# ANALYZING WEB LOGS

In HCI there is often an abundance of raw data, but making sense of it is hard. Computers can log every action. The issue is to find relevant information, which may be hidden like a needle in a haystack. Or even worse, it may be scattered all over the place in tiny tidbits with no apparent connection.

First, the server log was converted into a format that Theme can read. We used the path information of the retrieved files as event types in Theme. Every visiting IP-address was defined as a separate subject. We analyzed data for all subjects together.
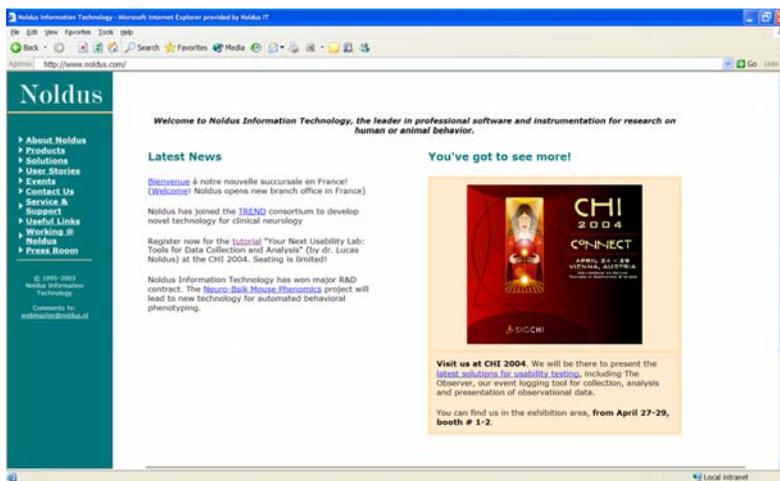
The result of the analysis was a set of patterns, reflecting the paths that visitors follow when surfing. One of the favorite patterns included the Company, Contact and Jobs sections in that order (figure 1). On average this took three minutes, ranging from 48 to 404 seconds. Another included Company, Resources (useful links), Applications (user stories) and Press (figure 2). In both cases the users followed the order of the menu on the left side of the main web page. Still other patterns included specialized sections, dealing with The Observer and psychology for instance. In between other sections may have been visited. The result of this analysis suggests that there are different types of visitors who are interested in different parts of the website.

In either case Theme can offer a solution. Theme can take both the order and timing of events into account, and can deal with variation in both. As an example a 24-hour server log was analyzed, realizing that a server log is far from a perfect observation. For reasons of convenience the Noldus website www.noldus.com was used. During the 24-hour logging period the website was visited more than 450 times. The visits were of very different nature. Some visitors retrieved only one page suggested by a search engine. Others browsed through multiple sections and spent quite some time on the site. We used Theme to search for similar navigation patterns among users.
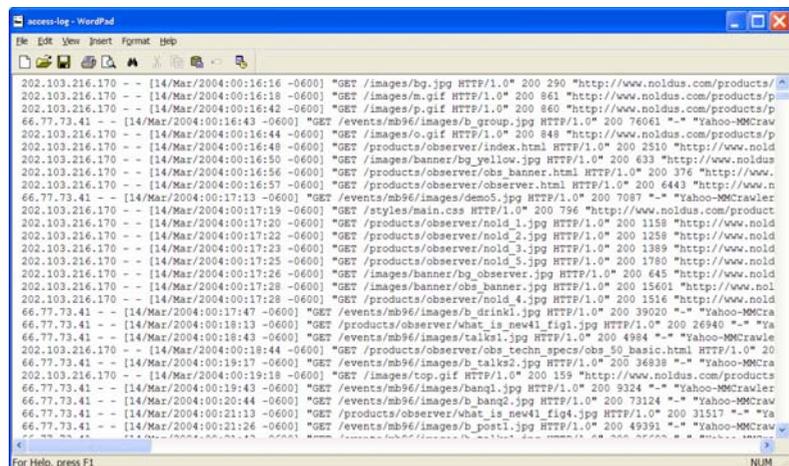
Figure 1 looks like the behavior of someone searching for a job, whereas fig. 2 looks like the behavior of a prospect who wants to know more about the company.

The use of Theme makes it possible to extract meaningful information from log files at the press of a button. No other method takes timing into account like Theme does, and most are seriously affected by variations such as other pages visited in between. Potential uses of this method are in the design and evaluation of web sites and applications.
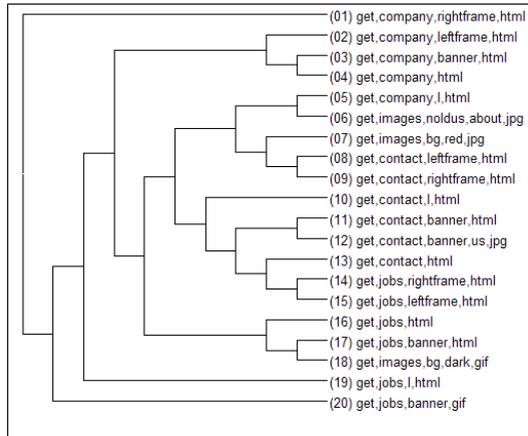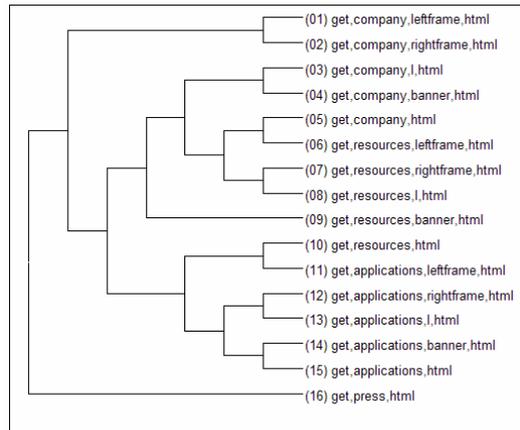

Figure 1. *'Job searcher' pattern.*

Figure 2. *'Prospect' pattern.*

How does Theme work?

Theme uses a unique algorithm that searches for relationships between events by taking into account both the order and relative timing of these events, and their hierarchical structure. Statistics such as the number of different patterns detected in a behavioral record, their length, abundance and number of actors involved can be used as objective measures of the level of organization or complexity of the behavior.
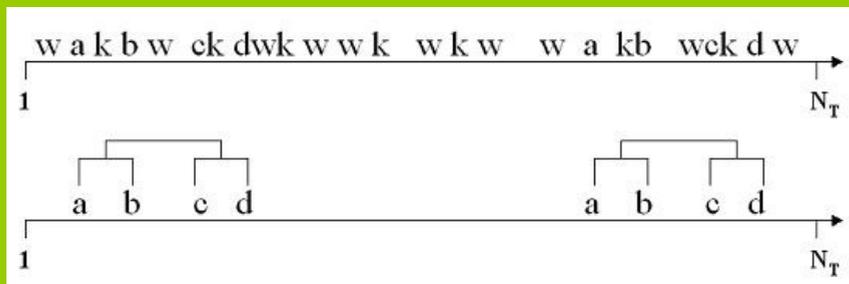

Figure 4. *Upper part: A record containing six types of events (a, b, c, d, k and w). Lower part: The same record after removing all occurrences of k and w. Two simple pattern (ab) and (cd) appear that were difficult to find when the other events were present. The patterns (ab) and (cd) are part of a larger pattern ((ab)(cd)) which may then become part of an even more complex pattern.*

Reference

Magnusson MS (2000), Discovering hidden time patterns in behavior: T-patterns and their detection. *Behavior, Research Methods, Instruments & Computers* **32**, 93-110.