



HAL
open science

Classification 2-3 hiérarchique de données du Web

Sergiu Theodor Chelcea, Brigitte Trousse

► **To cite this version:**

Sergiu Theodor Chelcea, Brigitte Trousse. Classification 2-3 hiérarchique de données du Web. Actes des 5ème journées Extraction et Gestion des Connaissances (EGC 2005), Revue des Nouvelles Technologies de l'Information (RNTI-E-3), Jan 2005, Paris, pp.219. inria-00000864

HAL Id: inria-00000864

<https://inria.hal.science/inria-00000864>

Submitted on 28 Nov 2005

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Classification 2-3 hiérarchique de données du Web

Sergiu Chelcea et Brigitte Trousse

Projet AxIS, INRIA Sophia Antipolis,
B.P. 93, 06902 Sophia Antipolis Cedex
http://www-sop.inria.fr/axis

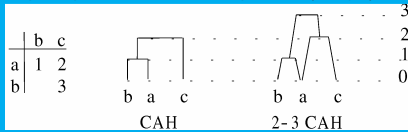


Motivation :

Classification des rubriques des URLs visitées du site Web de l'INRIA (équipes de recherche en particulier) à partir des navigations des internautes en vue d'étudier l'impact de la structure du site Web et de la structure organisationnelle de l'INRIA sur les comportements des internautes.

Méthode utilisée : 2-3 CAH

La Classification Ascendante 2-3 Hiérarchique génère une structure plus riche que la CAH classique et présente la même complexité $\Theta(n^2 \log n)$.



Références : - P. Bertrand (RR 0202, Université Paris IX, 2002)

- S. Chelcea, P. Bertrand, B. Trousse (SYNASC 2002, GFKL 2003, RFIA 2004)

Données du Web :

- deux fichiers Log issus de serveurs Web de l'INRIA (siège et l'UR de Sophia Antipolis) ;
- deux périodes de 15 jours concernées : avant et après le changement de l'organisation scientifique en thèmes de recherche de l'INRIA en avril 2004 (thèmes 1, 2, 3 et 4 \rightarrow thèmes Com, Cog, Sym, Num et Bio).

Prétraitement des données :

- prétraitement classique (fusion, nettoyage et structuration des données) ;
- prétraitement avancé :
 - sélection des navigations pertinentes (durée de navigation > 60 seconds, nombre de requêtes > 10, vitesse de navigation > 4) ;
 - prétraitement spécifique à chacune des trois analyses réalisées.
- utilisation d'indice de Jaccard sur les navigations des internautes.

Première analyse :

Données : rubriques syntaxiques de premier niveau dans les navigations passant par les deux sites à la fois, pendant la première période.

Objectif : Etude de l'impact global du site Web sur les navigations des internautes

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, chr SOP 4A, comore 4A, calman SOP 4B	orion SOP 3A, axis SOP 3A, orion 3A
prisme SOP 2B, prisme 2B	koala SOP 2A, kosa 2A, croap SOP 2A, croap 2A	odyssee 3B, dream SOP 3A, lemme 2A, opale SOP 4B, opale 4B, certilab 2A, patric 3B
orion SOP 3A, acacia SOP 3A, acacia 3A, osie SOP 3A, orion 3A, aid SOP 3A, aid 3A	coprin SOP 2B, saga SOP 2B, coga 2B	simon SOP 4B, simus 4B, croush 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, oasie SOP 2A, oasie 2A
rodos SOP 1B, rodos 1B, planete SOP 1B, planete 1B	homme SOP 2A, tropics SOP 1A, mascotte SOP 1B, omega SOP 4B, galad SOP 2B, cafe SOP 2B, certilab SOP 2A	mistral SOP 1B, mistral 1B
mefisto SOP 4B, mefisto 4B, mejo SOP 1C, mejo 1C	mascotte SOP 1B, mascotte 1B	safir SOP 2B, safir 2B

Répartition des équipes de recherche dans la classification obtenue.

Résultat :

- 190 rubriques visitées et 78 équipes de recherche ;
- 15/19 classes contiennent des équipes de recherche du même thème.

Deuxième analyse :

Données : - sélection des pages du serveur siège et des navigations passant au moins par une de ces pages pour les deux périodes ;

- classification des équipes de recherche en utilisant les URLs visitées \Rightarrow utilisation de connaissances complémentaires sur le site de l'INRIA pour l'attribution des URLs aux équipes de recherche (rubriques sémantiques).

Objectif : - analyse des classifications des équipes de recherche appartenant à l'ancien thème 3 de recherche pendant les deux périodes (cf. Figure 1 et 2) et du thème Cog pendant la deuxième période (cf. Figure 3) pour comparer les classes obtenues à l'organisation scientifique en thèmes de recherche.

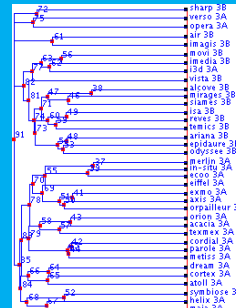


Figure 1 : Equipes du thème 3 avant

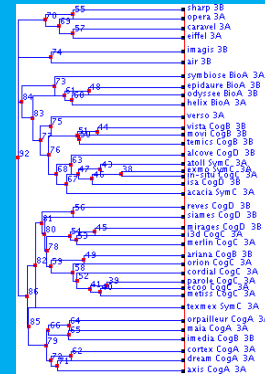
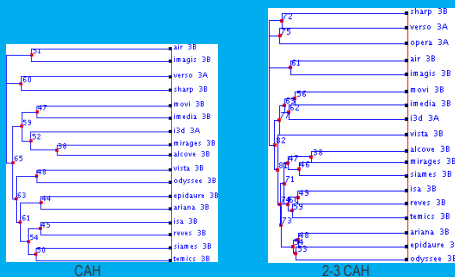


Figure 2 : Equipes du thème 3 après

Troisième analyse :

Données : cf. données de la deuxième analyse.

Objectif : comparaison des classifications d'une partie des équipes de recherche du thème 3 (sub-thème 3A) en utilisant la 2-3 CAH et la CAH classique.



Résultat : 15 contre 22 classes créées \Rightarrow plus d'information.

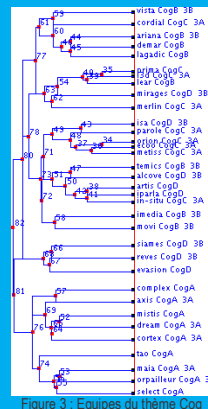


Figure 3 : Equipes du thème Cog

Résultat : les équipes de recherche sont classées conformément à l'organisation scientifique en thèmes dans les périodes analysées.

La classification des équipes de recherche du thème 3 pendant la première période correspond à l'organisation scientifique (cf. Fig. 1), et elle change dans la nouvelle organisation pendant la deuxième période (cf. Fig. 2).

La classification des équipes de recherche du thème Cog (deuxième période) est relativement proche de l'organisation scientifique (cf. Fig. 3).

Conclusions :

- première application de la 2-3 CAH sur des données du Web ;
- navigations des internautes influencées par la structure organisationnelle de l'INRIA présente au niveau de la structure de son site Web.

Travaux futurs :

Expérimentations avec d'autres liens d'agrégation, d'autres indices pour la classification et sur d'autres données du Web.