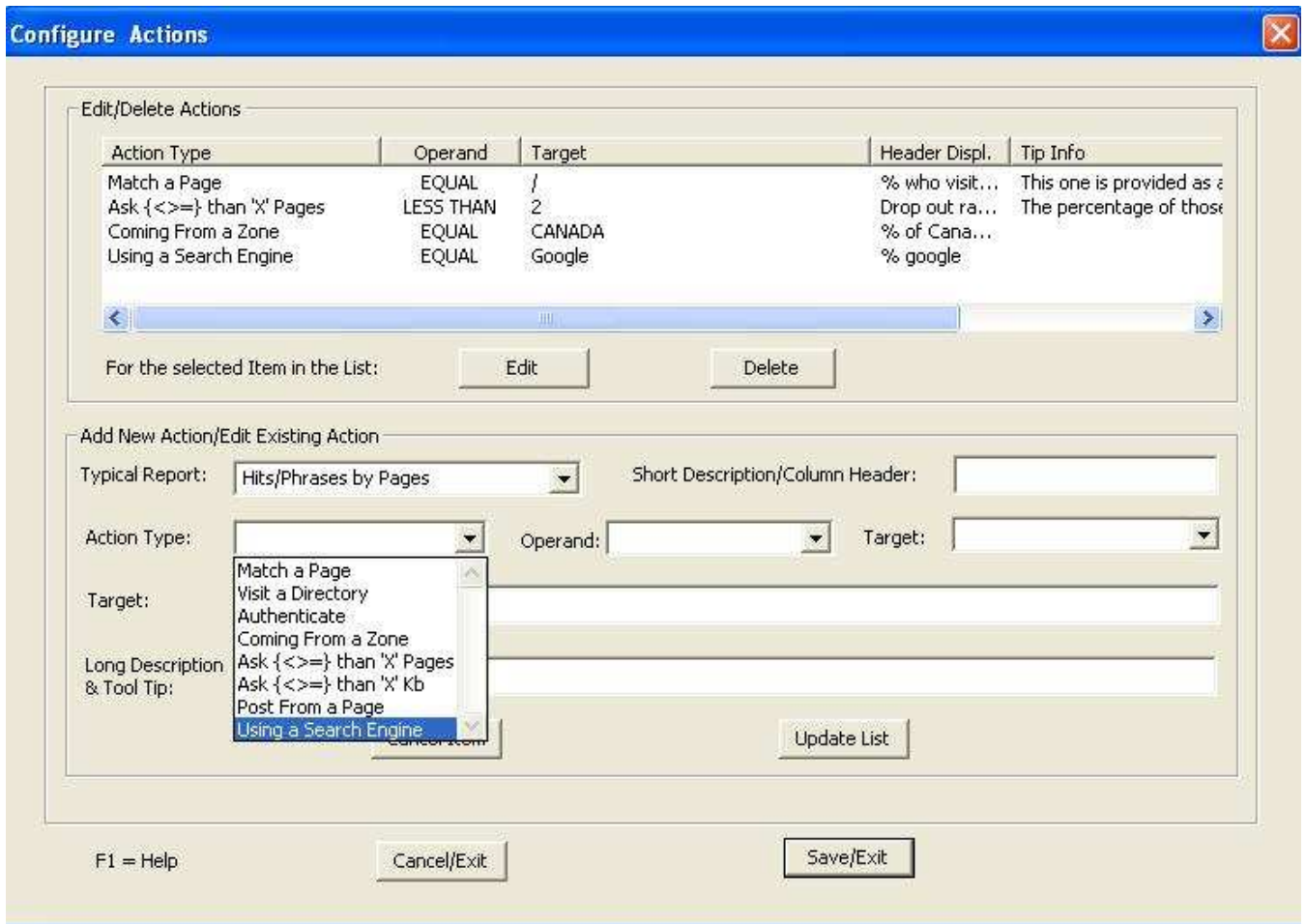
Two other reports in EDM can be very interesting; *Hits/Phrases by Pages* and *Hits by Pages by WordSets* . If you focus on some specific pages and word sets, it might be of some interest to see how different search engines perform for a group of keywords for a



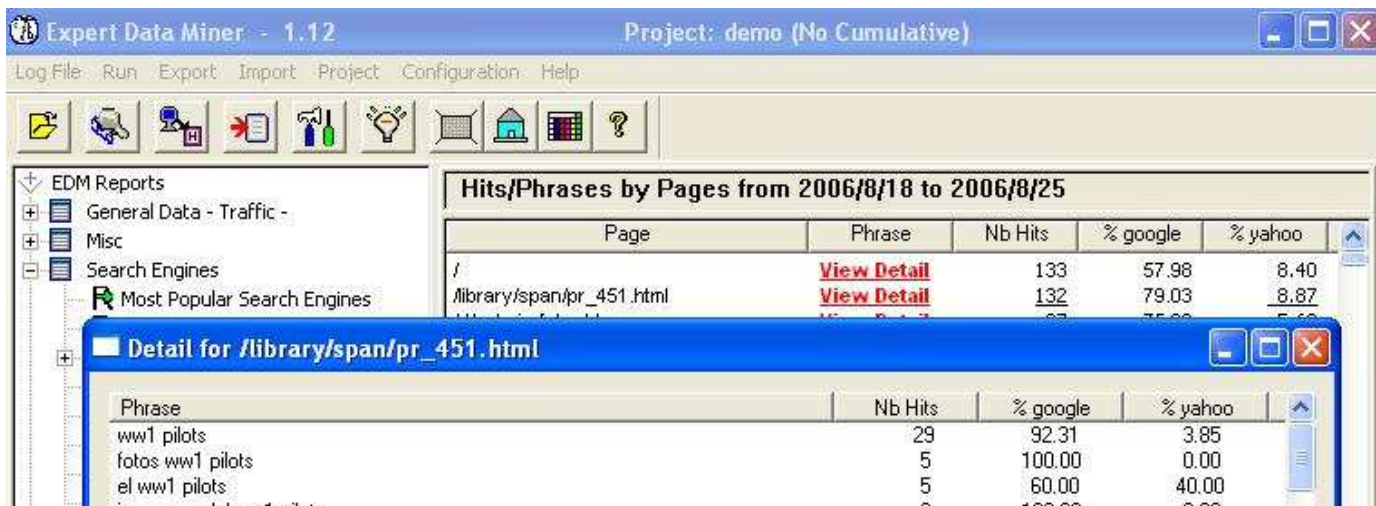
specific page. Let's choose the first of those two reports and click on the button associated with the local configuration.



If we press on the button *Define Action* the following group of customizable columns is available:



The option 'Using a Search Engine' is selected from the combo box. In doing so the other combo box (*Target*) is filled with the list of all possible engines. One can select Google, Yahoo, MSN, AOL search... the new columns will contain the percentage of the visitors who use Google and Yahoo. The log file is then re-scanned.



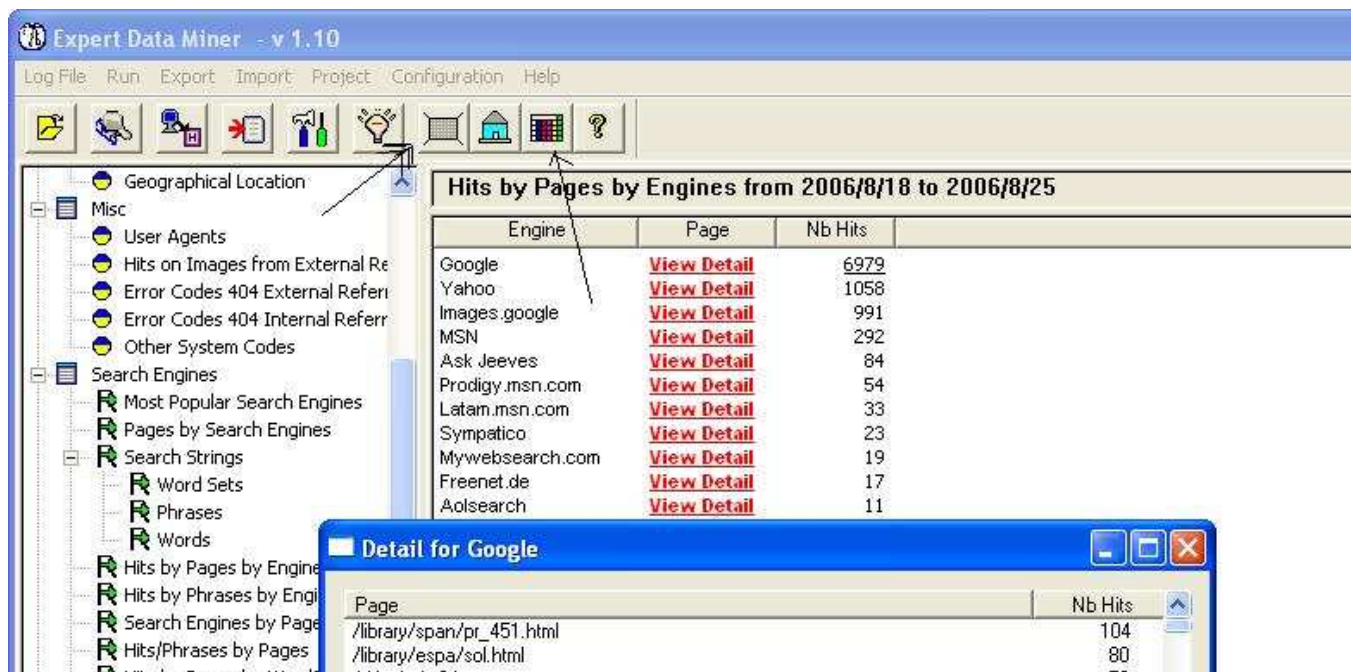
The main screen is now giving the percentage of the visitors who fetched a specific page and who used Yahoo and the percentage who used Google. With a sizeable sample, this percentage will be strongly connected with the position of the page for a search on each engine. But if you click on the hyperlink *View Detail*, the popup 'Detail for...' appears and show a distribution of this percentage on each group of words for that specific page.

As mentioned earlier, it is more profitable to get a 6 th position on Yahoo than a 88 th position on Google. Optimizing and changing your text to put more emphasis on some keywords will often work in the middle/long term, but you may loose ground on another front. Provided that you have a net gain regarding the visitors who matter everything is OK.

## 4- Applying Filters to Your Log Files

Filters can be very useful for those who work in a marketing/sales department. Even those who use the Freeware version of *Expert Data Miner - Log Analyzer* may need filters more than they think. If you are an individual who has just started to develop a website, it can be frustrating to see that your own constant modifications distort the statistics regarding the real number of visitors. If you upgrade your site five times per day, this will have a significant impact if the initial number of visitors is very small. There is a very broad choice of filters in EDM, you can even combine several filters with an AND/OR condition.

Filters can be used from two buttons:



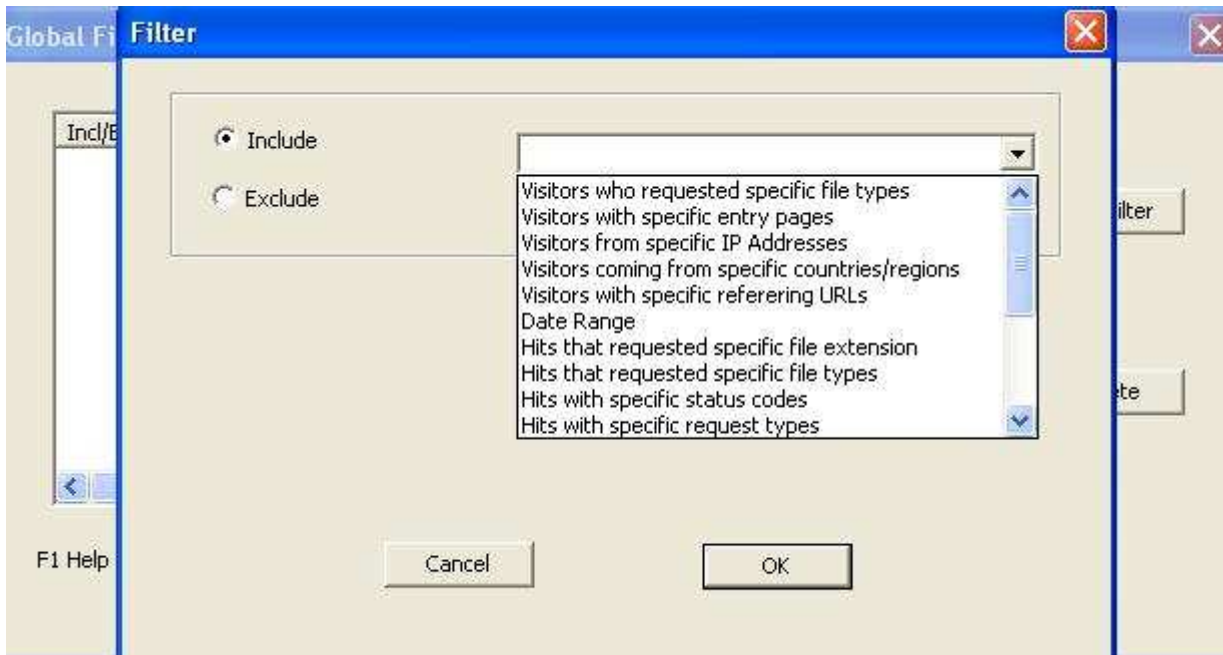
The screenshot shows the Expert Data Miner v 1.10 interface. On the left, a tree view lists various filters such as Geographical Location, Misc, User Agents, Hits on Images from External Re, Error Codes 404 External Referr, Error Codes 404 Internal Referr, Other System Codes, Search Engines, Most Popular Search Engines, Pages by Search Engines, Search Strings, Word Sets, Phrases, Words, Hits by Pages by Engine, Hits by Phrases by Engi, Search Engines by Page, and Hits/Phrases by Pages. A right button (calendar icon) is highlighted with an arrow. The main window displays a table titled "Hits by Pages by Engines from 2006/8/18 to 2006/8/25". A "Detail for Google" window is open, showing a table of page hits for Google.

Engine	Page	Nb Hits
Google	<a href="#">View Detail</a>	6979
Yahoo	<a href="#">View Detail</a>	1058
Images.google	<a href="#">View Detail</a>	991
MSN	<a href="#">View Detail</a>	292
Ask Jeeves	<a href="#">View Detail</a>	84
Prodigy.msn.com	<a href="#">View Detail</a>	54
Latam.msn.com	<a href="#">View Detail</a>	33
Sympatico	<a href="#">View Detail</a>	23
Mywebsearch.com	<a href="#">View Detail</a>	19
Freenet.de	<a href="#">View Detail</a>	17
Aolsearch	<a href="#">View Detail</a>	11

Page	Nb Hits
/library/span/pr_451.html	104
/library/espa/sol.html	80

The net will give you a large number of filters, but the right button (some kind of calendar) is just a quick filter on dates.

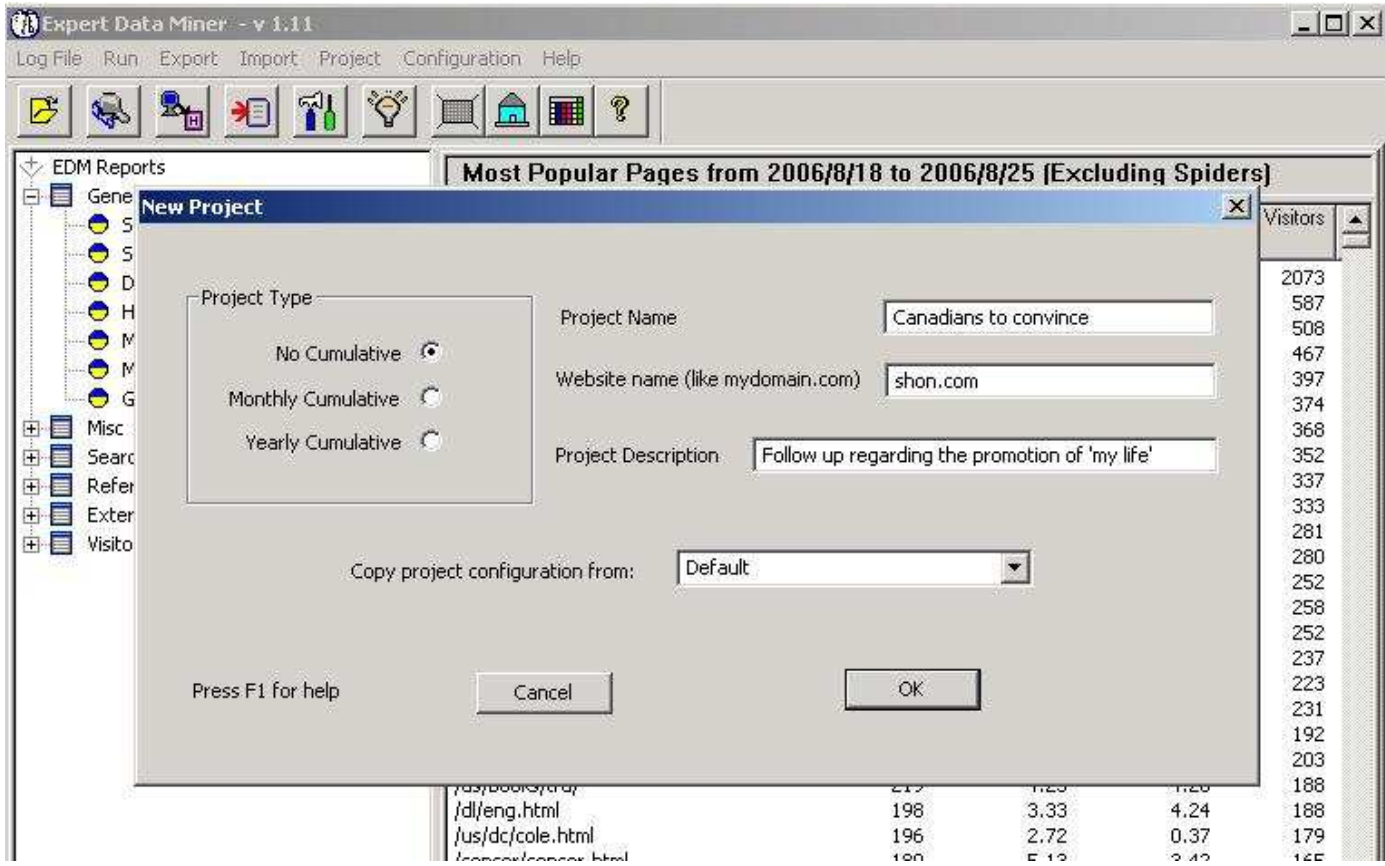


Here the button with a net was selected. You can choose a filter and re-scan your log file. When a filter is active a message appears in the status bar in order to remind you that the global filter mode is on.

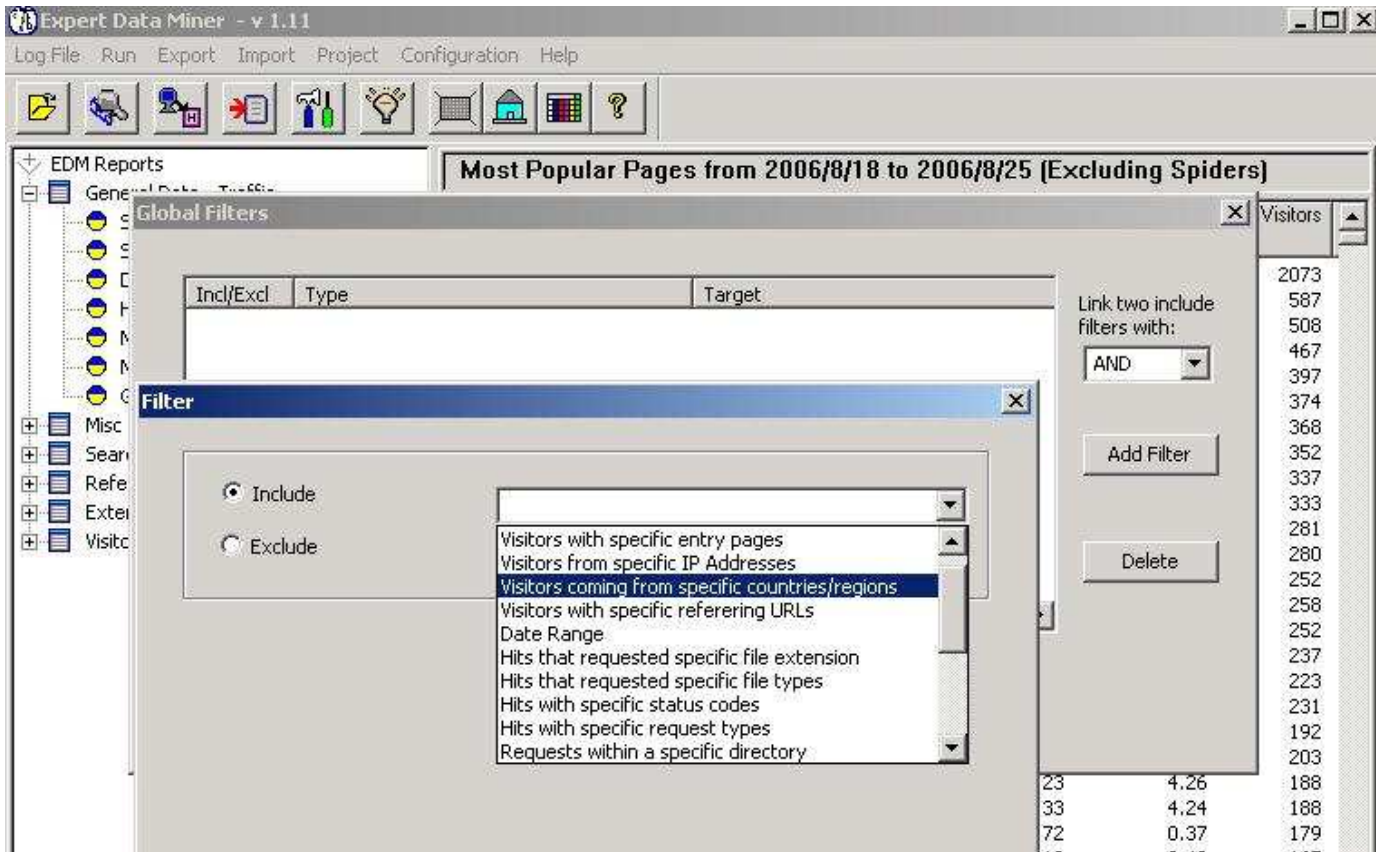
Now if you do not pay attention to this message, you may search for a while before to realize that your old filter needs to be deleted. To delete a filter one has just to select it, press the button *Delete* and then *Save*.

Let say that you have a company and you launched a promotion for Canadians for a specific product. You want to target the visitors from Canada who reached your site but who did not visit one of your page (it can be a page where they can sign up for something). In some cases you can contact them by email in order to make a follow up and explain them what they missed. If you have sold a large volume of cheap items, it can be preferable to study their behavior and adjust your web pages rather than to contact them one by one if your initial strategy failed.

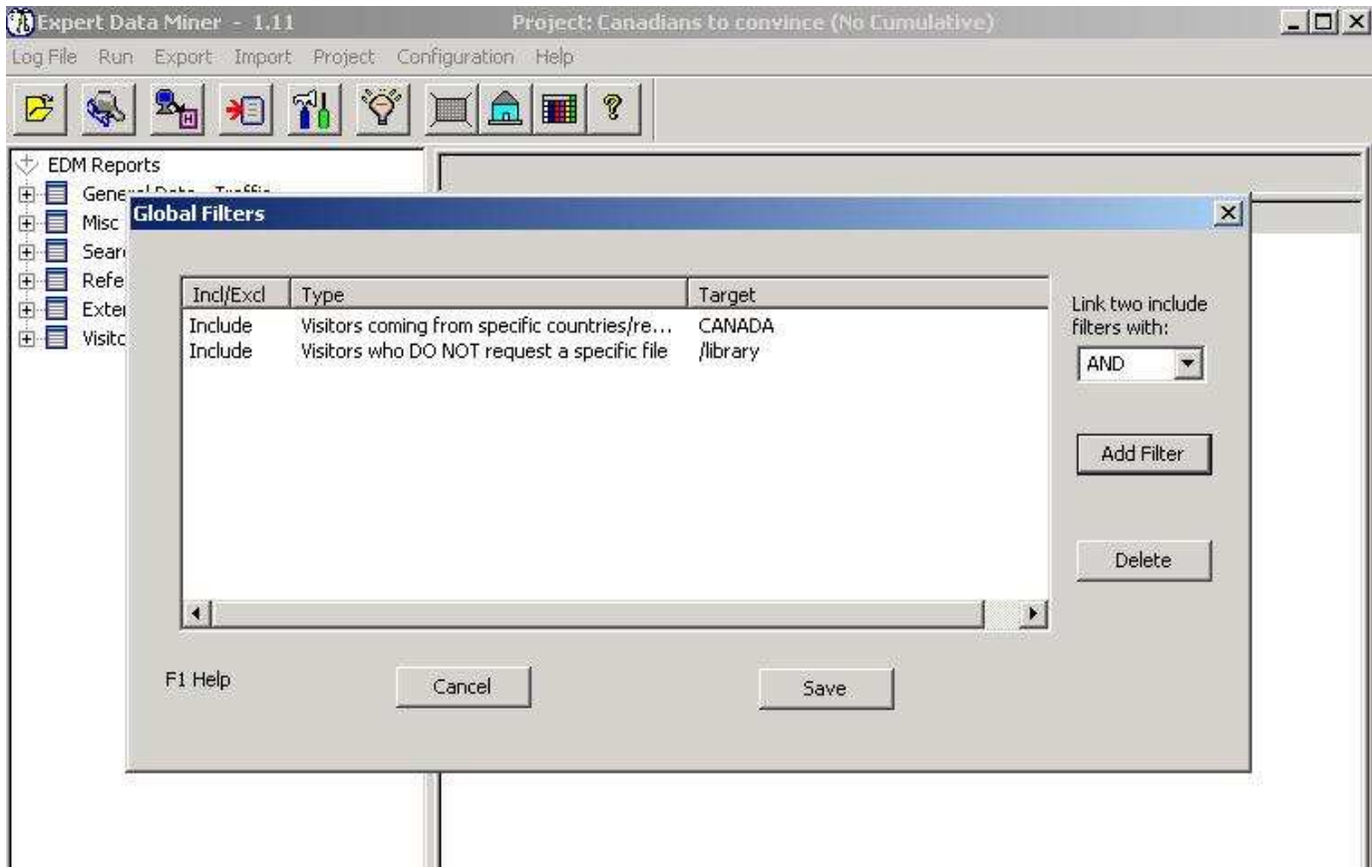
Lets say here that your product costs several hundreds of dollars; tracing back your customer one by one (when you can) is not a waste of time here. For that example, you are doing the promotion of a book to some potential resellers; if they purchase 1,000 of them they get a substantial discount. The first step is to define a non cumulative project for this specific goal:



The second step is to define an include filter on your visitors who are Canadians:



However you also want those who did not visit your target page; the combo box OR/AND allows you to define the condition between your two include filters. In this case it should be set to AND. If you need to define two include conditions linked with an OR, you cannot define an AND with your third include condition, the AND/OR flag is used between ANY include filter. So if this scenario happens one day, nothing prevent you from using an OR condition between two include filters plus an exclusive filter on people who DO visit the target page (in this example we target those who did not visit it with an include filter).



The next step is to save your filter and rescan your log with the proper button. When this is done, you go to the last report, visitors-details



Expert Data Miner - 1.11 Project: Canadians to convince (No Cumulative)

Log File Run Export Import Project Configuration Help

EDM Reports
 

- General Data - Traffic -
- Misc
- Search Engines
- Referrers vs Entry Pages
- External Referrers
- Visitors
  - Visitors - Details

Visitors - Details from 2006/8/18 to 2006/8/25

First Date In	DNS	User ID	User ID Type
2006/08/19 12:22:30		66.198.41.17	IP
2006/08/20 15:01:23		66.198.39.25	IP
2006/08/19 16:26:47		66.198.39.25	IP
2006/08/19 19:53:02		66.198.39.25	IP
2006/08/19 17:00:04		24.71.223.142	IP
2006/08/25 12:51:31		65.92.66.238	IP
2006/08/22 15:52:12		66.198.39.25	IP
2006/08/22 21:18:28		216.254.167.157	IP
2006/08/18 23:24:54		66.154.102.237	IP
2006/08/23 20:45:42		216.254.166.202	IP
2006/08/20 22:53:02		154.5.24.197	IP
2006/08/23 19:34:50		216.254.166.202	IP
2006/08/25 15:41:21		66.198.39.25	IP
2006/08/20 18:17:10		209.183.21.54	IP
2006/08/24 19:23:19		67.70.253.242	IP
2006/08/22 11:50:36	Unresolved	74.56.111.57	IP
2006/08/20 20:14:28		66.130.30.167	IP
2006/08/20 09:31:31		66.130.175.223	IP
2006/08/21 09:31:28		216.46.18.194	IP
2006/08/22 22:42:16		216.254.167.157	IP
2006/08/25 21:26:07		65.93.4.72	IP
2006/08/19 19:55:08		216.254.156.197	IP
2006/08/25 21:41:18		66.131.66.143	IP
2006/08/19 16:42:13		24.203.149.14	IP
2006/08/20 22:09:01		24.71.223.148	IP

It's not a bad idea to select the first 100 items and right click to fetch their DNS because you may need this information. Now let say that you want to know what happened with a specific customer; as usual you select the row, right click and choose to view the click stream:

**User Click Path**

IP: 154.5.24.197 Fetch DNS DNS:

Country: CANADA

Cookie: none

Authentic: none

Period: [2006/08/20] - [2006/08/20]

- [-]-[22:53:02]- \*\* (referer): http://www.google.com/search?hl=en&sa=X&oi=spell&resnum=0&ct=result&cc
- [-]-[22:53:02]-[22:53:12]- /us/dc/cole.html code: 200
  - [22:53:12]-[22:55:24]- /dl/eng/cole.ra code: 200
  - [22:55:24]-[-] /dl/eng/cole.ra code: 200

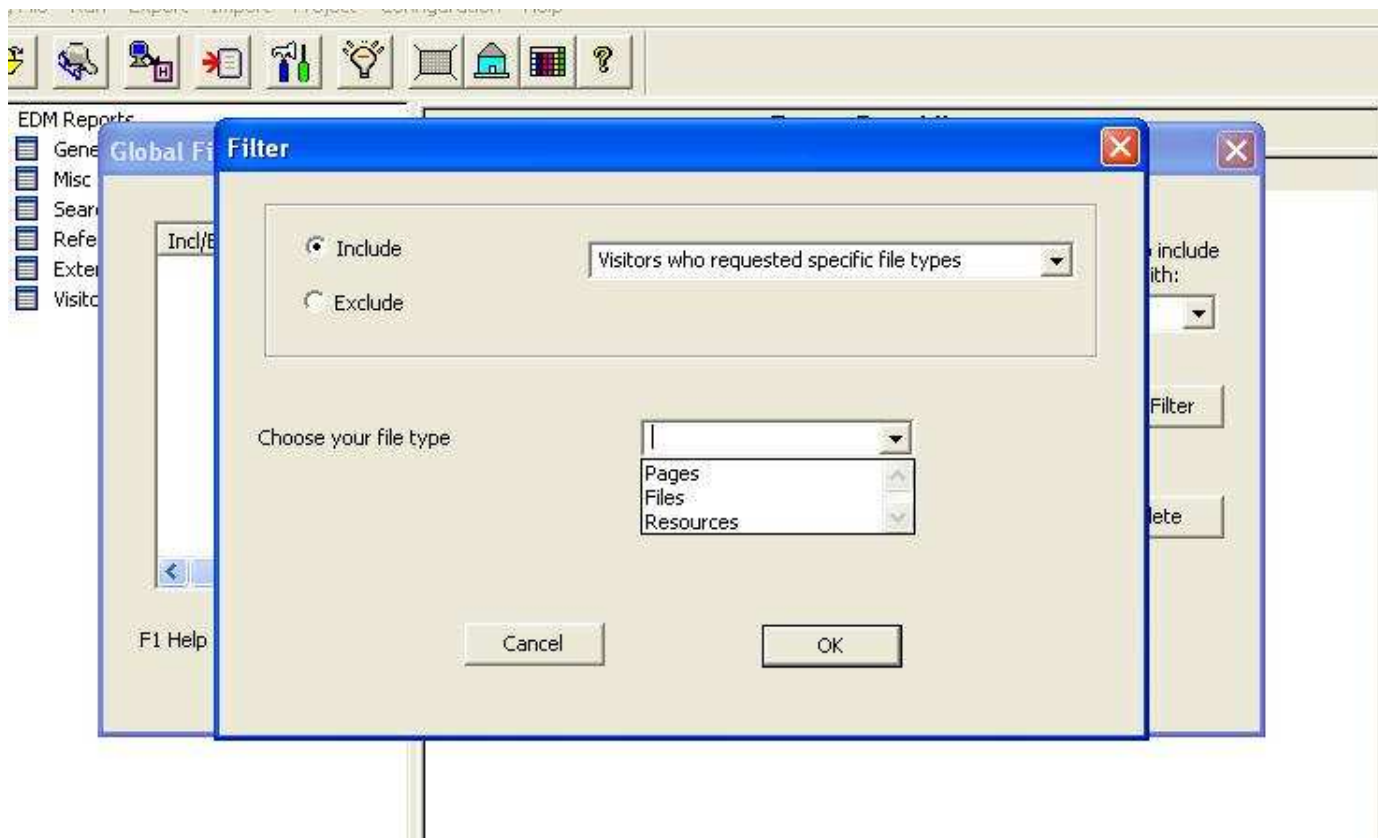
Downloaded Files (hits):

Time	File	Code	Size (kb)

In some cases the DNS itself contains the name of the company (your visitor). In other cases your visitor can be a registered visitor who logged on, so his authentication ID will be present; if he provided his email some weeks ago when he purchased some article you probably have a table that can link the cookie and email addresses of this visitor to a user session. In the Enterprise Edition of EDM you can import such links and build reports on the fly.

Lets now see another example; you may be puzzled about the real number of visitors that your website receives, in many situation this number is drastically inflated with some kinds of 'false visitors'. Many of the 'visitors' on <http://www.expertdataminer.com/> are false visitors. Presently most of them come from a website called *software dungeon* but in a month they will come from elsewhere. What happens is that several download websites will list this site among many others and each time that someone opens a specific page it will load an icon or a small picture from dozens of remote websites who announce their products in the same category. Most of those people will never ask a page from this website and many will still ignore its existence at the end of the day. Those

people can be filtered out if I create two inclusive filters, each on the *Visitors who requested specific file types*, except that from the combo box the first target will be *Pages* and the second *Files*; Resources are left out.



In this case an OR condition is selected from the combo box above the button Filter.

If you have a small website, you can also filter yourself out from the statistics. One possible filter is *Visitors with a specific authentication ID*; you probably use this ID when you try to use your file manager or some other tool. If you are prompted for a userID/password before to connect to those tools (Cpanel, file manager or anything like this) your user ID is probably present in your log file. Another possible filter that you can use is *Visitors who requested a specific file*; Nobody else than you should use your initial management page. In these cases you should use an *Exclusive* filter before to re-scan your log.

If you use a filter on a specific page or directory, you must use the standard that appears in your log file, so normally when you are prompted for a page name you should type `/mypage.html` or `/mypage.asp?id=9643` and not `http://mydomain/mypage.html`.

## 5 - Fraud Detection – Pay per Click

The *pay per click* system is an enormous source of income for Google, but also for many other web sites. The companies who choose to promote their website with the pay per click system need to be careful regarding the amounts for which they are billed. There is a growing number of frauds and they take miscellaneous forms.

Some web sites or affiliates can pay people from remote places like Botswana to make fraudulent clicks on an ad in order to inflate their customer's bills. Since the year 2006 click frauds are not limited to such methods, now malwares like click bots can be used for such a goal. These small pieces of code can be spread like viruses on many computers in order to generate clicks from different IP addresses. The most intelligent scams involve a malware that adopts a low profile and generate only a few clicks per computer in order to avoid detection. These bots are generally controlled remotely by the person who wishes to limit the clicks to ads that can generate a real profit.

Of course Google is not involved behind such a scam even if it happens often that PPC users ask themselves questions regarding the amount for which they are billed. For example you may wish to avoid paying too much if a visitor (let say a competitor) is clicking many times on the same ad from the same IP address in order to exhaust your budget. Google is doing itself an effort to reduce click fraud, but this company is posing itself as a judge even if it is also making a profit from PPC. An extra tool is not useless if you wish to see in details how Google is billing your clicks.

A log analyzer like Expert Data Miner has many functionalities that makes it unique. This software allows you to detect many cases of click fraud but also to understand in details the results for any referrer. EDM will check for the duplication of hits, verify if the IP address of a visitor is an anonymous proxy, check if the visitor's browser fetched the images associated with your target page (normally bots do not), verify if there is a pattern with cookies and perform a statistical analysis according to countries. Getting all those information IS important. You can try to get a refund from Google for past cases of click fraud, but reacting quickly when a dangerous pattern is underway is certainly wiser. From your Google account you can use two of their filters to prevent clicks from your competitors or some dishonest Internet user. The filter on IP ranges will work fine for visitors who have a static IP address, the filters on locations (like cities) can be used in the worst case if you target a broad market and your competitor is coming from a small city.

Lets take the report dealing with Google Syndication or either *Google AdWords*, (*Content Network*): In this report, EDM gives you the list of all the referring websites and the number of visitors who were sent to your site. If you right click on one of those referrers, you get the list of all the visitors and their unique IP or cookie. In the following screen a popup (the click trail) appears when one of the lines is chosen:

Expert Data Miner - V 1.37.1 Project: edm (No Cumulative)

Log File Run Export Import Project Configuration Help

Google AdWords from 2007/11/12 to 2007/11/25

Referrer	Nb Hits	Nb Visitors	Avg nb pages per visitor
www.mixmap.com	78	71	<b>1.11</b>
awstats.sourceforge....	45	40	<b>2.25</b>
www.mstracker.net	41	40	<b>1.18</b>

**Click Path**

Select the desired element and double-click to view the user path  
(click on a column header to sort)

IP	Time In	Nb Pages	Nb Hits-Files	Country
218.165.35.231	2007/11/19 00:10:56	2	1	TAIWAN
155.69.5.234	2007/11/16 00:26:01	1	0	SINGAPORE
222.151.92.240	2007/11/17 14:07:24	1	1	JAPAN
218.102.236.38	2007/11/17 18:28:29	1	0	HONG KONG
125.229.166.173	2007/11/18 02:19:40	1	0	TAIWAN
211.20.132.224	2007/11/18 07:30:30	1	0	TAIWAN
61.59.233.57	2007/11/18 11:18:22	1	0	TAIWAN
61.224.247.134	2007/11/18 15:01:07	1	0	TAIWAN
220.131.169.146	2007/11/18 15:52:31	1	0	TAIWAN
210.146.45.34	2007/11/18 16:37:22	1	1	JAPAN
210.200.105.228	2007/11/18 21:01:59	1	0	TAIWAN
124.8.44.212	2007/11/18 21:30:54	1	0	TAIWAN
211.20.150.76	2007/11/18 21:50:07	1	0	TAIWAN
219.68.240.72	2007/11/20 05:15:12	1	0	TAIWAN
210.64.245.187	2007/11/21 22:34:59	1	0	TAIWAN
210.200.105.228	2007/11/22 15:07:59	1	0	TAIWAN

F1 HELP Right click on a line

H Hits par Pages par Groupes

Moteur de Recherche Intern

When someone clicks on one of the lines in this popup he gets the details for this visitor:

**User Click Path**

IP: 218.165.35.231      Fetch DNS      DNS:

Country: TAIWAN

Cookie: 010-1690085903      Comment:

Authentic: none

Period: [2007/11/19] - [2007/11/19]

```

[-]-[00:10:56]- ** (referer): http://pagead2.googleadsyndication.com/pagead/ads?client=ca-pub-012550180452351
[-]-[00:10:56]-[00:11:03]- /download.asp?lang=eng&param8=none&query1=4 code: 200
[-]-[00:11:03]-[?]- / code: 200

```

Downloaded Files (hits):

Time	File	Code	Size (kb)
[00:11:03]	/fra/trial/edminst.msi	200	28.48

Hits with error codes for pages or hits on resource files and images not shown

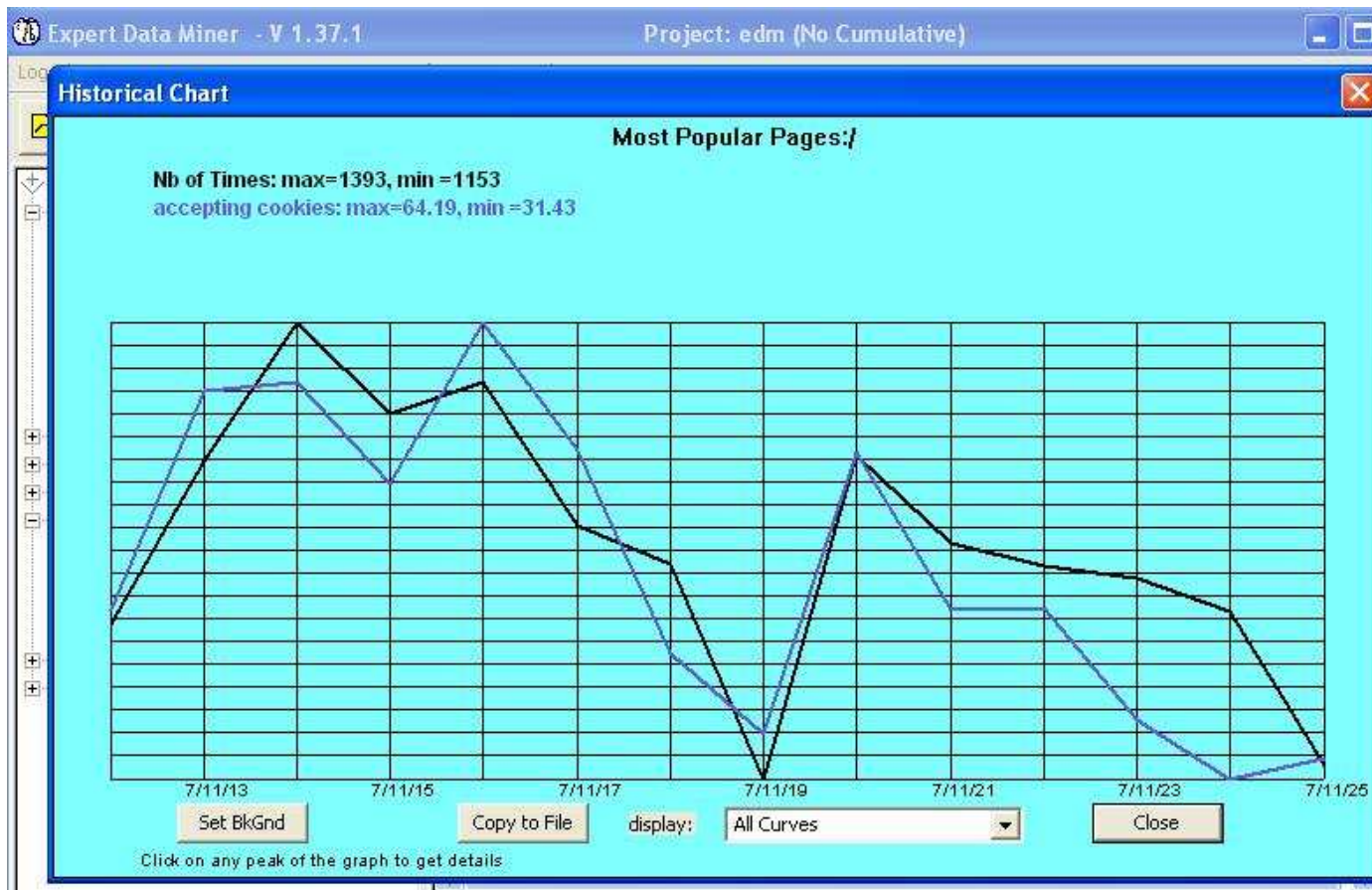
Previous Visit      Close      Next Visit

Here this visitor started to ask for the page *download.asp*; then he clicked on a link inside it and fetched the root page at 00:11:03 (the "/" is the root page ). If you store the cookies of your visitors in your log files the software allows you to fetch their behavior during a subsequent visit (the button *Next Visit* at the bottom of the window near the button *Close* ). Since this visitor didn't come back the button is grayed. So it is easy to detect suspicious behaviors associated with click frauds. If you fetch the DNS you can sometimes get the name of the server of your visitor (in some cases it can be a competitor). You can also sort your visitors according to their IPs or their cookies if you click on the header of the relevant column in the popup.

It is also possible to fetch the same kind of information for external websites who will charge you for an ad (not Google Ad Words here) in another report. In the fact this information is available by keywords, landing pages, referrers,, etc...

It can be somewhat more difficult to detect a click bot. If such a malware propagated itself you'll get many clicks from miscellaneous IP's and some click bots will adopt a low profile and perform a few clicks per day to avoid detection. However these robots have some features that distinguish them from human visitors and the periodicity of the clicks is one of them. The probability to see a human being clicking on your ad every 900 seconds 4 times in row is weak. But bots have also another characteristic; they do not support JavaScript, so cookies. They are not true browsers. If you configure your server to store the cookies that you assign to your visitors at the end of each line in your log files (something easy for Google Analytic's cookies) it is possible to detect indirectly an abnormal behavior for a set of visitors.

Indeed, Expert Data Miner allows you to apply filters based on a segment of the name of your pages, your whole pages, or a segment of the referrer. It is thus possible to isolate those who find your site with PPC and those who find it through an organic search - a search that doesn't cost you a penny -. It is the presence of a abnormal fraction of visitors whose browsers refuse cookies that will raise a concern. In many reports, you can build a column from scratch and one of them concerns the number of visitors whose browser accepted your cookies. This column is available as a percentage but if you right click you get an historical chart for any of your pages:



The blue curve gives the percentage of the visitors who ask for the page "/" (the root page) and whose browser accept cookies. This percentage varies from 31.43 to 64.19% in this example. In the fact this percentage, for human visitors, should be above 90%.

But you can also get details regarding click frauds from the report Google AdWords. If you press the F6 key, the software will scan for strange patterns and provide you a description of such cases (including the IP address of the visitors). EDM will regroup your visitors with their IP, their cookie, or provide useful data regarding the top countries. If 1.4% of your visitors are coming from a country but own 18% of the PPC clicks, this is quite abnormal. If a competitor clicks on your ads 10 times a day and repeats the same operation the day after (when his provider assign him a new dynamically allocated IP address) you can often know it from the cookie.

Most of the fraud cases come from unscrupulous webmasters who are AdSense members (the network content) and clicks on your ads to boost their profits. In a growing number of cases the task can be delegated to third world countries, especially if your bids per clicks are high. There is very little cases of fraud from searches done on Google except from competitors. But still if you have a doubt, if your competitor is not using cookies and doesn't use a static IP address, you can still get his ISP name and get the IP range from this provider. Especially when it's a small ISP you can apply a filter in EDM to see what's going on.

Two other useful reports are also available when you launch the task 'SCAN FOR CLICK FRAUD'. One is mainly targeting potential click bots. Most bots will not load the images associated with a target page; EDM can check what is the maximum number of images that at least one visitor fetched with that page and give you the IP address, the cookie and the time associated with any visitor who did not ask for the images or resource files (or a small fraction of them). The second report will scan for IP addresses related to anonymous proxies.

If you use another system than Google AdWords, EDM can also be used to scan for click fraud in the report *External Referrers, Summary*. All you need to do is to right click on the referrer of your choice and mark it as a paid referrer. The next time that you scan your log file(s), EDM will build some extra statistics for this referring website. Retrieving them is quite easy with the key F6.

## **Getting the ROI in a Pay per Click campaign**

Someone may wish also to optimize his investment and suppress the keywords that do not generate enough profits or to know the percentage of the visitors who purchase something. In Web Analytics this is called the conversion rate. Unlike many other log analyzers, EDM doesn't limit itself to the conversion rate for the current session, it can also check if a referred visitor purchased something several days later. Let's take the report 'pages accessed from a search engine':



Expert Data Miner - V 1.37.1 Project: edm (No Cumulative)

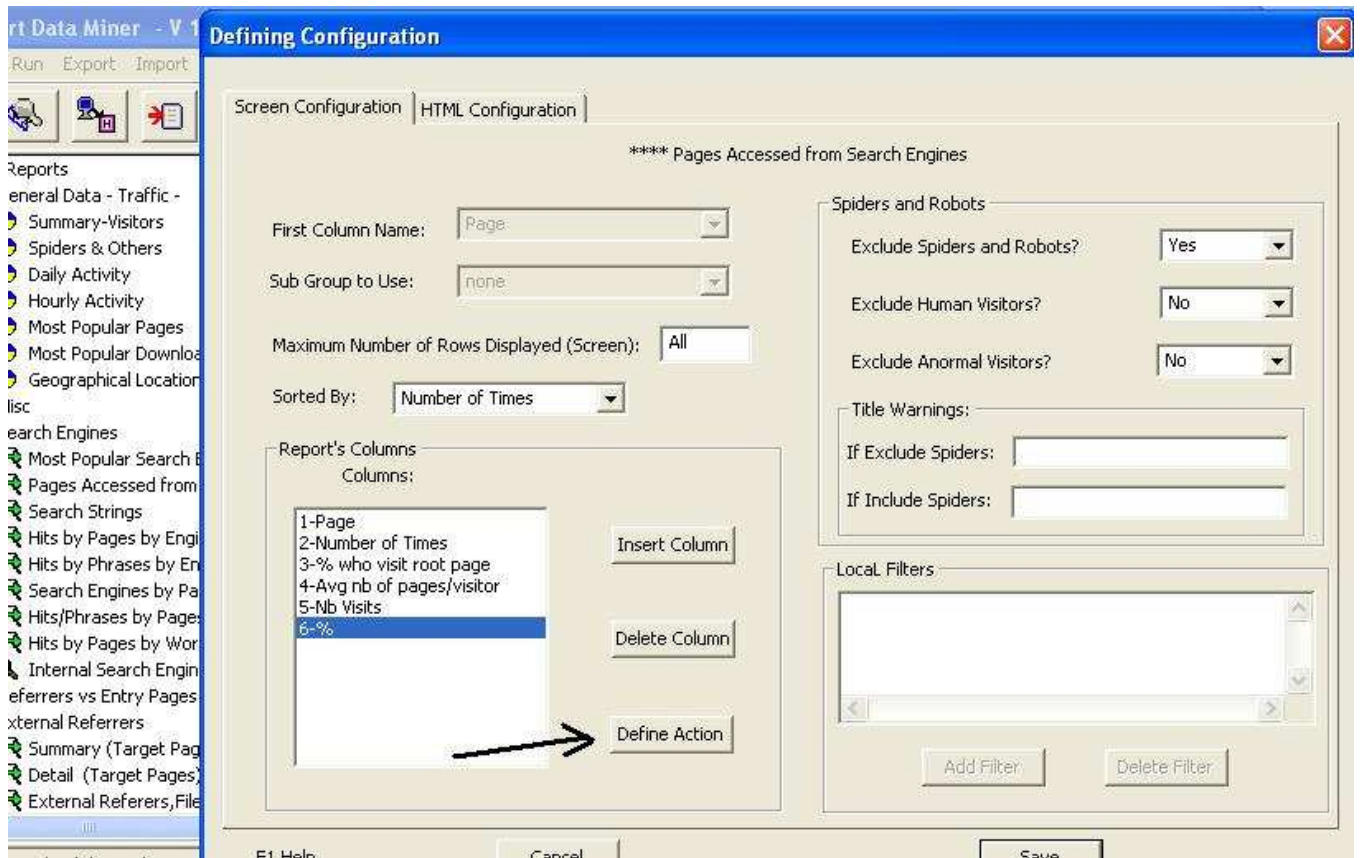
Log File Run Export Import Project Configuration Help

EDM Reports

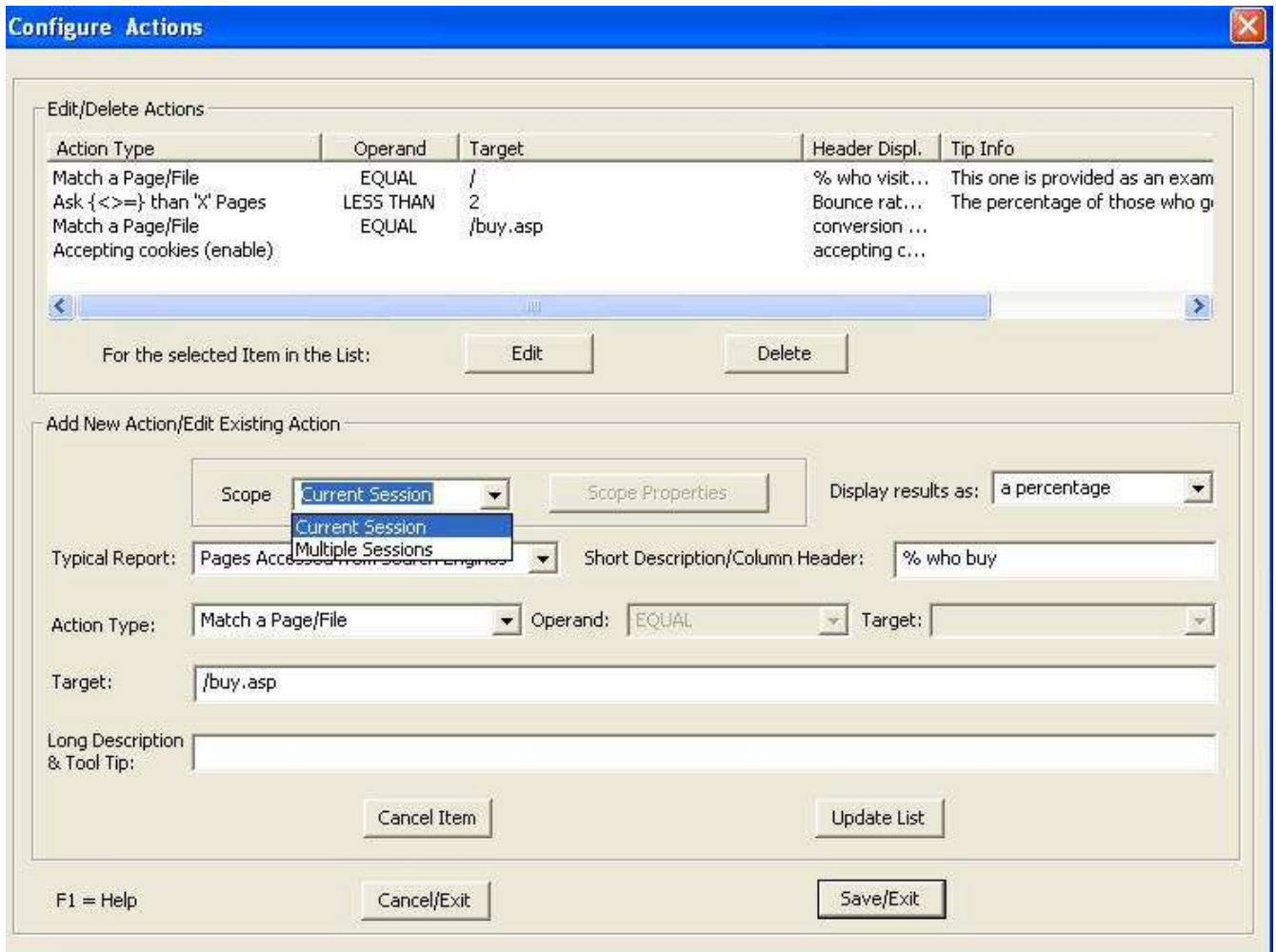
- General Data - Traffic -
  - Summary-Visitors
  - Spiders & Others
  - Daily Activity
  - Hourly Activity
  - Most Popular Pages
  - Most Popular Downloads
  - Geographical Location
- Misc
- Search Engines
  - Most Popular Search Engines
  - Pages Accessed from Search En**
  - Search Strings
    - Hits by Pages by Engines
    - Hits by Phrases by Engines
    - Search Engines by Pages
    - Hits/Phrases by Pages
    - Hits by Pages by WordSets
  - Internal Search Engine
- Referrers vs Entry Pages
- External Referrers

Pages Accessed from Search Engines (pages do not need always to be entry)						
Page	Number of Times	% who visit root page	Avg nb of pages/visitor	Nb Visits	%	
/download.asp	511	9.20	<u>2.73</u>	435	61.44	
/	260	100.00	<u>3.22</u>	198	27.97	
/buy.asp	8	37.50	<u>4.63</u>	8	1.13	
/website-tracking.asp	8	0.00	<u>3.00</u>	7	0.99	
/faq.asp	6	16.67	<u>1.33</u>	6	0.85	
/log-viewer.asp	6	0.00	<u>2.80</u>	5	0.71	
/log-file-viewer.asp	5	0.00	<u>2.25</u>	4	0.56	
/performance-tracking-log.asp	4	0.00	<u>1.00</u>	4	0.56	
/track-website-visitors.asp	4	0.00	<u>2.67</u>	3	0.42	
/web-site-usage-tracking.asp	4	50.00	<u>7.00</u>	2	0.28	
/contacts.asp	3	33.33	<u>2.33</u>	3	0.42	
/analyze-log-files.asp	2	0.00	<u>1.00</u>	2	0.28	
/log-analyzer.asp	2	0.00	<u>2.00</u>	2	0.28	
/modifications.asp	2	0.00	<u>3.50</u>	2	0.28	
/track-website-stats.asp	2	0.00	<u>1.00</u>	2	0.28	
/web-stats-software.asp	2	0.00	<u>3.50</u>	2	0.28	
/windows-log-file-analysis.asp	2	0.00	<u>1.00</u>	2	0.28	
/articles.asp	1	0.00	<u>2.00</u>	1	0.14	
/download/edm.pdf	1	0.00	<u>1.00</u>	1	0.14	
/edm.asp	1	0.00	<u>1.00</u>	1	0.14	
/history.asp	1	0.00	<u>2.00</u>	1	0.14	

If you click on the button with a hammer and a screwdriver, you get the configuration screen for the current report:



If you click on the button *Define Action* you get the following page:



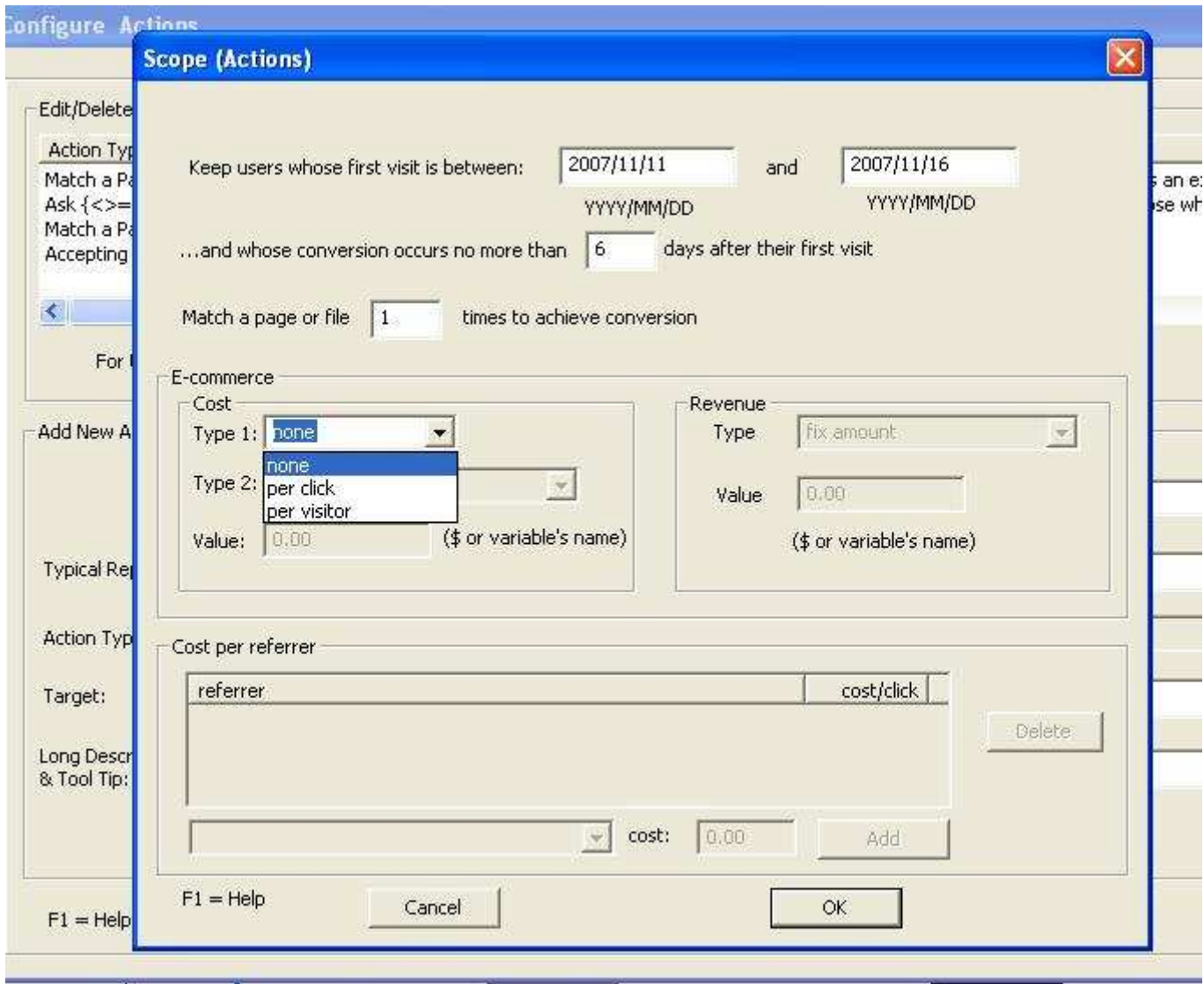
The fields were filled to tell the software that:

1) Asking for the page /buy.asp ("/" being the root page of the site) is considered as a purchase. In the fact it would be more accurate to select a page that is displayed once a visitor *paid* for something (*Thank you for purchasing with us!*).

2) The result will be displayed as a percentage.

In short, you ask the software to build a new column that will display the percentage of the visitors who asked for the page buy.asp during their session regardless of their landing pages. In other reports this column could be available for search keywords, the referring websites, etc...

But you don't want just to know if these visitors purchased something during their first visit but also if they bought something some days or weeks after. Since you store the cookies of these people in your log files you just need to go in the box *scope* and choose 'Multiple Sessions' rather than 'Current Session' from the combo box and click on the button *Scope Properties*.



Let say that you wish to find the visitors who found your website with a search engine between November,11 2007 and November, 16 of the same year. You want to isolate those who purchased your products at most 6 days after their first visit. Since you are using the PPC system and not organic search your invoice is determined either by the number of clicks (even if the same visitors is asking often for the same page) either by the number of visitors who performed at least one click. From the second combo box you can choose how much it costs you in two ways: if the URL is asked in such a way : <http://www.mydomain.com/mapage.asp?p=9222&source=Google> you assign an amount of 1.76\$ or euros when you tell the software to use the variable 'cost' each time that it appears in your log file. You can also assign a fix cost for each referring website. Revenues are calculated the same way, you can assign a fix amount each time that buy.asp is called (let say your average sales) or either modify your pages so that a variable 'price' appears when someone calls that page, like in /buy.asp?price=43.44. This variable will be ignored by your application but stored in your log files.

One can build several columns like this in several reports and play with some parameters like the initial date, the necessary time to achieve a conversion, etc....

Once you re-analyze your log files you get the following results:

**Expert Data Miner - V 1.37.1** Project: edm (No Cumulative)

Log File Run Export Import Project Configuration Help

**Pages Accessed from Search Engines (pages do not need always to be entry pages)**

Page	Number of Times	% who visit root page	Avg nb of pages/visitor	Nb Visits	%	% who buy after 6 days
/download.asp	511	9.20	<u>2.73</u>	435	61.44	9.32
/	260	100.00	<u>3.22</u>	198	27.97	24.00
/buy.asp	8	37.50	<u>4.63</u>	8	1.13	100.00
/website-tracking.asp	8	0.00	<u>3.00</u>	7	0.99	33.33
/faq.asp	6	16.67	<u>1.33</u>	6	0.85	0.00
/log-viewer.asp	6	0.00	<u>2.80</u>	5	0.71	33.33
/log-file-viewer.asp	5	0.00	<u>2.25</u>	4	0.56	0.00
/performance-tracking-log.asp	4	0.00	<u>1.00</u>	4	0.56	0.00
/track-website-visitors.asp	4	0.00	<u>2.67</u>	3	0.42	100.00
/web-site-usage-tracking.asp	4	50.00	<u>7.00</u>	2	0.28	0.00
/contacts.asp	3	33.33	<u>2.33</u>	3	0.42	0.00
/analyze-log-files.asp	2	0.00	<u>1.00</u>	2	0.28	0.00
/log-analyzer.asp	2	0.00	<u>2.00</u>	2	0.28	0.00
/modifications.asp	2	0.00	<u>3.50</u>	2	0.28	0.00
/track-website-stats.asp	2	0.00	<u>1.00</u>	2	0.28	0.00
/web-stats-software.asp	2	0.00	<u>3.50</u>	2	0.28	100.00
/windows-log-file-analysis.asp	2	0.00	<u>1.00</u>	2	0.28	0.00
/articles.asp	1	0.00	<u>2.00</u>	1	0.14	0.00
/download/edm.pdf	1	0.00	<u>1.00</u>	1	0.14	0.00
/edm.asp	1	0.00	<u>1.00</u>	1	0.14	0.00
/history.asp	1	0.00	<u>2.00</u>	1	0.14	0.00

9.32% of those who found the page /download.asp of this website between the 11 th and the 16 th of November with a search engine did ask for the page buy.asp within the following 6 days. Those who entered your website directly on /buy.asp obviously asked for the same page in 100% of the cases. In case of doubt, you can always right click on a line to get the list of the visitors, their IPs, etc...

If you press the F9 key, it's no longer the percentage that is displayed but the net revenue:

Expert Data Miner - V 1.37.1 Project: edm (No Cumulative)

Log File Run Export Import Project Configuration Help

Pages Accessed from Search Engines (pages do not need always to be entry pages)

Page	Number of Times	% who visit root page	Avg nb of pages/visitor	Nb Visits	%	Net Revenue
/download.asp	511	9.20	2.73	435	61.44	1171.00
/	260	100.00	3.22	198	27.97	1219.00
/buy.asp	8	37.50	4.63	8	1.13	343.00
/website-tracking.asp	8	0.00	3.00	7	0.99	47.00
/faq.asp	6	16.67	1.33	6	0.85	(3.00)
/log-viewer.asp	6	0.00	2.60	5	0.71	46.00
/log-file-viewer.asp	5	0.00	2.25	4	0.56	(1.00)
/performance-tracking-log.asp	4	0.00	1.00	4	0.56	(2.00)
/track-website-visitors.asp	4	0.00	2.67	3	0.42	49.00
/web-site-usage-tracking.asp	4	50.00	7.00	2	0.28	(1.00)
/contacts.asp	3	33.33	2.33	3	0.42	(2.00)
/analyze-log-files.asp	2	0.00	1.00	2	0.28	(2.00)
/log-analyzer.asp	2	0.00	2.00	2	0.28	(1.00)
/modifications.asp	2	0.00	3.50	2	0.28	(1.00)
/track-website-stats.asp	2	0.00	1.00	2	0.28	(1.00)
/web-stats-software.asp	2	0.00	3.50	2	0.28	49.00
/windows-log-file-analysis.asp	2	0.00	1.00	2	0.28	(2.00)
/articles.asp	1	0.00	2.00	1	0.14	(1.00)
/download/edm.pdf	1	0.00	1.00	1	0.14	(1.00)
/edm.asp	1	0.00	1.00	1	0.14	(1.00)
/history.asp	1	0.00	2.00	1	0.14	0.00
/measuring-web-site-visitors.asp	1	0.00	1.00	1	0.14	0.00
/network-traffic-analyzer.asp	1	0.00	2.00	1	0.14	49.00
/privacy-policy.asp	1	0.00	7.00	1	0.14	49.00
/screenshots.asp	1	100.00	2.00	1	0.14	(1.00)
/site-traffic-stats.asp	1	0.00	5.00	1	0.14	0.00
/software.asp	1	0.00	1.00	1	0.14	0.00
/web-metrics.asp	1	0.00	1.00	1	0.14	(1.00)

F1 HELP F9 Toggle: ROI|Net Revenue:% users Right click on a line = Details F7 DISPLAY PAGES ALIASES

177 45 Head|Uwr|Block|Syn

Note that such results could be obtained in other reports where the first column displays reerring websites, search phrases or either or Google AdWords. When you press on F9 again you get the ROI, then the cost before to fall back on the percentage of the visitors who ask for the target page.

The demo version of Expert Data Miner doesn't allow you to get conversion rates based on cookies; this feature is available only in the Enterprise version.