Classification 2-3 Hiérarchique de données du Web

Sergiu Chelcea, Brigitte Trousse

Projet AxIS, INRIA Sophia-Antipolis,
B.P. 93, 06902 Sophia Antipolis Cedex, France
Prénom.Nom@sophia.inria.fr

Dans ce papier nous présentons une classification des rubriques des URLs visitées du site Web de l'INRIA (équipes de recherche en particulier), en vue d'étudier l'impact de la structure du site Web et de la structure organisationnelle de l'INRIA sur les navigations des internautes. Pour cela nous avons utilisé notre algorithme de Classification Ascendante 2-3 Hiérarchique (Chelcea et al. 2004) qui génère une structure plus riche que la CAH classique (cf. Figure 1), en présentant une complexité identique soit $\Theta(n^2 \log n)$.

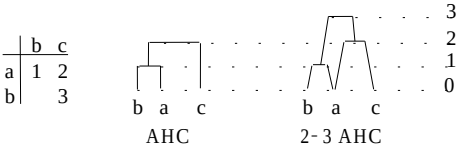

FIG. 1 – Exemple de CAH et de 2-3 CAH sur un petit jeu de données.

Nous avons réalisé trois analyses sur les fichiers log provenant de deux serveurs Web de l'INRIA, sur deux périodes de 15 jours : avant et après le changement de l'organisation scientifique de l'INRIA en avril 2004 (cf. http ://www.inria.fr/recherche/). Pour classer les rubriques des URLs (dont certaines représentent des équipes de recherche) nous avons utilisé l'indice de Jaccard sur les navigations (ensembles des URLs) des internautes. Tout d'abord nous avons analysé toutes les rubriques visitées de premier niveau ce qui a révélé l'impact *global* de la structure du site Web sur les navigations des utilisateurs (16 classes parmi les 19 contenaient des équipes de recherche du même thème scientifique (Chelcea et Trousse 2004)). Ensuite, nous avons comparé la classification des équipes de recherche de l'ancien Thème 3 (première période) et du nouveau Thème Cog (deuxième période) où apparaît le projet AxIS : nous avons pu montrer l'impact de la *nouvelle structuration en thèmes*. Finalement, nous avons sur un thème donné comparé le résultat de notre algorithme de 2-3 CAH avec celui de la CAH classique : nous avons pu mettre en évidence des groupes plus homogènes de projets. Nos travaux courants et futurs concernent l'utilisation des autres indices de dissimilarité (e.g. Jaccard généralisé) et la prise en compte du champ "referer" dans nos analyses.

Références

Chelcea S., Bertrand P., and Trousse B. (2004), Un Nouvel Algorithme de Classification Ascendante 2-3 Hiérarchique. In RFIA 2004, volume 3, Centre de Congrès Pierre BAUDIS, Toulouse, France, pages 1471-1480, 28-30 Janvier 2004.

Chelcea S. and Trousse B. (2004), Application of the 2-3 Agglomerative Hierarchical Classification on Web usage data. In Petcu D. et al, editors, SYNASC 2004, Romania, pages 107-118, September 2004. Mirton Publisher, ISBN 973-661-441-7.