Application of the 2-3 Agglomerative Hierarchical Classification on Web Usage Data.

Sergiu Chelcea, Brigitte Trousse

Application of the 2-3 Agglomerative Hierarchical Classification on Web usage data

Sergiu Chelcea and Brigitte Trousse

AxIS Research Team, INRIA[1] Sophia-Antipolis, France
BP 93, 06902 Sophia–Antipolis Cedex
FirstName.LastName@inria.fr

Abstract. In this paper we present a clustering of INRIA[1] Web sites' visited topics, using an hierarchical classification method, the 2-3 Hierarchical Ascending Classification proposed by Patrice Bertrand in [1]. The obtained clusters are then analyzed for their relevance.

Keywords and phrases: Clustering, Web usage data, Classification, Data Mining, 2-3 Hierarchies

1 Introduction

Nowadays the rapid and continuous growth of the Web can be a serious obstacle for users in their information search, despite the improvement of existing search engines. Personalizing the web space by analyzing users behaviour can reduce their quest for information and help them find more easily what they are looking for. For this, Data Mining techniques can be used for Web data analysis, in one of the three main *Web Mining* domains: *Web Content Mining*, *Web Structure Mining* and *Web Usage Mining* (WUM).

The Web usage data used during the WUM process are generally the users' navigational paths gathered in Web server logs, sometimes correlated with informations from the other Web Mining processes, e.g. the site structure. The users behaviour analysis helps in the Web site(s) re-conception process and also facilitates users information search (through dynamic inserted links).

In this paper we present a clustering of the INRIA Web sites' visited topics, using an hierarchical classification method, the 2-3 Hierarchical Ascending Classification proposed by Patrice Bertrand in [1]. The 2-3 Agglomerative Hierarchical Clustering (2-3 AHC) [1], generalizes the Agglomerative Hierarchical Clustering (AHC) by giving each cluster the possibility of intersecting at most another cluster, when the obtained intersection is distinct from the two clusters. This characteristic allows the resulting clusters structure to highlight groups of objects (clusters) that have the common characteristics of two other groups (which is not possible with the classical AHC).

Using our new 2-3 AHC algorithm [3] with the same $\Theta(n^2 \log n)$ algorithmic complexity as the classical AHC, we have analyzed INRIA's Web sites' visitors activities based on their behaviour.

[1] The French National Institute for Research in Computer Science and Control

2 2-3 Agglomerative Hierarchical Classification

In the first subsection we present some hierarchical classification notions along with the classical AHC algorithm. The second subsection presents the 2-3 AHC concept introduced in [1] followed by a short description of our 2-3 AHC algorithm [3], used later along with the classical AHC algorithm to classify the Web usage data (Section 3.2).

2.1 Agglomerative Hierarchical Classification

Clustering or the *unsupervised classification* is a Data Mining technique used to group together similar object into *classes* also known as *clusters*. Among the well known clustering techniques one can find: the neural networks, the hierarchical classification methods, the fuzzy networks, the decision trees, etc. Each of these clustering techniques generates a cluster set organized by its specific structure (partitions, hierarchies, pyramids, etc.).

In the *agglomerative hierarchical methods*, starting form the initial elements (the *singletons*), the clusters are successively merged into higher level clusters, until the entire set of analyzed objects becomes a cluster. These resulting hierarchical structures (hierarchies, 2-3 hierarchies, pyramids) can then be easily visualized using a graphic called *dendogram*.

In order to present the 2-3 Agglomerative Hierarchical Classification [1] and our 2-3 AHC algorithm [3], we first remind the classical AHC principle.

Sokal and Sneath proposed in 1963 [6] the first version of the classical AHC algorithm, consisting in two phases: initialization and merging.

During the first phase, the singletons distance (dissimilarity) matrix is computed using a distance (dissimilarity) measure (e.g. the *Euclidean distance*), while the singletons represent the initial set of clusters. In the second phase, successive mergings are performed between the two *closest clusters*, until the initial objects are all merged into a final cluster. The two clusters are closest in the sense of a chosen *aggregation link*, denoted μ and simply called *link* (i.e. *single link*, *complete link*, *average link*). Together with this criteria, others like the cardinality and the lexicographical order can be used to determine the closest clusters.

The link $\mu(X, Y)$ between two merged clusters X and Y, represents the heterogeneity degree of the resulting cluster $X \cup Y$, and is denoted $f(X \cup Y)$. In the first phase of the AHC algorithm, the f values of the singletons are set to 0. At the end of the second phase, the resulting structure is a sequence of nested partitions which can be visualized using a *dendogram* (graphic based on the heterogeneity degree f of the clusters).

A small example of a classical hierarchy is presented in Figure 1. In order to chose a partition from the resulting hierarchy, the differences between the f level of the created clusters are analyzed. Usually the partitioning level is chosen whenever there's a "*big*" difference between the created cluster's levels. For example in Figure 2, the first partition $\{\{a, b, c\}, \{d, e, f\}, \{g\}\}$ generated

by the level P_1, is more appropriate than the second one $\{\{a,b,c\},\{d,e,f,g\}\}$ generated by P_2, since the clusters are more homogeneous.

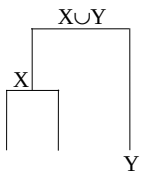

Fig. 1. Classical hierarchy

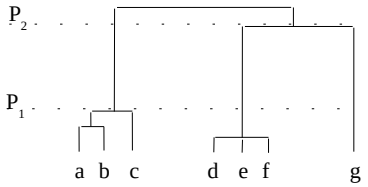

Fig. 2. Hierarchy partitioning

Whenever a new cluster $X \cup Y$ is created by merging two other clusters X and Y, we say that $X \cup Y$ is *predecessor* for X and Y, while X and Y are *successors* of $X \cup Y$ (cf. Figure 1). The clusters found on the same f level as their predecessor are non-relevant clusters from the viewpoint of cluster analysis, and can be eliminated from the hierarchy after its creation. This final elimination step is called the *refinement step* and will produce a *strictly indexed hierarchy*, making the corresponding dendogram easier to visualize.

The main characteristic of the classical AHC algorithm is that after each merging, only the resulting cluster is kept for future mergings. Thus the resulting structure (the *hierarchy*) will contain only nested or disjoint clusters, also denoted as *hierarchical* clusters.

A hierarchy will induce a new distance matrix (*ultra-metric*) over the initial elements based on the distances at which they were first regrouped in a cluster. This matrix can be then compared with the initial matrix or with other methods induced matrices, for quality analysis [3] using different indices (e.g *Stress* [5] formula).

2.2 2-3 Hierarchies and 2-3 AHC Algorithm

We present here the 2-3 *hierarchy* concept along with the 2-3 *Agglomerative Hierarchical Classification* method introduced in [1] in order to generalize and to make more flexible the classical AHC.

As we saw before, the AHC generates disjoint clusters or clusters included one in the other. The *2-3 Agglomerative Hierarchical Classification* method proposed in [1] gives each cluster the possibility of intersecting at most another cluster, when the obtained intersection is distinct from the two clusters. This characteristic allows the obtained cluster structure to highlight groups of objects having the common characteristics of two other groups (not possible with the AHC).

The resulting cluster structure is called an 2-3 *hierarchy*, term justified by the following property equivalent with the aforementioned characteristic:

Given any three clusters, at least two out of the three possible formed clusters pairs are hierarchical (nested or disjoint).

The term *2-3 hierarchy* specifies how the set of 2-3 hierarchies is an extension of the hierarchies set - indeed, from the definition above it clearly results that a hierarchy is a particular case of 2-3 hierarchy. This happens when all three possible cluster pairs are hierarchical, thus leading to a classical hierarchy.

A 2-3 hierarchy example is presented in Figure 3 bellow. When two clusters are not hierarchical, we say that they *properly intersect* themselves: X *properly intersects* $Y \Leftrightarrow X \cap Y \notin \{X, Y, \emptyset\}$. For instance on the example from Figure 3, the cluster $\{ba\}$ *properly intersects* $\{ac\}$, while the clusters $\{ba\}$ and $\{bac\}$ are hierarchical.

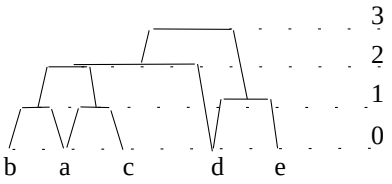

Fig. 3. Example of an 2-3 Hierarchy

Since the 2-3 hierarchies allow clusters to properly intersect themselves, their structures are richer compared with the classical hierarchies obtained on same data sets [3]. For example the maximal number of created clusters by the classical AHC is $(n-1)$, compared with $[\frac{3}{2}(n-1)]$ for the 2-3 AHC [1], where n is the initial number of elements. Comparative tests [3] have showed an 11% (with *single-link*) to 45% (with *complete-link*) increase in the created clusters number for our 2-3 AHC algorithm. Figure 4 shows a small example of created hierarchy and 2-3 hierarchy on a three points dataset, using the *single-link*.

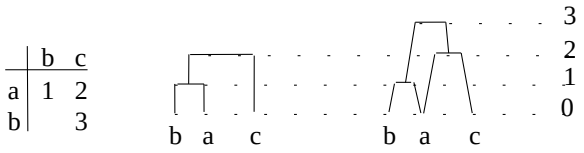

Fig. 4. AHC and 2-3 AHC

During our experimentations on the Web usage data presented in the next section, we have used our simplified 2-3 AHC algorithm proposed in [2]. Compared with the initial 2-3 AHC algorithm [1], our 2-3 AHC algorithm's advantages has a smaller complexity ($\mathcal{O}(n^2 \log n)$ instead of the initial $\mathcal{O}(n^3)$), and also a simplified principle which is similar to the classical AHC algorithm.

3 Applying 2-3 AHC on Web usage data

3.1 Motivations

Nowadays the rapid and continuous growth of the Web can be a serious obstacle for users in their information search, despite the improvements of existing search engines. Personalizing the web space by analyzing users behaviour can reduce their quest for information and help them find more easily what they are looking for. For this, Data Mining techniques can be used for Web data analysis, in one of the three main *Web Mining* domains: *Web Content Mining*, *Web Structure Mining* and *Web Usage Mining* (WUM).

The Web usage data used during the WUM process are generally the users' navigational paths gathered in Web server logs, sometimes correlated with information from the other Web Mining processes, e.g. the site structure. The users behaviour analysis helps in the Web site(s) re-conception process and also facilitates users information search (through dynamic inserted links).

Users searching for information on INRIA's Web site, are transparently browsing through the interconnected pages of INRIA's Web servers. To trace their behaviour, we have analyzed the log access files from two INRIA Web servers: the national Web server (*http://www.inria.fr*) and also the Sophia Antipolis research unit Web server (*http://www-sop.inria.fr*).

Since INRIA's scientific organization into research teams has recently changed (1^{st} April 2004), we have chosen to study the Web users behaviour on two different 15 days time periods:

- from 01 until 15 January 2003, period denoted in the followings as Per_1,
- and from 27 May until 10 June 2004, period denoted as Per_2.

Indeed, the main motivation of our study here was to analyze the impact of the changes in the Web site structure (see Appendix A), on users behaviour when searching for information. More particularly, our study concerned the clustering of INRIA Web sites' visited topics, using the 2-3 Hierarchical Ascending Classification presented in the previous section and was done in two phases:

- first, the *preprocessing* of the Web access logs (based on the work of Tanasa et al. [7]) and presented in section 3.2,
- secondly, the data mining (using our 2-3 AHC algorithm) and the result analysis phase (section 4).

3.2 Data preprocessing

In this subsection, we will shortly explain the data preprocessing methodology proposed by Tanasa et al. (see. [7] for more details). We used the AxIS LogMiner[1] tool developed within the AxIS research team at INRIA Sophia Antipolis.

The aim of the preprocessing phase was to identify and extract *user navigations* (sequences of user actions) from the raw Web logs, and was done in four steps: *data fusion*, *data cleaning*, *data structuration* and *data summarization*.

[1] Tool description available at: http://www-sop.inria.fr/axis/axislogminer

Classical data preprocessing :

During the *data fusion* step, the Web logs files were joined together for each analyzed period (resulting in log L_1 for Per_1 and log L_2 for Per_2), in order to reconstruct the cross-server users' navigations.

Thus the two joined logs contained all the requests (chronologically sorted) made by different users for different *resources* on the two Web servers, over the given periods of time. Some of these requests were made for non-relevant resources from our analysis viewpoint and were eliminated in the *data cleaning* step. For example, we do not interest ourselves in requests for images, since they usually are implicit requests (images contained in the accessed page [7]). Also, the requests made by web crawlers (robots) have been eliminated from the web logs for obvious reasons. Filtering out all these requests has reduced L_1 to 11% and L_2 to 15%, from their original size. For example for L_2, the number of requests was reduced from 4.473.228 to 686.084, equivalent to a log file size reduction from 901Mb to 135Mb.

Next, the *data structuration* step grouped the unstructured log files requests by *user*, *user session*, *page view*, and *visit* (*navigations*). Since log files contain only the computer IP and (sometimes) the user agent, we've considered as a *user*, the couple (IP, [User Agent]) and as a *user session* all the actions performed by the user over the analyzed period.

Then the users *navigations* were obtained by splitting every user session using a 30 minutes threshold: 173.015 navigations for L_1 and 145.454 navigations for L_2.

Finally, the obtained log files were stored in a relational database in the *data summarization* step.

Advanced data preprocessing :

We performed a general *data selection* step in which we selected from the relational DB the navigations (users visits) to analyze later, using the following criteria:

- navigation duration > 60 seconds,
- number of requests in the navigation > 10,
- browsing speed (duration/number of requests) > 4.

This has reduced the number of analyzed navigation to 9625 for L_1 and to 9309 for L_2.

Next, depending on the analysis, we performed secondary data selections. For example, in the secondary data selection associated with our first analysis (section 4) we decided to keep only the visits on both INRIA's servers and to cluster the visited first level topics (from the visited URLs). Thus the number of analyzed navigations was reduced to 3905 for L_1 and to 3513 for L_2. Also, the Web pages returned by the Web server with an error status code (≥ 400) were ignored in our analysis. We have found a total of 190 visited topics for Per_1 (78 were research teams). For Per_2 we found 210 topics from which 86 were research teams (49 actual research teams and 37 old research teams from Per_1).

Next, we performed a *data generalization* step, in which the visited URLs were assigned to different research teams for later clustering. Since INRIA research teams organization has changed (starting from 1^{st} of April 2004), its Web site structure changed accordingly. The research teams were reorganized from the four existing research themes, into five new research themes (Appendix A).

We decided to analyze the impact of the Web site structure on users navigations before and after this change (during Per_1 and Per_2). This was done by performing two different analyses of users visits: one on INRIA's first level topics and another on INRIA's research teams. For the second analysis, since a user visit is actually a set of visited URLs, we needed to determine which URLs belong to different research teams.

Each URL can have several topics associated with different semantic topics:

$$http://\underbrace{www-sop.inria.fr}_{Site}/\underbrace{axis}_{topic1}/\underbrace{personnel}_{topic2}/\underbrace{Doru.Tanasa}_{topic3}/doru{-}eng.html$$

Thus in a first step, a URL was assigned to a research team when one of its topics was the research team itself. After this, complementary information on INRIA's Web site was used to assign URLs to research teams. For example, the URL: `http://www.inria.fr/recherche/equipes/axis.en.html` does not contain any research team topics, but is the AxIS research team presentation page from INRIA's main server.

After the data generalization step and in order to cluster the obtained topics, we needed to *compute the dissimilarity matrix* used as input for the AHC and our 2-3 AHC algorithms. For this, we used the Jaccard similarity index on the visited topics as defined in [4]. As in [4], we represented each navigation by a binary vector of the visited topics: the position i in the vector is 0 if $topic_i$ was not visited and 1 if $topic_i$ was visited during the navigation. Based on these vectors and aiming to define a similarity/dissimilarity between two topics R_i and R_j, we define the four following quantities:

- a as the number of counts when $R_i^k = R_j^k = 1$,
- b as the number of counts when $R_i^k = 0$ and $R_j^k = 1$,
- c as the number of counts when $R_i^k = 1$ and $R_j^k = 0$,
- d as the number of counts when $R_i^k = 0$ and $R_j^k = 0$.

Then the similarity between two topics R_i and R_j is computed using: $S(R_i, R_j) = \frac{a}{a+b+c}$, which represents the probability of visiting both topics when at least one of them is visited. The dissimilarity matrix used as input for the classical AHC and our 2-3 AHC, was computed using the dissimilarity: $\mu(R_i, R_j) = 1 - S(R_i, R_j)$.

4 Results

For our first analysis, we have focused on the research teams distribution in the server-crossed visited topics. We have selected from Per_1 all server-crossed navi-

robotvis SOP **3B**, robotvis **3B**, epidaure SOP **3B**, odyssee SOP **3B**, epidaure **3B**, ariana SOP **3B**, ariana **3B**	comore SOP **4A**, icare SOP **4A**, icare **4A**, miaou SOP **4A**, reves SOP **3B**, miaou **4A**, chir SOP **4A**, comore **4A**, caiman SOP **4B**	orion SOP **3A**, axis SOP **3A**, orion **3A**
prisme SOP **2B**, prisme **2B**	koala SOP **2A**, koala **2A**, croap SOP **2A**, croap **2A**	odyssee **3B**, dream SOP **3A**, lemme **2A**, opale SOP **4B**, opale **4B**, certilab **2A**, pastis **3B**
orion SOP **3A**, acacia SOP **3A**, acacia **3A**, axis SOP **3A**, orion **3A**, aid SOP **3A**, aid **3A**	coprin SOP **2B**, saga SOP **2B**, saga **2B**	sinus SOP **4B**, sinus **4B**, smash SOP **4B**
robotvis SOP **3B**, robotvis **3B**, odyssee SOP **3B**	mimosa SOP **1C**, mimosa **1C**, tick SOP **1C**, tick **1C**	sloop SOP **1A**, sloop **1A**, oasis SOP **2A**, oasis **2A**
rodeo SOP **1B**, rodeo **1B**, planete SOP **1B**, planete **1B**	lemme SOP **2A**, tropics SOP **1A**, mascotte SOP **1B**, omega SOP **4B**, galaad SOP **2B**, cafe SOP **2B**, certilab SOP **2A**	mistral SOP **1B**, mistral **1B**
mefisto SOP **4B**, mefisto **4B**	mascotte SOP **1B**, mascotte **1B**	safir SOP **2B**, safir **2B**
meije SOP **1C**, meije **1C**		

Table 1. INRIA's Web site topics clustering using 2-3 AHC for Per_1

gations (visiting both Web sites: main and Sophia's), and then clustered all obtained visited topics using our 2-3 AHC algorithm. Table 1 presents the repartition of only the research team topics in the obtained clusters (the other topics are not presented here). Also, we did not represent, the one element clusters (the "outliers"): `caiman 4B, saga 2B, meije SOP 1C, sysdys SOP 4B, chir 4A, cafe 2B, codes 2B, visa SOP 4A, tropics 1A, omega 4B`. We added after the name of each research team, their theme and sub-theme, as well as their site (empty for the main site and "SOP" for Sophia's site).

As we can see the research teams distribution usually corresponds to their theme membership (16 out of the 19 non-trivial clusters contain research teams from the same theme). Also, old research teams that have been replaced by new research teams, are in the same clusters as the corresponding new ones, since their pages are strongly interconnected. For example: *aid* was replaced by *axis* (cluster 7), *rodeo* by *planete* (cluster 13), etc.

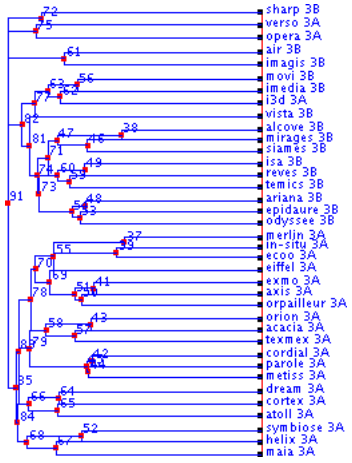

Fig. 5. 2-3 Hierarchy on theme 3 projects during Per_1

For the second analysis we have selected from the obtained navigations in section 3.2, only those visiting at least one page on INRIA's main server. From these navigations, we choose to analyze the visited topics (research teams) only from the main server pages and to cluster only the research teams topics for theme 3 (Per_1 and Per_2) and for theme Cog (Per_2).

Thus, we have first selected only those navigations containing at least one visit of the theme 3 pages on INRIA's main server during Per_1. From these navigations, we have clustered the visited topics on the main server only, corresponding to research teams from theme 3 (Figure 5). In the resulting classification we can distinguish two main clusters (82 and 85) that group almost all elements from the two sub-themes of theme 3 (see Appendix A).

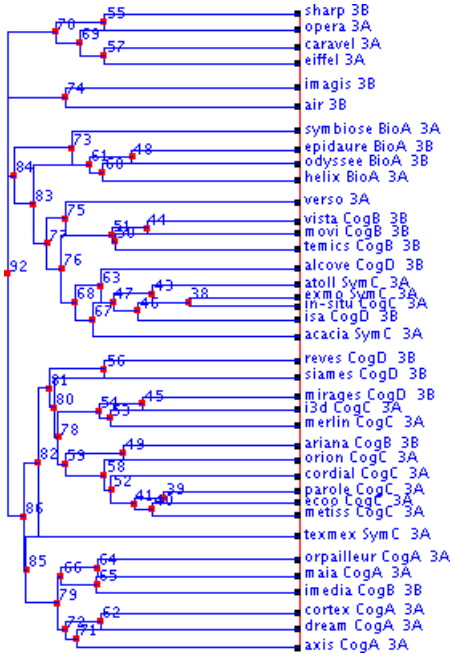

Fig. 6. 2-3 hierarchy on theme 3 projects during Per_2

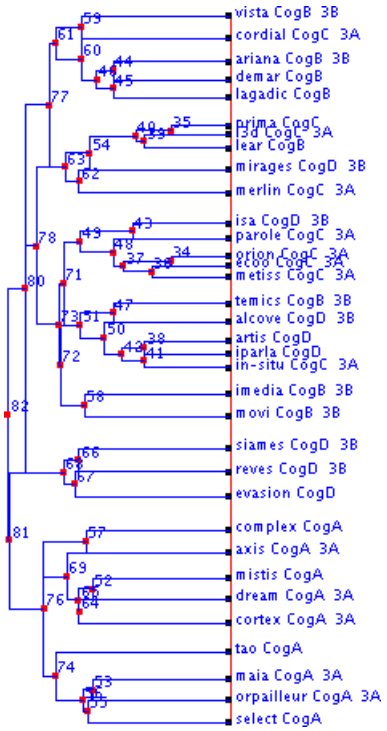

Fig. 7. 2-3 hierarchy on theme Cog projects during Per_2

Next, we have selected for the second period the navigations visiting at least one of the new themes pages on the INRIA's main server. From these navigations we have focused again on the topics corresponding to research teams from the "old" theme 3, only on the main server's pages (cf. Figure 6). We refer in this classification to the new theme Bio, which groups together in cluster 73 the four research teams that were separated in the previous organization (Figure 5). Some

of the research teams have not been assigned to one of the new five themes since they were either replaced or stopped, but their Web pages are still accessible on the Internet (i.e. *sharp*, *opera*, *verso*).

Finally, we have selected only the navigations that had at least one visit of the new theme Cog pages during Per_2. We have then clustered only the topics on the main servers pages corresponding to a research team from theme Cog (cf. Figure 7). We note that users have the tendency to visit all CogA research teams, while for the others sub-themes there is a certain variability and a deeper analysis is needed (possible causes: events like conferences or seminaries presented on INRIA's Web site that affects users visits, the recent change of structure, etc.). Also we found that our research team, *AxIS*, is grouped in the two time periods with different research teams.

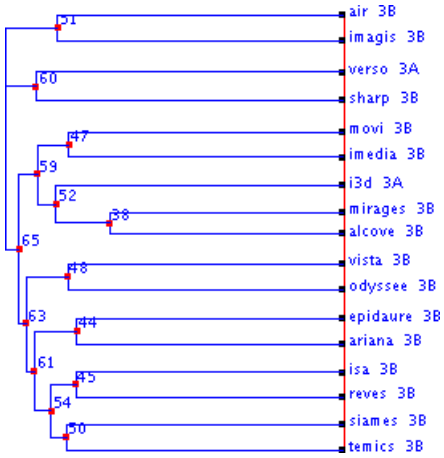

Fig. 8. Classical hierarchy on theme Cog projects

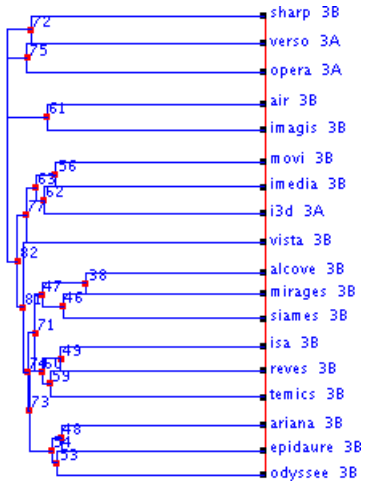

Fig. 9. 2-3 hierarchy on theme Cog projects

In our final analysis we have compared the classical AHC method and our 2-3 AHC algorithm, by clustering the research teams from theme Cog (during Per_2). The data selection was the same as in the previous analysis: navigations visiting at least one of the theme Cog pages, topics only from the main server's pages and representing research teams from theme Cog.

Figures 8 and 9 present a partial output (containing all 3B research teams) of the classical AHC respectively of our 2-3 AHC algorithm. The 2-3 hierarchy obtained contains more created clusters than the classical hierarchy (22 against 15), and thus more information. For example, analyzing cluster 54 in Figure 9, we can say that research teams *ariana*, *epidaure* and *odyssee* have a "stronger" probability of being visited together, compared with the one given by the classical hierarchy (Figure 8).

5 Conclusions and future work

This paper presents the first application of a 2-3 AHC algorithm on Web usage data, and shows the potentials of our algorithm compared with the classical AHC algorithm. In our study, we interested ourselves in clustering the visited topics of INRIA's Web site, which recently (1^{st} April 2004) changed its structure.

We have studied the impact of INRIA's Web site structure on users navigations, during two time periods (before and right after the site structure change). Although the second analyzed period was shortly after the change, we have found that usually users navigations are influenced by the Web site structure.

Our ongoing and future work concern the following topics:

- deeper analysis of the comparison between 2-3 AHC and AHC on same Web usage data, using different aggregation links,
- a better dissimilarity measure. For example the generalized Jaccard index, which takes into account the number of visited pages for a topic, not just its presence (count vs. binary),
- application of our 2-3 AHC algorithm on other data inferred from the activities reports of INRIA's research teams, and comparison of the results with the ones obtained here.

Acknowledgements : We would like to thank Mihai Jurca for his help on the preprocessing phase and Sophie Honnorat for her help with the english corrections.

A INRIA research teams organization

Before 1^{st} of April 2004, INRIA's research teams were organized in four different research themes, namely:

- Theme **1**: Networks and systems:
  - **A** : Architectures and Systems,
  - **B** : Networks and Telecommunications,
  - **C** : Distributed and Real-Time Programming.
- Theme **2**: Software engineering and symbolic computing:
  - **A** : Semantics and Programming,
  - **B** : Algorithms and Computational Algebra.
- Theme **3**: Human-computer interaction, images processing, data management, knowledge systems:
  - **A** : Databases, Knowledge Bases, Cognitive Systems,
  - **B** : Vision, Image Analysis and Synthesis.
- Theme **4**: Simulation and optimization of complex systems:
  - **A** : Control, Robotics, Signal,
  - **B** : Modelling and Scientific Computing.

After this date, the research teams were reorganized in the following five research themes:

- Theme **Com**: Communicating systems:
  - **A** : Distributed systems and software architecture,
  - **B** : Networks and telecoms,
  - **C** : Embedded systems and mobility,
  - **D** : Architecture and compiling.
- Theme **Cog**: Cognitive systems:
  - **A** : Statistical modeling and machine learning,
  - **B** : Perception, indexing and communication for images and video,
  - **C** : Multimedia data: interpretation and man-machine interaction,
  - **D** : Image synthesis and virtual reality.
- Theme **Sym**: Symbolic systems:
  - **A** : Reliability and safety of software,
  - **B** : Algebraic and geometric structures, algorithms,
  - **C** : Management and processing of language and data.
- Theme **Num**: Numerical systems:
  - **A** : Control and complex systems,
  - **B** : Grids and high-performance computing,
  - **C** : Optimization and inverse problems for stochastic or large-scale systems,
  - **D** : Modeling, simulation and numerical analysis.
- Theme **Bio**: Biological systems:
  - **A** : Modeling and simulation in biology and medicine.

References

1. P. Bertrand. Set systems for which each set properly intersects at most one other set - Application to Cluster Analysis. Research Report Ceremade 0202, Universite Paris-9, France, 2002.
2. S. Chelcea, P. Bertrand, and B. Trousse. Agglomerative 2-3 Hierarchical Clustering: theoretical improvements and tests. In *27th Annual Conference of the Gesellschaft fur Klassifikation*, Cottbus, Germany, 12 - 14 mars 2003.
3. S. Chelcea, P. Bertrand, and B. Trousse. Un Nouvel Algorithme de Classification Ascendante 2-3 Hiérarchique. In *Reconnaissance des Formes et d'Intelligence Artificielle (RFIA 2004)*, Centre de Congrès Pierre BAUDIS, Toulouse, 28-30 Janvier 2004.
4. A. El Golli, B. Conan-Guez, F. Rossi, D. Tanasa, B. Trousse, and Y. Lechevallier. Les cartes topologiques auto-organisatrices pour l'analyse des fichiers logs. In *11èmes Rencontre de la Société Francophone de Classification*, Bordeaux, 8-10 septembre, 2004. to appear.
5. A.R. JOHNSON and D.W. WICHERN. *Applied Multivariate Statistical Analysis*, chapter 12. Prentince Hall, 1982.
6. R.R. Sokal and P.H.A. Sneath. Principles of numerical taxonomy. Freeman, San Francisco, 1963.
7. D. Tanasa and B. Trousse. Advanced data preprocessing for intersites web usage mining. *IEEE Intelligent Systems*, 19(2):59–65, March-April 2004.