Le Web Usage Mining, méthodes et expérimentation pour l'extraction d'information et de connaissance sur les données du Web.

Christine Michel

Caractérisation d'usages et personnalisation d'un portail pédagogique. État de l'art et expérimentation de différentes méthodes d'analyse du Web Usage Mining.

Christine MICHEL

Maître de conférence

Centre d'Étude des Médias
MSHA - Université Bx3
10, Esplanade des Antilles
33607 PESSAC Cedex
Tel : 05 56 84 68 13
E-Mail Christine.Michel@montaigne.u-bordeaux.fr

✍ **Résumé**

La caractérisation et la personnalisation des systèmes d'accès à l'information se font généralement par des techniques de Web Data Mining. Dans cet article nous en présentons trois volets principaux : le processus de pré-traitement sur les données, le choix et l'application de plusieurs méthodes analytiques et l'interprétation des résultats d'analyse. Dans la première partie de cet article nous présentons le processus de pré-traitement ainsi qu'un aperçu des différentes méthodes statistiques. Dans la seconde partie nous présentons une recherche expérimentale en cours consistant à rechercher des profils d'usage ou d'usagers pour la consultation d'un site Web Pédagogique. A terme, l'objectif est de pouvoir personnaliser l'information présentée selon le type de l'usager. Le logiciel utilisé est *Clémentine*© (société SPSS).

Mots clefs :
Caractérisation, usages, personnalisation, Web Mining, méthodes d'analyse

✍ **Abstract**

Information access systems like Web Sites or Intranet are generally characterized and personalized by using Web Data Mining techniques. The evaluation process is decomposed into 3 steps : first circulation data (stored in log files) are pre-treated in order to reduce noise, then, in function of the aim of the study, analytical methods are choice and applied, and finally results are interpreted. In the first part of this paper, we describe different pre-treatments and present an overview of mathematical and statistical methods used. In a second part, we present, thought our own experiment, the difficulty of interpreting the results. Our experimentation's aim is to discover interesting usage patterns from a pedagogical Web Site data in order to adapt and personalize the presentation. The software used is *Clémentine*© (SPSS sociéty).

Key-words:
Characterization, usages, personalization, Web Mining, analytical methods

Introduction

Le Data Mining s'applique dans différents contextes industriels pour l'évaluation de processus et la prise de décision. Lorsque les données analysées proviennent ou ont pour objectifs d'extraire des informations sur l'Internet on parle plutôt de Web Mining. Selon l'objectif visé, plusieurs types d'études peuvent être menées : les premières concernent l'analyse du contenu des pages Web, on parle de Web Content Mining, les secondes concernent l'étude des liens entre les site Web, on parle alors de Web Structure Mining enfin les dernières s'intéressent à l'usage, c'est à dire aux traces laissées sur les sites Web lors des connexions , on parle alors de Web Usage Mining.

Srivastava (Srivastava, 2000) considère qu'il y a cinq secteurs ou activités dans lesquels on retrouve des études de *Web Usage Mining* :

- **Évaluer et caractériser de manière générale l'activité d'un site Web** : l'objectif étant ici de faire de l'observation, pas de modélisation ou d'induction. Les techniques d'analyse sont souvent simples, elles relèvent en effet du dénombrement et des statistiques simples (moyennes, histogramme, indice, tri croisés).

- **Améliorer les modes d'accès aux informations** : le Web Usage Mining est alors utilisé pour comprendre comment les utilisateurs se servent d'un site, s'il y a des failles dans la sécurité, des accès non autorisés identifiés.

- **Modifier la structure** : dans le même esprit que précédemment mais avec des techniques différentes, l'amélioration du système va se faire ici par une restructuration des pages et des liens. Les pages considérées comme similaires par des techniques de classification, seront reliées de manière hypertextuelle si ce n'est pas déjà le cas.

- **Personnaliser la consultation** : il s'agit de faire des recommandations dynamiques à un utilisateur en se basant sur son profil et une base de connaissance d'usages connus. C'est l'un des enjeux les plus important de bon nombre d'applications Internet ou de site de e-commerces.

- Faire de l'intelligence économique : pour les sites marchands, l'objectif est ici de comprendre : quand et comment l'utilisateur a été attiré sur le site, qu'est ce qui l'a retenu, quelles sont les ventes croisées que l'on doit lui proposer et qu'est ce qui a motivé son départ ?

Notre propos ici, est d'une part d'évaluer et de caractériser de manière générale l'activité du portail pédagogique I-Mont@igne (NTE, 2002), d'en améliorer l'accès aux informations et d'autre part de mettre en évidence des profils d'usagers ou des règles d'usage de manière à personnaliser la présentation.

Ce portail donne accès à plusieurs types de services pédagogiques, documentaires, administratifs, techniques comme :

« -Cours, supports de cours, exercices, texte de référence, base de données, bibliographie, webographie ;

- Autoformation aux logiciels bureautiques, statistiques … didacticiels de langues étrangères …

- Ressources documentaires (infothèque, médiathèque) ;

- Logithèque, réservation d'équipement pédagogique ;

- Catalogue des enseignements, emploi du temps, formulaires administratifs

- Conseils d'orientations universitaires, professionnelles

- Assistance. » (Ducasse, 2001)

Nous pouvons les classer selon 5 types d'usages : l'information, la formation, l'organisation, la communication et la gestion. Le logiciel utilisé est Clémentine© et est développé par la société SPSS.

1. Corpus expérimental, pré-traitement des données

Il est constitué de l'ensemble des connexions au portail pédagogique I-Mont@igne (NTE, 2002), sur la période allant du 30 janvier 2001 au 5 décembre 2001. Le fichier de log est composé de 438372 requêtes enregistrées à la norme ECLF (Extended Common Log Format) (Kimdall, 2000). Pour chacune d'elles, on connaît : la date, l'heure, l'adresse IP de l'utilisateur, son login (username), l'adresse IP du serveur, le port de connexion, la méthode, la ressource demandée, la question posée, le statut de réponse du serveur et l'agent de l'utilisateur.

Nous n'avons pas pu bénéficier d'une technologie plus précise, par cookies ou par tags, permettant d'identifier chaque visite ou visiteurs, la durée de consultation ou un chemin de navigation. Nous devons donc procéder à un certain nombre de pré-traitements pour identifier les visites et les durées de connexions.

Cooley (Colley, 2000) défini le processus de pré-traitement des données initiales selon le schéma suivant (figure 1).

Nous allons en détailler les premiers modules : le nettoyage des données, l'identification des visites et le calcul du temps de consultation. Nous terminerons en évoquant le processus d'échantillonnage nécessaire aux calculs et à la validation des modèles.

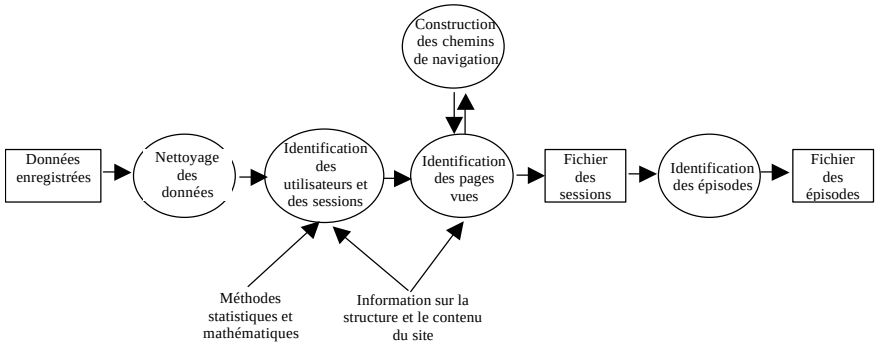

Figure 1 : Pré-traitement obligatoire avant analyse des usages

1.1 Nettoyer les données

Nous avons supprimé tous les enregistrements relatifs aux requêtes d'images[1] (extension jpg ou gif), le fichier de travail effectif comptait donc au final **211488** enregistrements.

La structure du site est telle que les différents services sont présentés dans des pages regroupées dans des répertoires ayant des noms significatifs : toutes les informations relatives à l'enseignement à distance sont stockées dans un répertoire *EAD*, celle concernant les cours dans un répertoire *SuppCour*. En extrayant le nom du premier répertoire dans l'URL, on peut attribuer à chaque page une catégorie de service, seule les pages d'accueil (comme *ead.html*), présentées à la racine constituent des catégories à part entière. Nous obtenons 287 catégories. Cet effectif, largement supérieur au nombre de service défini ci-dessus, s'explique par le fait que ces derniers ne sont pas directement accessibles au public. Il existe en effet des répertoires de travail pour les responsables des services (*resaudio* est le répertoire de travail et canal audio est le répertoire de présentation). De plus, le site *i-Mont@igne* étant en phase de développement, l'administrateur crée, supprime ou déplace des répertoires, les anciens correspondant à des usages effectifs ne devant pas être supprimés. Enfin, il ne faut pas oublier les fautes de frappes quand les URL sont directement tapées par l'utilisateur (par exemple pour accéder aux supports de cours, on retrouve les orthographes /*SuppCours*/, /*supp+cours*, /*supcours*, /*sup*/, /*soppcours*).

1.2 Identifier l'utilisateur et les visites.

Nous avons reconstruit les visites selon les critères empiriques de Kimball [KIM2000] : une visite est caractérisée par une série d'enregistrements séquentiellement ordonnés, ayant la même adresse IP et le même nom d'utilisateur, ne présentant pas de rupture de séquence de plus d'une demi-heure. Soulignons que nous n'avons pas pris en compte ici les recommandations de caractérisation sur l'agent spécifiées par Cooley (Cooley, 2000), nous le ferons dans une étude ultérieure. Selon ces principes nous comptons 25829 visites.

[1] Sachant que l'affichage de chacune des images nécessite une requête au serveur, les requêtes d'appel de pages, qui sont réellement porteuses d'informations, se retrouvent noyées et inexploitables statistiquement.

1.3 Calculer la durée de la visite

Théoriquement, la durée de consultation d'une page est le temps qui sépare deux requêtes http, diminué du temps nécessaire au chargement de la page (surtout lorsqu'elles sont alourdies par beaucoup d'images). Dans notre cas, la durée de consultation d'une page a été calculée par les différences entre les dates et heures des enregistrements successifs. Nous n'avons pas pris en compte les temps nécessaires au chargement des pages, même si certaines ressource sons et vidéo sont lourdes, car nous avons considéré d'une part que l'usage maximum se ferait sur le réseau haut débit de l'université et d'autre part que les données, présentées en format *RealMedia*, ne nécessitaient pas un téléchargement complet pour commencer à être visualisées. De plus, nous avons considéré que le temps de consultation de la dernière page était la moyenne des temps de consultation des pages pendant la visite.

Nous n'avons, pour cette première analyse pas poursuivi le processus de pré traitement et n'avons pas cherché à : identifier les pages complètement téléchargées, reconstitué les chemins de navigation et identifier des épisodes. Ce travail sera fait ultérieurement.

1.3.1 Construction de l'échantillon de travail

Lincoln (http://www.web-datamining.net) spécifie qu'*« aucune méthode ou indicateur n'est apte à ce jour à démontrer les qualités prédictives et la stabilité d'un réseau à partir simplement des résultats fournis sur la base d'apprentissage du modèle (c'est à dire la base ayant servi à sa constitution) »*. Argumentant de plus que le nombre de facteurs internes est très important, il préconise que la méthodologie la plus efficiente doit s'appuyer sur la création d'un échantillon large (40% et plus de la base initiale) pour l'apprentissage, la création d'un nouvel échantillon pour tester la validité du modèle

(30% de la base initiale) et enfin la création d'un échantillon final (70% des connexions) qui testera globalement et permettra de comparer le ou les modèles (4-5 modèles maximum). La création de l'échantillon doit se faire dans le respect de la complétude des visites. De plus, la création des échantillons de validation et de test doit se faire sur une année entière.

Notre échantillon a été constitué par extraction de manière aléatoire de 25% des visites (40% représentant en effet un volume trop important pour notre logiciel de traitement *Clémentine*©).

2. Méthodes d'analyses du Web Usage Mining

Les méthodes d'analyse des données se répartissent selon quatre grandes familles :

- **les méthodes statistiques unidimensionnelles** (dénombrement, indice, moyenne, …)
- **les méthodes statistiques multidimensionnelles** (factorisation, segmentation, classification, régression, …)
- **Les méthodes de visualisation des données** (règles d'association).
- **Les méthodes s'appuyant sur l'intelligence artificielle** (réseaux de neurones)

Ces méthodes sont utilisées dans deux contextes différents :

- pour faire de la modélisation descriptive, on retrouve les techniques de classification (centre mobile, classification ascendante hiérarchique), factorielles (ACP) et d'association (règle d'association, réseaux bayesiens)

- pour faire de la modélisation prédictive, on retrouve les techniques de segmentation (d'arbre de décision et de réseaux de neurones) et d'estimation (régression, réseaux de neurones).

Notre propos n'est pas ici de présenter en détail chacune des méthodes, mais d'expliquer globalement à quoi elles servent et comment elles s'exploitent.

2.1 Analyse statistiques unidimensionnelles

Les indicateurs d'audience éditoriale selon la terminologie du CRESP (Centre d'étude des supports de Publicité) [CRE2002] sont :

- Le nombre de pages demandées ou vues (c'est à dire totalement téléchargées).
- Nombre de pages provenant de mémoires-caches ou de serveurs proxy.
- Nombre de visiteurs.
- Nombre de pages vues par visite.
- Origine géographique des consultations
- Durée de consultation par visite

2.2 Les méthodes statistiques multidimensionnelles

Elles permettent, en réduisant l'espace, en fournissant des représentations graphiques, d'exploiter, de fouiller, de représenter de grands ensembles de données.

2.2.1 Méthodes factorielles

Les méthodes factorielles se proposent de fournir des représentations synthétiques, souvent sous forme graphique, de vastes ensembles de valeurs numériques. *L'analyse en composante principale* (ACP) s'applique sur des tableaux dont les lignes sont des individus et les colonnes des variables descriptives numériques, *l'analyse factorielle des correspondances* (AFC) s'applique sur des tableaux de contingence, c'est à dire des tableaux de comptage de co-occurrence de variables nominales. Son extension, *l'analyse factorielle des correspondances multiples* (AFCM) s'utilise sur de gros tableaux de variables nominales (résultat d'enquête par exemple) où les lignes sont les individus et les colonnes des variables descriptives (Lebart, 2001).

2.2.2 La segmentation : construction d'arbres de décision :

« Il s'agit de diviser la population successivement en sous-groupes selon les valeurs prises par les variables, qui à chaque stade discriminent le mieux la variable à modéliser. La variable à expliquer est par définition le sommet de l'arbre. L'objectif est de prédire la probabilité d'appartenance de chaque individu à l'une ou l'autre des catégories de la variable à expliquer. » (Web Data-Mining, 2001)

Il existe plusieurs méthodes de construction des arbres, la plus ancienne est la méthode *arbre C5*, la plus récente et la plus robuste (Lebart, 2001) est la *méthode CART*. L'arbre de décision est particulièrement efficace sur de gros volumes de données qualitatives et il est facilement interprétable par des non-statisticiens mais il se révèle en revanche moins efficace que les régressions sur les données quantitatives.

2.2.3 La classification

Ces techniques sont destinées à produire des groupements de lignes ou de colonnes dans un tableau de manière à voir apparaître des classes d'individus. Les représentations se font sous forme de partitions simples des ensembles étudiés (*méthode d'agrégation par les centres mobiles* ou *K-means*) ou de partitions hiérarchisées (algorithmes ascendants ou descendant) qui ont un aspect d'arbre (au sens de la théorie des graphes) (Lefébure, 2001).

2.3 L'association

2.3.1 Co-occurrence

Les associations les plus simples sont calculées en fonction du nombre d'occurrences simultanées de couples de modalités, c'est la co-occurence. Les résultats de ces comptages permettant de produire des réseaux d'association dans lesquels l'épaisseur du lien est fonction du nombre de co-occurrence. Pour rendre ce graphe plus signifiant, on peut grouper les termes géographiquement dans des « cluster » avec une technique de classification comme décrit ci dessus, on parle alors de diagramme stratégique. En sciences de l'information cette technique est assez largement utilisée sur des corpus textuels - on parle de l'analyse des mots associés - ou sur les bibliographies d'articles - c'est la méthode des co-citations.

2.3.2 Règle d'association

Les associations se feront ici sur des n-uples de modalités, c'est à dire les occurrences simultanées de plusieurs modalités, la force de l'association proprement dite est mesurée par la probabilité conditionnelle de retrouver une modalité y en co-occurrence avec un ensemble de n autres {a, b, c, d, …} qui le sont déjà. Cette probabilité s'appelle la confiance. La règle correspondante est « si {a, b, c, d, …} sont en cooccurrence alors y l'est aussi avec eux » et s'exprime comme {y}<={a, b, c, d, …}. Une règle est considérée comme bonne et est sélectionnée si son seuil de confiance est élevé et si elle est présente souvent, c'est à dire si elle a un support élevé. Parmi les différents algorithmes, celui qui est le plus souvent utilisé est APRIORI et a été développé par IBM. Cette technique est utilisée dans le cas de personnalisation d'une offre ou de vente croisée (Bazsalicza, 2001).

2.3.3 Les réseaux Bayesiens

Les réseaux Bayesiens est un modèle graphique probabiliste de connaissance. C'est une distribution de probabilité (appelée quantification de la connaissance) factorisée suivant un graphe (appelé structure de la connaissance) (Bazsalicza, 2001).

2.4 Modèle de dépendance basé sur l'intelligence artificielle : Réseaux de neurones.

Les réseaux de neurones sont utilisés pour prédire les valeurs prises par une ou plusieurs variables. Fondée au départ sur des analogies biologiques, le réseau de neurones est utile pour mettre en évidence des relations non linéaires entre les variables. Il sert aussi à faire une bonne approximation des relations si l'on sait qu'il existe des relations non-linaires mais non identifiées entre plus de trois variables. (Web Datamining 2001) :

3. Résultats

Nous présenterons ici les résultats des indicateurs de statistiques simples et de quatre méthodes de modélisations :

- une méthode de classification par les K-means.

- deux méthodes d'association ; la première basée sur la représentation graphique des co-occurrences et la seconde s'appuyant sur la méthode des règles d'association APRIORI.

- une méthode neuronale s'appuyant sur les cartes de Kohonen[2].

Nous n'avons pas pu tester de méthodes factorielles, la seule de ce type proposée par *Clémentine*© est l'ACP et nos valeurs, en grande majorité qualitatives, ne lui convienne pas. De plus, les deux méthodes de construction d'arbre (C5 et CART) ont échoué car notre échantillon contient trop de valeurs uniques.

3.1.1 Résultats généraux

Le bilan général est de :

- 212388 pages demandées sur site
- 25829 visites
- 103,02 visites par jour en moyenne
- 17,26 pages vues par jour en moyenne
- 8,18 pages demandées par visite
- 1401 secondes, c'est à dire 23 minutes et 21 secondes de temps de consultation par visite.

N'ayant pu caractériser correctement les visiteurs, il est préférable de ne pas les considérer.

Nous avons ensuite comptabilisé le nombre et les fréquences des requêtes faites pour les catégories de services (tableau 1), ainsi que les visites faites par utilisateurs (tableau 2) (nous ne présentons, pour des raisons de place que les 10 premiers résultats).

Les usages principaux observés sont la formation (principalement la consultation des supports de cours 44% des requêtes), la gestion (principalement celle des impressions papier des étudiants), l'organisation (en particulier l'accès à l'outil de travail collaboratif Exchange) et l'information avec l'accès à la médiathèque. Les utilisateurs sont en grande partie (73%) anonymes.

Existe-il des règles d'usage ? Nous allons le vérifier en utilisant 4 modèles classiques de Web Usage Mining.

---

2 « Les "cartes de Kohonen" sont une classe de réseaux de neurones ayant principalement la particularité de prendre en compte des propriétés de continuité spatiale ou temporelle. Ce type de réseau s'appuie sur une dynamique de propagation multi-directionnelle avec de fortes interactions entre neurones d'un même voisinage. Les cartes de Kohonen conviennent bien au traitement de problèmes intégrant une dimension spatiale ou temporelle (ou les deux).» (Web Datamining, 2001) :.

catégorie 2001	Nb Requête	pourcentage
/SuppCours/	95429	44,9420264
/intranet.css	20374	9,59507954
/index.php	15468	7,28461227
/Printings/	12774	6,01588034
/public/	10664	5,02218162
/ead/	8565	4,03366331
/exchange/	7031	3,31123021
/exchweb/	6826	3,21468602
/scripts/	2661	1,25319067
/Canalaudio/	2332	1,09824902
/Canalvideo/	2199	1,03561303

Tableau 1 Nombre d'accès pour chaque catégorie

cs-username	Nb visite	pourcentage
Total	25829	100
-	18985	73,50265
moniteur	2657	10,28689
ducasse	918	3,554145
desclaud	765	2,961787
arcelin	714	2,764335
ead	579	2,241666
svalat	453	1,753843
CMichel	177	0,685276
srouissi	131	0,507182
dubois	109	0,4220062

Tableau 2 Nombre de visite par visiteur

3.1.2 La méthode des K-means

La méthode des K-means a été choisie pour mettre en évidence des classes d'individus. Les critères de description des classes sont : les catégories des pages, les noms des utilisateurs, la durée moyenne de consultation d'une page et le nombre de pages vues. Cette méthode a permis de construire 5 classes (voir tableau 3). Pour chaque modalité, la fréquence d'apparition est présentée (pour des raisons de place nous ne reportons ici que les valeurs les plus fortes). Les classes 1 et 4 correspondent à des consultations de personnes anonymes (99%) vers des catégories assez hétérogènes, la classe 1 étant cependant plutôt orienté vers de la formation alors que la classe 4 vers de l'information. La différence principale entre ces deux classes porte sur le temps de consultation, faible dans le premier cas (25 secondes en moyenne) et fort dans le second (6083 secondes).

Les classes 2 et 5 regroupent des usages de gestion, la classe 2 correspond presque exclusivement au service d'impression papier alors que la classe 5 regroupe tous les accès liés à la maintenance, l'alimentation et mise à jour du site. La classe 3 regroupe les connexions motivées par l'organisation du travail via l'intranet (consultation des emplois du temps, du calendrier, réservation de ressources, consultation des dossiers publics.

Classes 1: 42149 requêtes	**Classes 2: 2681 requêtes**	**Classes 3: 2152 requêtes**	**Classes 4: 1181 requêtes**	**Classes 5: 2184 requêtes**
catégorie : > /SuppCours/ -> 0.497307 /intranet.css -> 0.119006 /index.php -> 0.082873 /exchweb/ -> 0.029514 /EAD/ -> 0.027901 /Canalaudio/ -> 0.016204 /Printings/ -> 0.015872 /scripts/ -> 0.014876 /Canalvideo/ -> 0.01089 /Annuaire/ -> 0.009538 /Divers/ -> 0.008944 /PtesAnn/ -> 0.008802 /Infotheque/ -> 0.008541 **username :** - -> *0.999408* ducasse -> 0.000142 moniteur -> 0.000142 arcelin -> 0.000119 CMichel -> 0.000095 ead -> 0.000071 desclaud -> 0.000024 **duree_sec :** 25 **ordre_page :** 78	**categorie :** /Printings/ -> 0.989183 /MAJactu/ -> 0.006341 GET -> 0.001119 /EAD/ -> 0.000746 /Annuaire/ -> 0.000373 /Canalaudio/ -> 0.000373 /Canalvideo/ -> 0.000373 /Divers/ -> 0.000373 /SuppCours/ -> 0.000373 /logitheque.php -> 0.000373 /public/ -> 0.000373 **username :** *moniteur -> 0.961582* - -> 0.027229 majactu -> 0.005595 MONITEUR -> 0.00373 desclaud -> 0.001119 barbier -> 0.000746 **duree_sec :** 295 **ordre_page :** 5	**categorie :** /ExAdmin/ -> 0.00697 /Exchange/ -> 0.000929 /exchange/ -> 0.000929 /public/ -> 0.991171 **username :** arcelin -> 0.438197 ead -> 0.19238 ducasse -> 0.106877 desclaud -> 0.093401 svalat -> 0.069238 - -> 0.063662 srouissi -> 0.013011 CMichel -> 0.003253 alexis -> 0.003253 EAD -> 0.002788 administrateur -> 0.002788 cdubois -> 0.002788 moniteur -> 0.002788 ARCELIN -> 0.001394 dubois -> 0.001394 nakam -> 0.001394 **duree_sec :** 17 **ordre_page** : 11	**categorie : >** /index.php -> 0.244708 /intranet.css -> 0.171888 /SuppCours/ -> 0.083827 /EAD/ -> 0.039797 /Divers/ -> 0.028789 /Canalaudio/ -> 0.027096 /Infotheque/ -> 0.026249 /index.htm -> 0.022862 /Resaudio/ -> 0.021169 /Canalvideo/ -> 0.019475 /FLE/ -> 0.017782 /Printings/ -> 0.016935 /logitheque.php -> 0.016935 /scripts/ -> 0.016088 /texte.txt -> 0.015241 /Annuaire/ -> 0.013548 /EAD.htm -> 0.011854 /PtesAnn/ -> 0.011008 /_vti_bin/ -> 0.011008 /Depinfo/ -> 0.009314 /fle/ -> 0.009314 **username :** - -> *0.994073* ead -> 0.003387 moniteur -> 0.00254 **duree_sec** : 6083 **ordre_page** : 8	**categorie : >** /exchange/ -> 0.599817 /exchweb/ -> 0.221153 /exchange -> 0.048535 /Printings/ -> 0.024267 /ExAdmin/ -> 0.02152 /MAJactu/ -> 0.016941 /Annuaire/ -> 0.008242 /intranet.css -> 0.008242 /index.php -> 0.006868 /Resaudio/ -> 0.005037 /Canalvideo/ -> 0.004121 /fle2/ -> 0.004121 **username :** ducasse -> 0.294414 desclaud -> 0.171246 - -> 0.094322 svalat -> 0.090659 ead -> 0.067766 CMichel -> 0.054945 srouissi -> 0.045788 arcelin -> 0.036172 moniteur -> 0.027473 MBH -> 0.016484 **duree_sec : 9** **ordre_page :** 4

Tableau 3 : Classe d'usage pour la variable catégorie

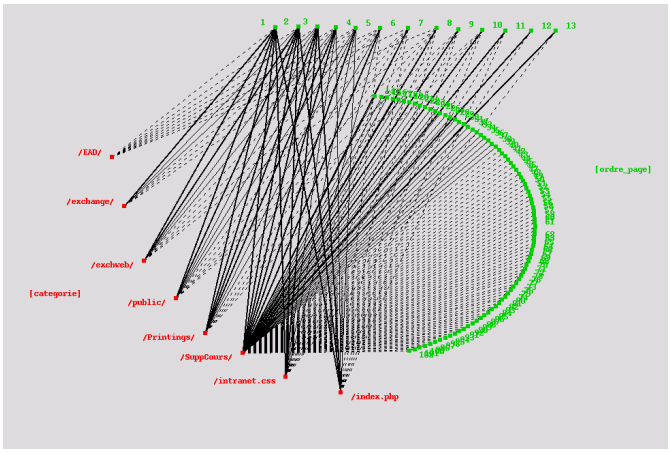

Figure 2 : Co-occurrence entre les catégories et l'ordre de la page dans la visite

3.1.3 Co-occurrence

Le nœud relation du logiciel *Clémentine©* permet de présenter des co-occurrences sous forme graphique. Notre propos était d'utiliser cet outil pour mettre en évidence un éventuel usage détourné du site. La figure 2 présente les co-occurrences[3] entre les catégories et l'ordre de la page dans la visite[4]. On peut clairement voir que la page d'accès générale (*index.php*), les pages contenant les supports de cours (*Suppcours*), la gestion de l'impression (*printing*) et l'accès à l'intranet (*exchange* et *exchweb*) sont consultées assez directement. Les utilisateurs ne viennent pas « flaner » mais consulter dans un but précis une ressource. On peut supposer que cet outil est massivement utilisé par sa communauté (enseignants, administratifs et étudiants) qui en connaît le contenu et l'organisation et relativement peu comme une source d'information pour des utilisateurs externes.

3.1.4 Règle d'association

Comme pour la classification, nous essayons de mettre en évidence les règles d'usages déterminant la consultation des différentes catégories. L'algorithme APRIORI est appliqué avec une couverture minimale de 10%, une précision minimale de 50% et un maximum de 5 préconditions. Il propose 16 règles (voir tableaux 4 et 5).

On peut remarquer que parmi les différentes catégories, seule *SuppCours*, c'est à dire la consultation de cours en ligne, est expliquée. Cela est du à sa fréquence d'apparition (44% voir tableau 1). La fiabilité est à peu près toujours équivalente (0,5 soit 50% environ). Les règles 1 et 2 et 3 concernent 80% des requêtes ce qui n'est pas étonnant si l'on regarde la généralité du profil : un utilisateur anonyme se connecte sans avoir posé de question, avec la méthode get ! Les requêtes 7,8,9,11,12,13,14 et 15 signalent de plus que le profil général des agents est une configuration Microsoft (IE5 avec Windows NT ou 98).

La durée de connexion (de 0 seconde ou de 2 à 5 secondes) est sujette à plusieurs interprétations.

Généralement, les requêtes de 0 seconde correspondent à des processus automatiques (indexation par un robot ou enregistrement par un aspirateur) ou à des requêtes comportant des erreurs (requête interrompue avant chargement complet de la page). Les codes d'état de la catégorie *SuppCours* le montrent bien (voir tableau 6) : 4660 requêtes sont servies en code 304 c'est à dire « page non modifiée », 4908 requêtes ont été servies correctement (code 200). On peut donc très certainement dire que ce sont des consultations de robots.

N°	Règle
1	categorie=/SuppCours/ <= (question=-)(methode=GET)
2	categorie=/SuppCours/ <= (question=-)(utilisateur=-)
3	categorie=/SuppCours/ <= (question=-)(methode=GET)(utilisateur=-)
4	categorie=/SuppCours/ <= (question=-)(methode=GET)(duree_echelle=0s)
5	categorie=/SuppCours/ <= (question=-)(utilisateur=-)(duree_echelle=0s)
6	categorie=/SuppCours/ <= (question=-)(methode=GET)(utilisateur=-)(duree_echelle=0s)
7	categorie=/SuppCours/ <= (question=-)(methode=GET)(utilisateur=-)(agent=Mozilla/4.0+(compatible;+MSIE+5.5;+Windows+NT+5.0))
8	categorie=/SuppCours/ <= (question=-)(utilisateur=-)(agent=Mozilla/4.0+(compatible;+MSIE+5.0;+Windows+98;+DigExt))
9	categorie=/SuppCours/ <= (methode=GET)(utilisateur=-)(agent=Mozilla/4.0+(compatible;+MSIE+5.0;+Windows+98;+DigExt))
10	categorie=/SuppCours/ <= (question=-)(methode=GET)(utilisateur=-)(agent=Mozilla/4.0+(compatible;+MSIE+5.0;+Windows+98;+DigExt))
11	categorie=/SuppCours/ <= (duree_echelle=0s)(agent=Mozilla/4.0+(compatible;+MSIE+5.5;+Windows+NT+5.0))
12	categorie=/SuppCours/ <= (methode=GET)(duree_echelle=0s)(agent=Mozilla/4.0+(compatible;+MSIE+5.5;+Windows+NT+5.0))
13	categorie=/SuppCours/ <= (question=-)(duree_echelle=0s)(agent=Mozilla/4.0+(compatible;+MSIE+5.5;+Windows+NT+5.0))
14	categorie=/SuppCours/ <= (question=-)(methode=GET)(duree_echelle=0s)(agent=Mozilla/4.0+(compatible;+MSIE+5.5;+Windows+NT+5.0))
15	categorie=/SuppCours/ <= (question=-)(duree_echelle=2-5s)
16	categorie=/SuppCours/ <= (utilisateur=-)(duree_echelle=0s)(agent=Mozilla/4.0+(compatible;+MSIE+5.5;+Windows+NT+5.0))

Tableau 4 : Règles d'association pour la variable « catégorie »

N°	occurrence	pourcentage	fiabilité
1	41536	82.5%	0.503
2	41133	81.7%	0.512
3	39422	78.3%	0.529
4	19736	39.2%	0.509
5	19233	38.2%	0.525
6	18679	37.1%	0.537
7	8408	16.7%	0.508
8	7149	14.2%	0.513
9	7149	14.2%	0.515
10	6898	13.7%	0.531
11	6243	12.4%	0.526
12	5991	11.9%	0.546
13	5941	11.8%	0.551
14	5740	11.4%	0.573
15	5639	11.2%	0.526
16	5639	11.2%	0.582

Tableau 5 : Nombre d'occurrence des règles d'association

Les requêtes de 2 à 5 secondes correspondent à des consultations effectives par des opérateurs humains, et sont effectivement bien servies par le serveur (2006 en code 200 sur 2976 requêtes) (voir tableau 7). Nous pouvons dire globalement que les supports de cours sont consultés toujours massivement par des opérateurs automatiques ou des utilisateurs ayant un profil très général (pas de *login* et/ou pas de question et/ou une méthode GET et/ou une configuration Microsoft) ne restant connectés que 2 à 5 secondes. Les cours ne sont donc pas directement lu sur le site, mais enregistrés ou imprimés pour consultation ultérieure. Pour confirmer cette hypothèse nous avons comptabilisé les accès à *SuppCours* en fonction de l'adresse IP de l'utilisateur. Le plus grand nombre de connexions (7,2%) correspond effectivement bien à des machines sur le campus (tableau 8).

De manière à voir si d'autres règles peuvent être trouvées nous avons appliqué la méthode APRIORI sur les enregistrements sans *SuppCours*. Avec une couverture minimale de 8% une seule règle est calculée : **categorie=/Printings/ <= (username=moniteur),** c'est à dire, lorsqu'un moniteur se connecte c'est pour gérer les impressions papier des étudiants.

3 Les liens pointillés correspondent à 200-500 co-occurrence, le liens pleins à 500-1000 co-occurrence et le liens gras à plus de 1000 co-occurrence entre les modalités de catégorie et de durée de connexion.

4 Pour chaque page et chaque visite on note si la page est la 1ère, 2ème , ..., nème page vue. Ce numéro correspond à l'ordre de la page dans la visite.

État	Occurrence
200	4908
206	190
207	17
302	32
304	4660
403	29
404	324
500	4
Total	10164

Tableau 6 : Ventilation des états pour les connexions de 0 seconde

État	Occurrence
200	2006
206	87
207	9
302	6
304	797
400	1
403	4
404	66
total	2976

Tableau 7 : Ventilation des états pour les connexions de 2-5 secondes

Adresse IP	Fréquence
147.210.245.2	7,2%
172.29.20.87	5,22%
172.29.20.70	3,96%

Tableau 8 : Fréquence des adresses IP (supérieures à 3%)

État	Occurrence	Fréquence
200	12203	57,62%
304	7714	36,42%
404	557	2,63%
206	536	2,53%

Tableau 9 : Occurrence et fréquences (en pourcentage) des états (supérieure à 3%)

3.1.5 Kohonen

Nous avions 75 neurones en couche d'entrée, et 35 en couche de sortie. 21 classes de description des connexions ont été crées. Les résultats sont reportés dans le tableau 10 et la figure 3.

Les résultats produits tels quels ne sont pas facilement interprétables. On peut voir que les classes 2-4 et 4-0 sont les plus grosses mais on ne sait pas quels sont les critères qui les définissent. Lefèvre (Lefèvre, 2001) recommande de procéder à des traitements complémentaires de classification pour les mettre en évidence. Nous le ferons ultérieurement.

classe	Nb hit	pourcentage
2-4	10579	21,0121755
4-0	9929	19,7211353
6-0	5057	10,0442926
0-0	4045	8,03424236
0-4	3907	7,7601446
4-4	3097	6,15130991
0-2	3052	6,0619302
2-0	2333	4,63384114
4-2	1802	3,57916063
2-2	1781	3,5374501
6-2	1755	3,48580849
6-4	1458	2,89590244
6-1	575	1,14207401
6-3	369	0,73291358
4-3	315	0,62565793
5-4	206	0,40916043
5-2	46	0,09136592
4-1	20	0,03972431
3-2	13	0,0258208
5-3	7	0,01390351
2-3	1	0,00198622

Tableau 10 : Classes Kohonen

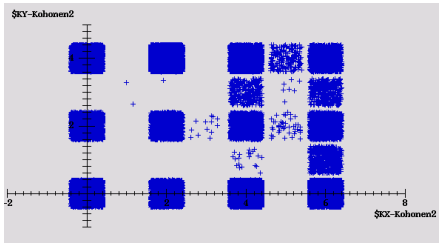

Figure 3 : Carte de Kohonen

4 Conclusion

L'évaluation générale montre une situation caricaturale où l'usage principal de ce site est lié à la formation (principalement la consultation des supports de cours), à la gestion (principalement celle des impressions papier) et dans une bien moindre mesure à l'organisation du travail (par l'utilisation d'un environnement de travail collaboratif). Ce déséquilibre biaise considérable le processus de création de connaissance, nous ne pouvons en effet pas dire que nous avons réellement extrait des informations ou des connaissances originales, auparavant inconnues, potentiellement utiles. Cependant, les méthodes d'analyses ont permis de mettre en évidence certaines caractéristiques très intéressantes :

- les utilisateurs anonymes se répartissent entre utilisateurs sur le campus et opérateurs automatiques (robots ou aspirateurs),
- les consultations sont directes, on donc peut supposer que les utilisateurs sont familiarisés avec l'interface (ce qui signifie à contrario que très peu de connexions ont une provenance complètement extérieure),
- les accès principaux se font sur le campus universitaire mais la consultation des cours ne se fait pas sur place (ils sont aspirés ou manuellement copiés et consultés ailleurs).

Les résultats des classifications auraient pu permettre la préconisation de quelques règles simples de personnalisation mais la généralité des résultats ne nous encourage pas à le faire. De plus, cela nous semble à l'heure actuelle inutile car ce site est en phase de développement et donc différents services vont être ouverts plus largement en particulier aux étudiants (à titre d'exemple, le service intranet collaboratif a été ouvert en janvier 2002 à une filière d'étudiants) et il est à gager que ces derniers vont massivement les utiliser. Les procédures de pré-traitement et d'analyses élaborées ici seront alors réutilisées. Nous pourrons procéder alors sur ces modèles à une analyse complète avec validation sur un échantillon plus large que l'échantillon test.

Remerciement *: Nous tenons à remercier le Conseil Régional d'Aquitaine qui a soutenu financièrement les projets « Territoire Virtuels de Communication » et « SAVANTE : Système Aquitain de Valorisation et d'Apprentissage des Nouvelles Technologies de l'Education)» dans le cadre desquels est réalisée cette étude.*

5 Références

(Alin, 1998) : Alin F, Lafont D et Macary J.F – *Le projet intranet. De l'analyse des besoins d'entreprise à la mise en œuvre des solutions* – Edition Eyrolles, Paris, 436p, 1998.

(Bazsalicza, 2001) : Bazsalicza M et Naim P – *Data Mining pour le Web : Profiling, filtrage collaboratif, personnalisation client* – Edition Eyrolles, Paris, 2001, 279p.

(Besse, 2001) : Besse P – *Statistique et datamining* - Cours version septembre 2001- Publication du laboratoire de statistique et probabilité, disponible en ligne (http://www.lsp.ups-tls.fr/Besse).

(Cooley, 2000) : COOLEY Robert W – *Web usage mining : discovery and application of interresting patterns from web data* – These soutenue à l'Université du Minnesota, Mai 2000, 170p.

(CRESP, 2002) : http://www.cesp.org/ site consulté le 15 janvier 2002

(Dazy, 1996) : DAZY Frédéric et LE BARZIC Jean-François - *Analyse des données évolutives, méthodes et applications*. Edition Technip, Paris. 1996

(Ducasse, 2001) Ducasse R – « Université virtuelle, campus numérique » in *Contact* – N°147, Mars 2001, p17.

(Kimball, 2000) : Kimball R et Merz R – *Le Data Webhouse. Analyser des comportements client sur le Web*. – Edition Eyrolles, Paris, 310p, 2000

(Lafouge, 2002) : Lafouge T, Le Coadic Y.F. et Michel C. – *Éléments de statistique et de mathématique de l'information*- Presses de l'Enssib, Lyon, 2002.

(Lebart, 2000) Lebart L, Morineau A et Piron M – *Statistique exploratoire multidimensionnelle* – Edition DUNOD, Paris 2000

(Lefébure, 2001) : Lefébure . et Venturi G - *Data Mining. Gestion de la relation client, personnalisation de sites Web* – Edition Eyrolles, Paris , 2001, 391p.

(NTE, 2002) http://www.nte.montaigne.u-bordeaux.fr

(Srivastava, 2000) : Srivastava J., Cooley R., Deshpande M. et Tan P. – "Web usage mining : discovery and applications of usage patterns from web data". in *SIGKDD Explorations* – Volume1, Issue 2 pp12-23 – editions ACM - Jan 2000.

(Web Datamining 2001) : http://www.web-datamining.net site consulté en décembre 2001

(Wonnacott, 95) : Wonnacott T.H. et Wonnacott R.J – Statistique – Edition Economica – 1995 – 919p.

.