UNIVERSITÉ KASDI MERBAH OUARGLA

Faculté des Nouvelles Technologies de

L'Information et de la Communication

Département d'Informatique et des Technologies de l'Information

Mémoire

En vue de l'obtention du diplôme de

MASTER ACADEMIQUE

Domaine : Informatique

Filière : Informatique

Spécialité : Informatique industrielle

Présenté par : Djerbaoui Imad Eddine

Herrouz Hichem

Thème

Exploration des traces de navigation sur le Web.

Soutenu publiquement le : 08/06/2015

Devant le jury composé de :

M.	Mahdjoub	Med Bachir	Université de Ouargla	**Président**
M.	Zga	Adel	Université de Ouargla	**Examinateur**
M.	Herrouz	Abdelhakim	Université de Ouargla	**Rapporteur**

Remerciement

Nous tenons tous d'abord à remercier le bon dieu tout puissant de nous avoir aidés à réaliser ce modeste travail.

Nous adressons nos sincères remerciements à notre encadreur monsieur ***Herrouz Abd El Hakim*** *; maitre-assistant A à Université Kasdi Merbah Ouargla, qui n'a ménagé aucun effort pour que ce mémoire puisse voir le jour ; nous lui exprimons notre gratitude de nous avoir dirigé, encouragé et surtout aidé afin de réaliser ce mémoire.*

Nous remercierons très chaleureusement le Professeur ***Boutarfaia Ahmed*** *; recteur de l'Université Kasdi Merbah Ouargla qui nous avoir encouragé et nous tenons à lui exprimer notre profonde gratitude pour l'attention et les conseils qu'il nous a prodigué malgré ces multiples responsabilités.*

Nos vifs remerciements vont également à tous les membres de jury pour l'honneur qu'ils nous ont fait en acceptant de juger notre travail, ainsi qu'aux tous les enseignants de notre cursus universitaire qui ont contribué à notre formation.

Enfin, il serait difficile d'omettre de remercier tous ceux qui ont contribué de près ou de loin à ce travail, qu'ils trouvent dans ses quelques lignes l'expression de nos sincères remerciements.

Imad Eddine & Hichem

Dédicace

Je dédie ce modeste travail :

*A la mémoire de mon cher papa **Ahmed** qui était et restera toujours mon grand exemple, et sa chaleur paternelle a été toujours pour moi un grand réconfort ; Qu'Allah bénisse son âme.*

*A ma chère maman **Nadjat** qui n'a jamais cessé de ménager ses efforts pour que j'atteigne ce niveau. Ses sacrifices et privations ne l'ont pas empêchée d'accomplir son devoir de mères soucieuses de l'avenir de ses enfants.*

*A ma sœur **Amina** et son époux **Abderraouf** qui étaient toujours à mes côtés et qui n'ont jamais cessé de me soutenir et de m'encourager ; jamais de simples mots ne permettront de vous exprimer mes remerciements.*

*A ma petite sœur **Ikram** que je lui souhaiterai le succès dans ses études et qu'elle aura un avenir scientifique prospère.*

*A ma fiancée **Sara** qui m'a toujours encouragé et m'a submergé de sa tendresse ; Je lui souhaite une bonne chance dans la soutenance de son mémoire de fin d'étude également.*

A ma grande famille grande et petite, Ainsi que mes chères ami (e)s et tous mes collègues de la promotion de 2014 /2015.

Imad Eddine

Dédicace

Je dédie ce travail :

A mon père et mon enseignant :

Mr Abdelhakim Herrouz *en signe de reconnaissance de l'immense bien que vous avez fait pour moi concernant mon éducation qui aboutit aujourd'hui à la réalisation de cette étude. Recevez à travers, Toute ma gratitude et mes profonds sentiments. Merci d'être toujours avec nous, Que dieu le tout puissant soit à vos côtés et vous accorde une meilleure santé (amen).*

A ma mère :

Mme Herrouz Nora Bouafia *pour m'avoir donnée la vie et la joie de vivre. Ta bonne éducation, tes conseils et tes bénédictions et tes prières n'ont jamais fait défaut, que dieu te bénisse.*

A mes chers frères, et ma chère sœur :

Hatem, Mustapha et Sana *pour leurs affections, compréhension et patience, ce modeste travail doit vous servir d'exemple pour réussir et faire mieux que votre grand frère ; je vous aime.*

Hichem

Table des matières

Chapitre 3: Panorama des travaux sur le domaine WUM

Chapitre 4: Coneption et Réalisation

Liste des figures

Liste des tables

Résumé

L'objectif de ce travail est la conception et la réalisation d'une application web, en utilisant les concepts du « Web usage mining », qui permettra au « Webmaster » d'avoir l'ensemble des connaissances sur le site web qu'il gère, en vue d'une amélioration et personnalisation.
Il s'agit en fait, d'extraire de l'information à partir du fichier Log du serveur Web, hébergeant le site Web, et prendre les décisions pour découvrir les habitudes des internautes, et de répondre à leurs besoins en adaptant le contenu, la forme et l'agencement des pages web.

***Mots-clés** : fouille des donnés, fouille d'utilisation web, fichiers logs, traces de navigation, prétraitement des logs.*

Abstract

The objective of this work is the design and implementation of a web application, using the concepts of "Web usage mining", which will allow the "Webmaster" to have the body of knowledge on the website that manages, for an improvement and customization.
This is in fact to extract information from the log file of a Web server, hosting the website, and decide to discover the habits of Internet users, and meet their needs by adapting the content, the shape and layout of web pages.

***Keywords:** data mining, web usage mining, log file, navigation trace, log pretreatment.*

ملخص

الهدف من هذا العمل هو تصميم تطبيق ويب وذلك باستخدام مفاهيم " استخدام التعدين في الويب"، والتي تسمح لمشرف الموقع استنباط مجموعة من المعارف، لتحسين لتخصيص الموقع.

وهذا التطبيق في الواقع لانتزاع معلومات من ملف سجل خادم الويب، استضافة الموقع، واتخاذ قرار لاكتشاف سلوك مستخدمي الإنترنت، وتلبية احتياجاتهم من خلال تكييف المحتوى، وشكل وتصميم صفحات الويب.

***كلمات البحث:** تعدين البيانات، تعدين استعمال الويب، ملف السجل، أثر الإبحار، تهيئة ملفات السجل.*

Introduction Générale

Introduction

En ce qui concerne les sources de données volumineuses et dynamiques, le Web est devenu l'exemple le plus pertinent grâce à l'augmentation colossale du nombre de documents mis en ligne et des nouvelles informations ajoutées chaque jour. Dans la perspective d'attirer de nouveaux clients et de répondre aux attentes des clients existants, un gérant de site Web bien avis é doit toujours garder à l'esprit que le fait d'offrir plus d'information ne constitue pas toujours une bonne solution. En réalité, les usagers d'un site Web apprécieront davantage la manière dont cette information est présentée au sein du site. L'analyse des traces d'usage (enregistrées dans les fichiers de type journaux par le serveur qui héberge le site Web) s'avère une pratique de plus en plus nécessaire pour mieux appréhender les pratiques des internautes. Dans ce contexte, la dimension temporelle joue un rôle très important car la distribution sous-jacente des données d'usage peut changer au cours du temps. Ce changement peut être provoqué par la mise à jour du contenu et/ou de la structure du site Web ou bien par le changement naturel d'intérêt des usagers d'un site Web.

Les modèles d'accès à un site Web ont une nature dynamique et peuvent être influencés par certains facteurs temporels, comme par exemple : l'heure et le jour de la semaine où se déroule la visite d'un site Web, des événements saisonniers (vacances d'été, d'hiver), des événements ponctuels dans le monde (compétitions sportives, élection présidentielle, etc.). La prise en compte de la dimension temporelle s'avère donc nécessaire pour l'analyse de ce type de données.

Durant les dernières années, toutes ces considérations ont motivé d'importants efforts dans l'analyse des traces des internautes ainsi que l'adaptation des méthodes de classification aux données du Web. Néanmoins, la plupart des méthodes consacrées à l'analyse des données d'usage prend en compte toute la période qui enregistre les traces d'usage. En con séquence, les modèles comportementaux ressortis par ces méthodes sont ceux qui prédominent sur toute la période de temps analysée.

Les comportements minoritaires passibles d'avoir lieu pendant de courtes périodes de temps restent ainsi inaperçus par les méthodes classiques. Dans le cadre du Web, quand un *webmaster* interroge les logs de son site, il souhaite que les résultats proposés en réponse à son interrogation soient fidèles à la période de temps analysée et non au comportement général remarqué tout au long de la période complète analysée. De plus, si une analyse des comportements des internautes ne réalise pas de suivi sur ces comportements au cours du temps, il serait impossible pour le *webmaster* de repérer la période de temps où le(s) possible(s) changement(s) de comportement d'usage ont eu lieu. Pour faire face à ce problème, une solution envisageable serait de définir une stratégie capable de fournir des moyens nécessaires pour que les responsables d'un site Web puissent être avertis lors de l'apparition, la disparition ou le changement des profils de comportement de leurs utilisateurs. Il serait également envisageable que l'administrateur du site puisse mesurer l'impact d'une nouvelle stratégie mise en ligne ainsi

que la popularité des pages à l'aide de l'analyse des traces laissées par les internautes lors des visites.

Motivation

Dans ce mémoire, nous nous intéressons plus particulièrement au Web Usage Mining, qui consiste à analyser, en utilisant les techniques d'ECD, les données issues de l'interaction des utilisateurs avec le Web. Ces données, dites traces de navigation, sont laissées par les internautes lors de leurs surfs, et consignées dans des fichiers de type log. Elles peuvent être relevées, selon la position de la sonde de recueil de données, par les serveurs web, les serveurs proxy, ou les postes des utilisateurs.

Les objectifs dans ce domaine sont aussi divers que la modélisation mathématique du trafic dans le but d'améliorer les systèmes de communication, les programmes et les équipements utilisés sur les réseaux pour les uns; la description et l'interprétation d'un point de vue sociologique des différentes catégories d'usages et d'utilisateurs pour les autres. Les applications, qui en résultent sont nombreuses. D'une part, l'ingénierie des réseaux constitue un débouché évident aux études visant à analyser et à modéliser le trafic. D'autre part, des analyses plus fines, souvent centrées sur des utilisateurs du Web, aboutissent à une meilleure compréhension des situations d'usage et des pratiques des internautes et peuvent être exploitées pour la conception de services plus adaptés aux profils des utilisateurs. Plus concrètement, le WUM est réputé bien accueilli dans la restructuration, l'adaptation et la personnalisation de sites Web, dans le marketing en ligne et les systèmes de recommandation en e-commerce, dans l'analyse de trafic et l'organisation d'architectures réseaux afin d'améliorer leurs performances, dans l'amélioration des systèmes de recherche d'information, et dans plusieurs autres applications basées sur le Web.

A l'instar des données de contenu du Web, les données de son usage ont atteint à leur tour des dimensions colossales. A titre indicatif, les fichiers logs de certains sites populaires collectaient des traces de navigation de l'ordre de giga-octets par heure. Ces volumes énormes de ce type de fichiers constituent une des difficultés majeures à leur manipulation par les algorithmes de fouille, même les plus astucieux entre eux.

De plus, il a été prouvé dans de nombreux travaux que ces données sont dans une forme très brute et inappropriée pour une application directe et fructueuse des techniques d'ECD. Ces aléas inhérents aux données d'usage du Web, qui les rendent incohérentes, erronées et non fiables et limitent leur exploitation naturelle, sont dus à plusieurs facteurs dont les plus prépondérants sont la nature même du protocole http, l'organisation hiérarchique de l'Internet, et les progrès dans la technologie de construction de sites web.

Contribution

Trois principales contributions sont introduites dans notre travail. La première est une partie des définitions générale sur le web et la fouille des données. La seconde est une étude des principaux travaux de prétraitement et d'extraction de connaissances de tous les types de fichiers logs, donnant lieu à un état de l'art dans ce domaine. La troisième est le développement d'un outil simple, mais efficace, de transformer les fichier log au format de base des données

et appliquer les processus de Web Usage Mining en s'appuyant sur la technique des bases de données relationnelles qui se base sur des langage des requêtes déclaratif et de haut niveau qui permet de spécifier les conditions à remplir par les données et restreindre l'analyse sur une partie de la base vérifiant certaines conditions.

Organisation du mémoire

Ce mémoire est organisé comme suit :

Nous commencerons par une introduction générale où on introduit notre thème de recherche ainsi que les objectifs de notre travail. Dans le premier chapitre nous présenterons des généralités sur le web et la fouille des données. Dans la première partie de ce chapitre nous allons présenter des notions sur le web et comment modéliser les données sur le web. La deuxième partie décrit l'application et les techniques sur ces données. Dans Le chapitre 2, Nous aborderons le Web Usage Mining (WUM) et on y montrera ses différents processus. Le chapitre 3 a pour objectif de présenter les travaux relatifs au WUM tout en faisant une étude comparative de ces travaux. Aussi, nous avons présenté les différents travaux dans ce domaine et nous avons tiré les conclusions utiles pour la suite. Dans le chapitre 4, nous avons fait une conception, proposé des diagrammes UML pour notre modèle et nous avons réalisé une application adéquate pour l'analyse de fichiers log. Dans Le chapitre 5, on a procédé à l'expérimentation et à l'évaluation d'usage du logiciel proposé. Enfin, nous avons conclu notre mémoire par une conclusion générale et nous avons exposé les perspectives pour de futurs travaux.

Chapitre 1:Généralités sur Le Web Et la fouille des données

Introduction

Au cours de ces dernières années, avec la croissance exponentielle du nombre des documents en ligne et des nouvelles pages chaque jour, le Web est devenu la principale source d'information. Ce développement a entraîné une croissance rapide de l'activité sur le Web, et une explosion des données résultant de cette activité. En effet, le nombre des utilisateurs d'Internet dans le monde a atteint 3,07 milliards au mois de janvier 2015, Cela représentera 42,4% de la population **[1]**, et le nombre de sites Web a atteint 1 billion au mois de septembre 2014, soit une augmentation de 250 milliards par rapport au l'année 2013 **[2]** selon l'enquête de Netcraft **[1]**. Pour analyser ce nouveau type de données, sont apparues de nouvelles méthodes d'analyse regroupées sous le terme Web Mining dont les trois axes de développement actuels sont le Web Content Mining (WCM) qui s'intéresse à l'analyse du contenu des pages Web, le Web Structure Mining (WSM), qui s'intéresse à l'étude des liens entre les sites Web et le Web Usage Mining (WUM) qui s'intéresse à l'étude de l'usage du Web.

1.1 Le World Wide Web

On appelle «Web» (nom anglais signifiant «toile»), contraction de «World Wide Web» (d'où l'acronyme www), une des possibilités offertes par le réseau Internet de naviguer entre des documents reliés par des liens hypertextes.

Le concept du Web a été mis au point au CERN (Centre Européen de Recherche Nucléaire) en 1991 par une équipe de chercheurs à laquelle appartenaient Tim-Berners Lee, le créateur du concept d'hyperlien **[5]**, considéré aujourd'hui comme le père fondateur du Web.

Le principe du Web repose sur l'utilisation d'hyperliens pour naviguer entre des documents (appelés «pages Web») grâce à un logiciel appelé navigateur (parfois également appelé fureteur ou butineur ou en anglais browser). Une page web est ainsi un simple fichier texte écrit dans un langage de description (appelé HTML), permettant de décrire la mise en page du document et d'inclure des éléments graphiques ou bien des liens vers d'autres documents à l'aide de balises.

Au-delà des liens reliant des documents formatés, le web prend tout son sens avec le protocole HTTP permettant de lier des documents hébergés par des ordinateurs distants (appelés serveurs web, par opposition au client que représente le navigateur). Sur Internet les documents sont ainsi repérés par une adresse unique, appelée URL (Uniform Resource Locator), permettant de localiser une ressource sur n'importe quel serveur du réseau internet **[3].**

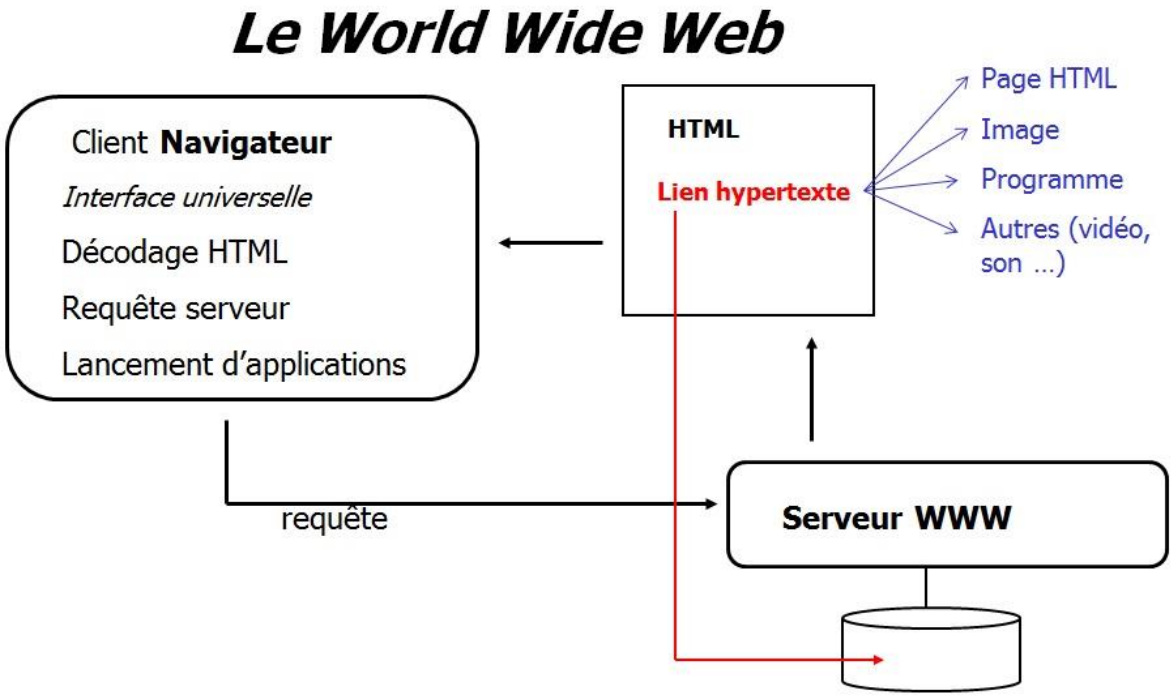


Figure 1 : fonctionnement du WWW

1.2. Quelques définitions (Notion de base)

1.2.1 Protocole http

Le hypertexte Transfer Protocol, en abrégé HTTP, littéralement protocole de transfert hypertexte, est un protocole de communication informatique client-serveur développé pour le World Wide Web. Il est utilisé pour transférer les documents (document HTML, image, feuille de style, etc.) entre le serveur HTTP et le navigateur Web lorsqu'un visiteur consulte un site Web **[4].**

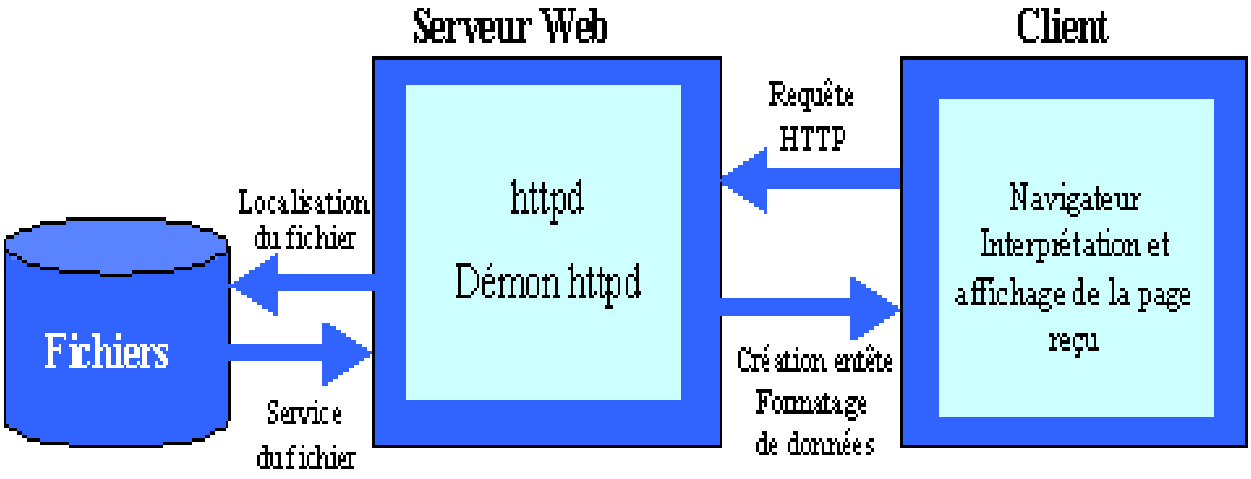


Figure 2 : Fonctionnement du protocole http

1.2.2 Un site web

C'est une ensemble des pages Web et d'éventuelles autres ressources, liées dans une structure cohérente, publiées par un propriétaire (une entreprise, une administration, une association, un particulier, etc.) et hébergées sur un ou plusieurs serveurs Web **[5].**

1.2.3 Une visite web

L'ensemble des clics utilisateur sur un seul serveur Web (ou sur plusieurs lorsque on a fusionné leurs fichiers logs) pendant une session utilisateur. Les clics de l'utilisateur peuvent être décomposés dans plusieurs visites en calculant la distance temporelle entre deux requêtes http consécutives et si cette distance excède un certain seuil une nouvelle visite commence [6].

1.2.4 Une requête http

Une requête HTTP est une demande effectuée par le navigateur WEB (ex: Internet Explorer, Firefox, Mozilla, Safari...) au serveur HTTP lorsqu'il souhaite téléchargerune page WEB **[7].**

1.2.5 Session http

Une session HTTP est une série de requêtes HTTP effectuées entre un client Web et un service Web. Utilisez des sessions HTTP pour enregistrer les données d'un client Web, comme les données d'identification et les variables, pour plusieurs requêtes HTTP. Par exemple, un utilisateur peut entrer un nom d'utilisateur au début d'une session HTTP, et le service Web peut enregistrer ces données pour les requêtes HTTP ultérieures **[8].**

1.2.6 Lien hypertext

Un hyperlien, ou lien hypertexte, ou lien web, ou simplement lien, est une référence dans un système hypertexte permettant de passer automatiquement d'un document consulté à un document lié. Les hyperliens sont notamment utilisés dans le World Wide Web pour permettre le passage d'une page Web à une autre à l'aide d'un clic **[9].**

1.2.7 Serveur

Un serveur informatique, appelé serveur lorsque le contexte s'y prête, est un ordinateur ou un programme informatique qui partage des ressources comme ses périphériques et ses disques durs avec d'autres ordinateurs clients sur un réseau informatique. Il est possible pour un ordinateur d'être client et serveur en même temps **[10].**

1.2.8 Client

On appelle ordinateur client l'ordinateur qui connecté avec l'ordinateur serveur. Une connexion en réseau est donc définie par des relations client/serveur **[10].**

1.2.9 Navigateur Web

Un navigateur Web est un logiciel conçu pour consulter le World Wide Web. Le «navigateur» est donc l'outil de l'internaute, lui permettant de surfer entre les pages web de ses sites préférés. Il s'agit d'un logiciel possédant une interface graphique composée de boutons de navigation,

d'une barre d'adresse, d'une barre d'état (généralement en bas de fenêtre) et dont la majeure partie de la surface sert à afficher les pages web **[11].**

			
FIREFOX	**Internet Explorer**	**SAFARI**	**CHROME**
Logiciel libre disponible pour Windows, mac Os X et Linux. Il est l'un des plus répandus	Développé par Microsoft et disponible pour windows. Les version 9 et 10 n'existent pas pour	Développé par Apple et disponible pour Mac OS X et Windows	Développé par Google, disponible pour Windows, Linux et Mac Os X. C'est le petit

Tab 1 : Les 4 Navigateurs web plus connus

1.3 Fouille de données et extraction de connaissances

L'exploration de données, connue aussi sous l'expression de fouille de données, forage de données, prospection de données, *data mining*, ou encore extraction de connaissances à partir de données, a pour objet l'extraction d'un savoir ou d'une connaissance à partir de grandes quantités de données, par des méthodes automatiques ou semi-automatiques.

Elle se propose d'utiliser un ensemble d'algorithmes issus de disciplines scientifiques diverses telles que les statistiques, l'intelligence artificielle ou l'informatique, pour construire des modèles à partir des données, c'est-à-dire trouver des structures intéressantes ou des motifs selon des critères fixés au préalable, et d'en extraire un maximum de connaissances utiles **[12].**

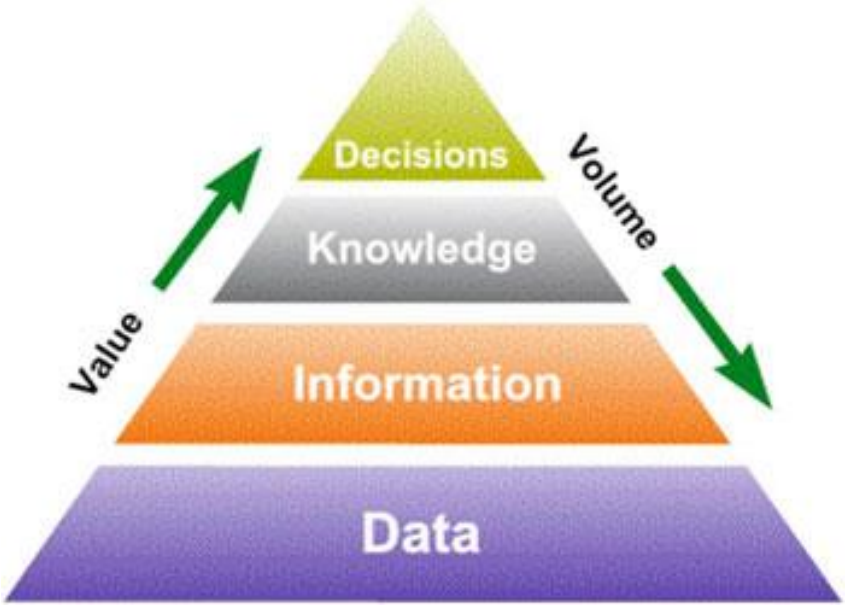


Figure 3 : Hiérarchie du Data Mining

L'ECD est un domaine de recherche pluridisciplinaire au confluent, entre autres disciplines, des mathématiques notamment la statistique et la théorie des probabilités, de l'intelligence artificielle particulièrement l'apprentissage automatique et la reconnaissance des formes, des bases de données et des techniques de visualisation…etc. Actuellement, les projets d'ECD sont peu ou prou standardisés. Ils consistent en un processus à plusieurs phases. Pour aboutir aux connaissances, ce processus comprend des étapes de définition du problème : en délimitant le champ de l'étude et fixant les objectifs à atteindre, de préparation des données : par l'application d'une série d'opérations sur les données telles que le nettoyage, la sélection et la transformation, de fouille proprement dite et en fin d'évaluation et de validation des résultats obtenus. L'étape de fouille de données (ou Data Mining) constitue le cœur du processus d'ECD.

1.4 Web Mining :

Le web mining (WM) est l'application des techniques du data mining pour l'extraction d'informations pertinentes à partir des ressources disponibles sur le Web ; une ressource Web peut être un document ou un service Web.

1.4.1 Les différents domaines du Web Mining :

Le domaine (WM) se divise en trois principales catégories : Web Content mining, Web Structure mining, Web Usage Mining **[6].**

L'analyse du contenu des pages Web (Web Content Mining) :
C'est le processus d'extraction des connaissances à partir du contenu réel des pages Web. Les informations provenant du Web sont stockées dans des bases de données. Ces dernières sont ensuite analysées en utilisant les langages d'interrogations des bases de données et les techniques de fouille de données.

L'analyse des liens entre les pages Web (Web Structure Mining) :
Il s'agit d'une analyse de la structure du Web, de l'architecture et des liens qui existent entre les différents sites. L'analyse des chemins parcourus permet par exemple de déterminer combien de pages consultent les internautes en moyenne et ainsi d'adapter l'arborescence du site pour que les pages les plus recherchées soient dans les premières pages du site. De même, la recherche des associations entre les pages consultées permet d'améliorer l'ergonomie du site par création de nouveaux liens.

L'analyse de l'usage des pages Web (Web Usage Mining) :
Cette dernière branche du Web Mining consiste à analyser le comportement de l'utilisateur à travers l'analyse de son interaction avec le site Web. Cette analyse est notamment centrée sur l'ensemble des clics effectués par l'utilisateur lors d'une visite au site. L'intéret est d'enrichir les sources de données utilisateur (bases de données clients, bases marketing, etc.) à partir des connaissances extraites des données brutes du click des utilisateurs et ce afin d'affiner les profils utilisateur et les modèles comportementaux.

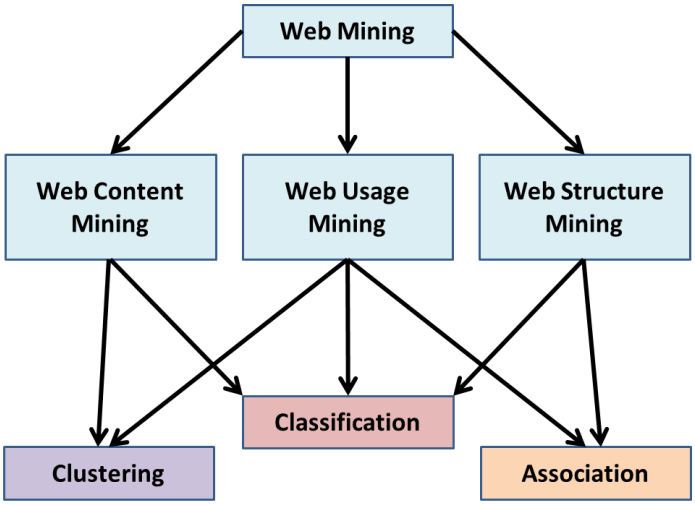


Figure 4 : La relation générale entre les catégories du web mining

1.4.2 Les opérations de base :

Clustering :

Le partitionnement de données (ou data clustering en anglais) est une des méthodes statistiques d'analyse des données. Elle vise à diviser un ensemble de données en différents « paquets » homogènes, en ce sens que les données de chaque sous-ensemble partagent des caractéristiques communes, qui correspondent le plus souvent à des critères de proximité (similarité informatique) que l'on définit en introduisant des mesures et classes de distance entre objets.

Pour obtenir un bon partitionnement, il convient d'à la fois :

- minimiser l'inertie intra-classe pour obtenir des grappes (cluster en anglais) les plus homogènes possibles.
- maximiser l'inertie inter-classe afin d'obtenir des sous-ensembles bien différenciés.

Classification :

Le processus de classification se produit comme suit :

- On donne une collection d'enregistrements, chaque enregistrement contient un nombre d'attributs dont un d'eux sera considérer comme classe
- On trouve un 'modèle' pour l'attribut classe comme une fonction des valeurs des autres attributs
- But : les enregistrements nouveaux doivent être assignés à la classe la plus adéquate.

Les règles d'association

C'est l'approche automatique pour découvrir des relations / corrélations intéressantes entre des objets.

Règles de la forme: $X \Rightarrow Y$ [support, confidence]

- X et Y peuvent être composés de conjonctions
- Support P(X ⇒ Y) = P(X et Y)
- Confidence P(X ⇒ Y) = P(Y | X) = P(X et Y)/P(X)

Applications:

- Détection des fraudes
- Gestion des stocks

$$X \rightarrow Y, \text{ où } X \in T,\ Y \in T, \text{ et } X \cap Y = \emptyset$$

Figure 5 : La forme d'une règle d'association

1.4.3 Processus de Web Mining

Le processus du Web Mining se déroule en trois étapes :

- Collecte des données sur l'utilisateur
- Utilisation de ces données à des fins de personnalisation
- Présentation à l'utilisateur d'un contenu ciblé

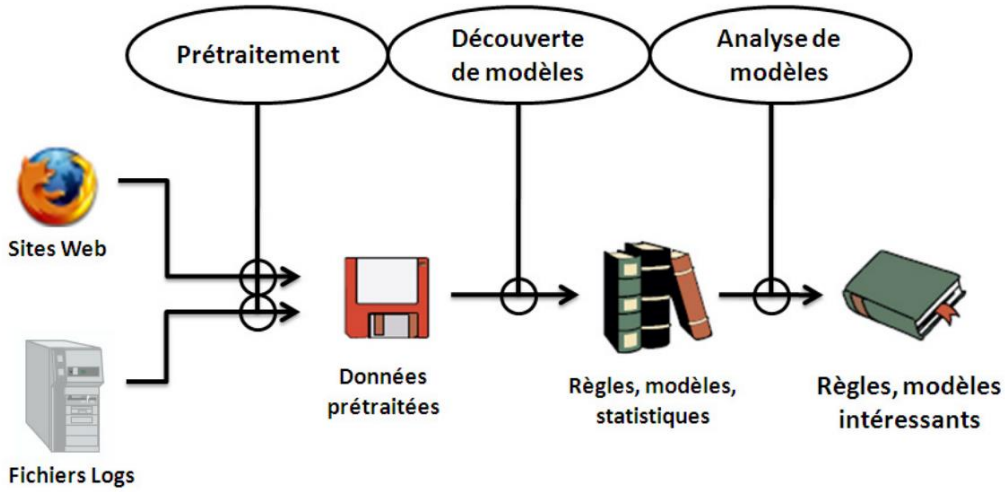


Figure 6 : Processus de web mining

Conclusion

Ce premier chapitre a servi d'introduction au domaine lié à notre étude .Nous avons défini les notions de base notions de base du World Wide Web et Data Mining, de plus, nous avons eu une idée sur la terminologie utilisée dans ces domaines. Ce qui nous permettra, par la suite, à comprendre les articles et travaux scientifiques du domaine Web Mining.

Chapitre 2: Web Usage Mining

Introduction

L'explosion du nombre de sites webs connectés sur Internet et la croissance accélérée du nombre d'internautes confirment de plus en plus la position du web comme un média de masse. Par conséquent, le problème de « l'audience » des sites web revêt de plus en plus d'importance. L'audience d'un site web est un subtil mélange entre le nombre de visiteurs de ce site(le quantitatif) et l'intérêt qu'ont les internautes à visiter le site(le qualitatif). Dans ce contexte, de nombreux travaux se sont intéressés à étudier les problématiques de l'auto adaptation des sites web et de la classification des internautes. L'application des techniques du data mining au web appelée web data mining [15] est devenue le centre d'intérêt d'un nombre grandissant de chercheurs. Il y a actuellement dans le web data mining trois principales directions de recherche:

(i) recherche d'informations,
(ii) analyse de la structure de liens du web,
(iii) analyse des comportements utilisateurs.

Ce dernier, qui présente apparemment beaucoup plus d'intérêts, est axé sur des techniques qui étudient les modèles de navigation des utilisateurs, permettant de personnaliser la présentation pour l'utilisateur individuel, et d'améliorer la structure du site selon les types d'utilisateurs.

2.1 Fichier Log

C'est un fichier informatique utilisé pour l'exploitation d'un serveur d'hébergement. Ce fichier conserve la trace de toutes les requêtes qui ont été adressées à ce serveur. Chacune des requêtes génère une ligne de codes dans le fichier journal. Ces fichiers sont très utiles pour analyser l'audience d'un site Internet, car ils fournissent des indications précises sur le trafic du site. Yseulys Costes **[13]** explique que : « Les fichiers de logs, encore appelés fichiers de traces ou journaux de connexions, sont des fichiers texte qui stockent, les unes à la suite des autres, les lignes d'informations générées par le serveur […] lorsqu'un utilisateur d'Internet, que l'on appelle alors un client, demande une page qui peut contenir du texte, du son, des images ou encore de la vidéo à un site Web hébergé sur un serveur, on parle alors de requête, une ligne de texte s'inscrit dans les fichiers de logs du serveur » **.**

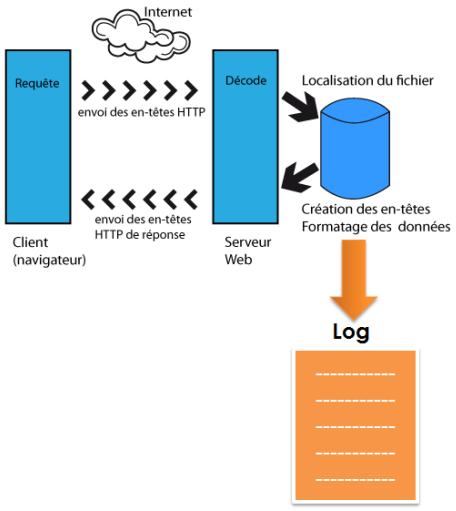


Figure 7 : l'enregistrement des informations dans un fichier log

2.1.1 Les formats du Fichier Log

Extended Log Format (ELF) :

Chaque ligne de ce fichier donne une information sur l'utilisateur, son matériel, la date et l'heure de la requête, la page requise, le statut de la page requise, la page de référence ainsi que quelques informations liées au protocole d'échange de données (figure 1).

Common Log Format (CLF) :

Ceci a le même structure que ELF (Extended Log Format) mais ne contient pas le « referrer » (désignant le navigateur, le système exploitation du l'ordinateur client et ainsi d'autres paramètres éventuelles.

2.1.2 Les composants d'un fichier log

- le nom du domaine ou l'adresse de Protocole Internet (IP) de la machine appelante,
- le nom et le login HTTP de l'utilisateur (en cas d'accès par mot de passe),
- la date et l'heure de la requête,
- la méthode utilisée dans la requête (GET, POST, etc.) et le nom de la ressource Web demandée (l'URL de la page demandée),
- le statut de la requête i.e. le résultat de la requête (succès, échec, erreur, etc.),
- la taille de la page demandée en octets.
- le navigateur et le système exploitation utilisé par le client.

```
161 .31.1 32 .11 6 - - [21 /Dec/2001:08:43:59 -0500] "GET
/images/flagfr.jpg HTTP/1.0" 304 - "-" "Mozilla/4.7 [en] (Win98)
209 .130.181.2 12 - - [21/Dec/2001:08:44:02 -0500] "GET /cs
HTTP/1.1" 301 236 "-" "Mozilla/4 .0 (compatible; MSIE 5.5;
Windows 98)"
```

Figure 8 : Extrait d'un fichier log.

Tout d'abord, il faut remarquer que les lignes arrivent dans un ordre chronologique au gré des différentes requêtes et non pas regroupées par visiteur. Chaque ligne a un format bien défini. La première ligne de la figure 1 servira d'exemple pour commenter les différents blocs de données.

161.31.132.116 : La première série de chiffres est l'adresse de Protocole Internet ou adresse IP. Cette adresse est unique lors d'une connexion. Ceci veut dire que lorsqu'un utilisateur se connecte à l'Internet, cette adresse sera déposée dans tous les fichiers log des sites que celui-ci visitera le temps de sa connexion. Cependant à chaque déconnexion, l'utilisateur perd cette adresse et en obtient une autre lors d'une connexion ultérieure1. Pour l'analyse du trafic, ceci a

deux conséquences importantes. Premièrement, il n'est pas possible de savoir, à partir d'un fichier log standard, si un utilisateur est déjà venu sur le site ou s'il s'agit d'une première visite. Deuxièmement, étant donné que le nombre d'adresses IP disponibles est limité, plusieurs personnes peuvent obtenir successivement la même adresse. En revanche plusieurs personnes ne peuvent pas obtenir la même adresse simultanément. L'adresse IP est unique durant toute la connexion et ne peut être partagée.

[21/Dec/2001:08:42:55 -0500] : Le deuxième groupe de données est relatif à la date et à l'heure de la requête.

GET /home.htm: Le troisième groupe de données concerne la requête. Ici la page requise est la page home.htm.

HTTP/1.0 : correspond au protocole utilisé.

200 : Viennent ensuite des données sur le statut de la page requise (200 pour disponible, 404 pour introuvable …).

4392 : correspond à la taille chargée.

Mozilla/4.7 [en] (Win98) : Le dernier bloc de données renseigne sur la configuration de l'utilisateur. Ici, le visiteur utilise le navigateur Netscape 4.7 version anglaise sous un environnement Windows 98.

2.1.3 Les requêtes principales

Les principales valeurs de types de requêtes sont :

Les requêtes généralement utilisées sont: GET, HEAD, PUT, POST, TRACE et OPTIONS.

La méthode GET : Est la première méthode qui a vu le jour. Elle indique au serveur que nous souhaitons obtenir une représentation de la ressource, en vue de la lire. De par sa nature, cette méthode suggère que la requête soit relativement concise, alors que la réponse, elle, peut être très lourde. En général, cette méthode déséquilibre la bande passante, le client envoyant très peu de données, mais pouvant en recevoir en retour beaucoup (ceci n'est qu'une remarque ayant pour but de justifier qu'aujourd'hui, en grande majorité, un internaute possède une bande passante plus large en recéption qu'en émission).

C'est la méthode utilisée par les navigateurs webs lorsqu'on entre une URL dans la barre d'adresse et qu'on la valide, ou encore lorsqu'on clique sur un lien dans une page web présentée en HTML. Cette méthode représente ainsi un très large pourcentage du total de requêtes web sur un serveur.

Même s'il est possible de passer des paramètres dans la requête, (appelés "query string", les fameuses variables derrière le "?" spéarées par des "&"), GET ne doit jamais faire intervenir une modification de la ressource coté serveur.

La méthode HEAD : Elle est très semblable à GET, si bien que si l'on supporte l'un, en théorie on supporte l'autre (sauf cas très rares). HEAD possède la même signification que GET, à la différence que la réponse ne comportera pas de corps.

Seuls les en-têtes de la réponse seront envoyés au client, sans le corps (souvent lourd) de celle-ci. C'est très pratique pour tester une URL ou encore pour "pinger" un serveur HTTP, car la réponse sera très légère.

On se sert de cette méthode en général pour tester l'existence d'une page (en analysant le code de réponse) : on n'a pas besoin du corps et de la représentation de la réponse, seuls les en-têtes nous importent.

La méthode POST : La méthode POST est utilisée lorsque le client doit faire transiter un nombre conséquent de données, vers un script sur le serveur qui va les traiter, ceci dans le but de créer ou mettre à jour une ressource.

Aujourd'hui, la seule manière pour un navigateur web d'envoyer du POST, est de le faire via un formulaire : une des grosses améliorations introduites par HTTP 1.1 par rapport à HTTP 1.0. Javascript aussi est capable d'envoyer des requêtes de type POST, mais ceci sort du cadre de cet article.

La méthode PUT : Elle permet de télécharger un document, dont le nom est précisé dans l'URI, ou d'effacer un document, toujours si le serveur l'autorise.

La méthode HTTP PUT utilisée par les navigateurs pour stocker des fichiers sur un serveur, les requêtes de type PUT sont beaucoup plus simples que les chargements de fichiers en utilisant le type POST

La méthode TRACE : Le but de cette méthode est assez particulier, et précisons tout de suite qu'elle peut introduire des problèmes de sécurité. La méthode TRACE demande au serveur de renvoyer dans le corps de sa réponse, les en-têtes qu'il a reçu dans sa requête.

C'est un genre de "loopback", on envoie "ping" au serveur, qui doit répondre par "ping".

La méthode OPTIONS : elle permet de demander au serveur les méthodes autorisées pour le document référencé, La méthode OPTIONS permet de lister les méthodes acceptées par le serveur (GET, POST, OPTIONS, TRACE, etc.). L'exemple qui suit montre que la méthode OPTIONS est autorisée par le serveur (code HTTP 200).

2.1.4 Les réponses du serveur http

A toute requête correspond au moins une réponse. Il peut y avoir plusieurs réponses pour une seule requête, comme ce sera le cas dans les transactions partielles **[14].**

100 et 101 : Codes d'information / Information codes : Cette classe de réponse est actuellement réservée pour de futures applications, et consiste en des messages avec une ligne d'état, des champs d'en-têtes éventuels, et terminés par une ligne vide. HTTP/1.0 ne définit actuellement aucun de ces codes, lesquels ne constituent pas une réponse valide à des requêtes HTTP/1.0. Ils restent cependant exploitables à titre expérimental, et dépassent le contexte des présentes spécifications.

200 à 206 : Codes de succès / Succès codes : La requête a abouti. L'information retournée en réponse dépend de la requête émise, comme suit:

GET : Une entité correspondant à l'URI visée par la requête est renvoyée au client.

HEAD : La réponse au client ne doit contenir que les champs d'en-tête à l'exclusion de tout corps d'entité.

POST : Une entité décrivant le résultat de l'action entreprise.

300 à 305 : Codes de re-direction / Redirection codes : Cette classe de messages précise que le client doit provoquer une action complémentaire pour que la requête puisse être conduite jusqu'à sa résolution finale. L'action peut être déclenchée par l'utilisateur final si et seulement si la méthode invoquée était GET ou HEAD. Un client ne peut automatiquement rediriger une requête plus de 5 fois. Il est supposé, si cela arrive, que la re-direction s'effectue selon une boucle infinie.

400 à 417 : Erreurs du client / Client Errors : La classe 4xx de codes d'état est définie pour répondre au cas où il semble que le client ait commis une erreur. Si le client n'a pas encore achevé la transmission de sa requête lorsqu'il reçoit le code 4xx, alors il doit cesser toute transmission. Excepté lorsque ce code répond à une requête de type HEAD, le serveur pourra y inclure une entité explicitant la nature de l'erreur, et indiquant s'il s'agit d'une condition d'erreur temporaire ou permanente. Ces codes sont valides pour tous les types de requête.

500 à 505 : Erreurs du serveur / Server Errors : Les réponses de code d'état 5xx indiquent une situation dans laquelle le serveur sait qu'il est la cause de l'erreur, ou est incapable de fournir le service demandé, bien que la requête ait été correctement formulée. Si le client reçoit cette réponse alors qu'il n'a pas encore terminé d'envoyer des données, il doit cesser immédiatement toute émission vers le serveur. Excepté lorsque la requête invoquée est de type HEAD, le serveur peut inclure une entité décrivant les causes de l'erreur, et s'il s'agit d'une condition permanente ou temporaire. Ces réponses s'appliquent quelque soit la requête, et ne nécessitent pas de champs d'en-tête particuliers.

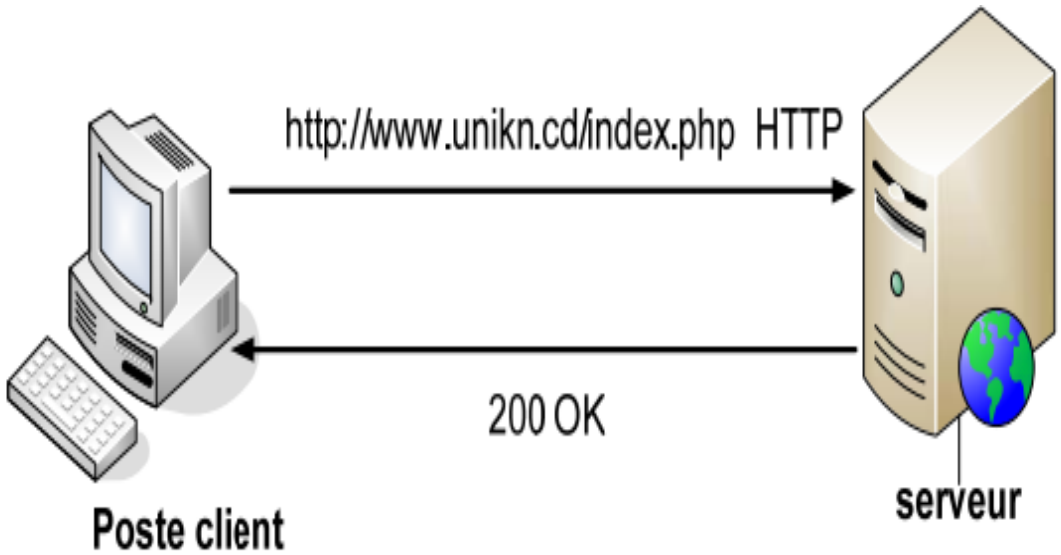


Figure 9 : schéma d'une réponse http

2.2 Le web usage mining (WUM)

La fouille de données d'usage du Web (Web Usage Mining (WUM), en anglais) est définie comme étant l'application du processus d'Extraction des Connaissances à partir de bases de Données (ECD) aux données issues des Fichiers Logs http afin d'extraire des modèles comportementaux d'accès au Web en vue de répondre aux besoins des visiteurs de manière spécifique et adaptée (personnaliser les services) et facilite r la navigation [Tan05]. Comme les analyses se font à partir des Fichiers logs (traces) de serveurs Web, on parle également de Web Log Mining. L'objectif de cette analyse est d'étudier le comportement de l'utilisateur dans son interaction avec le site Web. Elle est centrée sur l'ensemble de clics effectués lors d'une visite au site. On parle aussi d'analyse de ClickStream. L'intérêt de cette analyse est d'enrichir les sources de données utilisateur afin d'affiner les profils utilisateurs et les modèles comportementaux. Le passage en revue des techniques de traitement des Fichiers Logs ainsi que des travaux existants sur l'exploitation de l'information contenue dans les Fichiers logs constituent le corps de ce chapitre.

2.2.1 Processus de WUM :

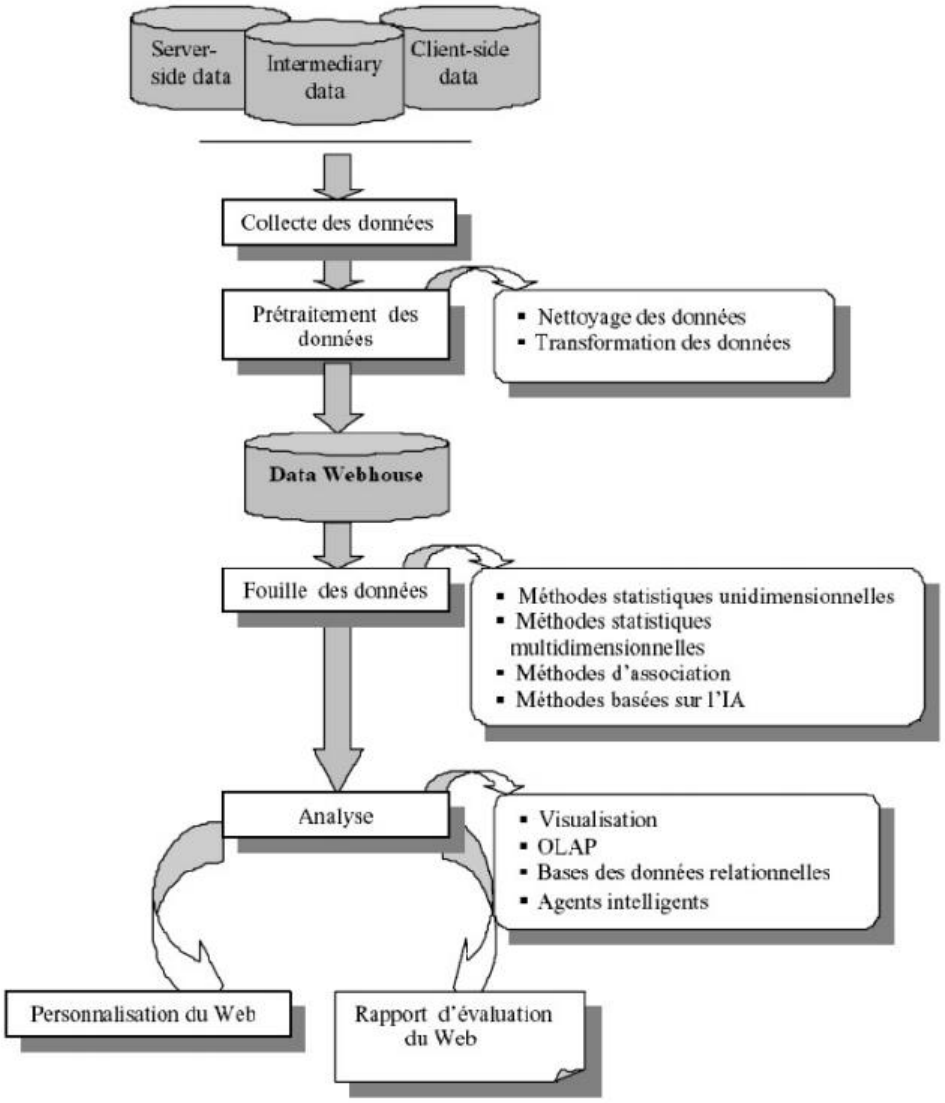


Figure 10 : Processus de web mining

2.2.1.1 Collection des données

La première phase dans le processus du WUM consiste à collecter les données du Web à analyser. Les deux sources principales des données collectées sont les données enregistré es au niveau du serveur et les données enregistrées au niveau du client. Une autre source consiste aux données enregistrées au niveau du serveur Proxy, intermédiaire dans la communication client-serveur.

Données enregistrées au niveau du client : Les données sont collectées au niveau du poste client à travers des agents implémentés en Java ou en Java script. Ces agents sont incorporés dans les pages Web (sous forme d'applets java, par exemple) et utilisés pour une collecte directe des informations à partir du poste client (exemples d'informations : le temps d'accès et d'abandon du site, l'historique de navigation). Une autre technique de collecte des données consiste à utiliser une version modifiée du navigateur [16]. Cette technique permet d'enregistrer les pages Web visitées par un utilisateur ainsi que le temps d'accès et le temps de réponse et les envoyer au serveur. La première méthode permet de collecter des données sur un utilisateur navigant sur un seul site Web. Par contre, un browser modifié permet la collecte des données sur un utilisateur navigant sur plusieurs sites Web. Le problème qui se pose dans le second cas est comment convaincre les internautes d'utilise r c e navigateur modifié dans leurs navigations sachant qu'il peut être considéré comme une menace de leur vie privée [17]. Les informations enregistrées au niveau du poste client sont plus fiables que les informations enregistrées au niveau du serveur puisqu'elles permettent de résoudre le problème du caching et l'identification de sessions [18].

Données enregistrées au niveau du Proxy : Le serveur Proxy joue le rôle d'intermédiaire entre des clients Web et des serveurs Web. C'est également un vaste espace disque servant au stockage des pages Web consultées par les utilisateurs (Web-cache server). En effet, pour toute requête émise sur une page, le Proxy, après consultation de son disque local, transmet la requête au serveur Web si le document n'est pas disponible à son niveau. Une fois l'information retournée par le serveur, le Proxy en effectue une copie locale sur son disque puis la transmet à l'initiateur de la requête. Le serveur Proxy garde la trace de toutes les communications établies dans des fichiers Logs semblables à ceux des serveurs Web. Ces traces peuvent révéler les requêtes http émises par plusieurs clients vers plusieurs serveurs Web et servir ainsi de source de données pour caractériser le comportement de navigation d'un groupe anonyme d'utilisateurs partageant un même serveur Proxy Cependant, les mêmes problèmes cités précédemment (problème du caching et d'allocation des adresses IP) sont présents au niveau du Proxy. [17] [19].

Données enregistrées au niveau du serveur : A cours de sa navigation sur le site, l'utilisateur consulte des pages Web. La demande de ces pages déclenche des requêtes (affichage, téléchargement..) qui sont enregistrées en format texte et stockées de manière standardisée dans un fichier log, appelé log web. Ce fichier est maintenu par le serveur HTTP hébergeant le site. L'enregistrement des données dans les Logs du serveur (server-side Log files) permet d'identifier l'ensemble d'utilisateurs accédant au site Web. De plus, les Logs du serveur fournissent des données sur le contenu, des informations sur la structure et des méta-informations sur les pages Web (taille du fichier, date de la dernière modification) [SCDT00] [19].

2.2.1.2 Prétraitement des données

Le prétraitement de fichiers logs a comme objectif la structuration et l'amélioration de la qualité des données contenues dans les fichiers pour les préparer à une analyse des usages. Cette étape est souvent la plus laborieuse et qui demande le plus de temps à cause de l'absence de structuration et la présence du bruit dans les données brutes d'usage. Les objets à identifier dans un processus de prétraitement de fichiers logs web sont les requêtes effectuées par des utilisateurs humains, les requêtes des robots Web ainsi que les sessions, les navigations et parfois les épisodes [20]. Cette étape dans le processus du WUM, a été traitée dans de nombreux travaux dont les plus importants sont les travaux de Cooley [21] qui propose une méthode complète de traitement des Fichiers logs basée sur les données contenues dans les fichiers logs et la carte du site, les travaux de Berendt et al [22] qui introduit dans leur méthodologie de prétraitement la durée des visites et les travaux de Tanasa et al [20] dont la méthodologie proposée réunit l'ensemble des méthodes et heuristiques classiques et propose une étape de prétraitement avancé. Dans tous ces travaux, le prétraitement des données se décompose en deux phases principales : une phase de nettoyage des données et une phase de transformation.

Nettoyage des données : L'étape du nettoyage consiste à filtrer les données inutiles à travers la suppression des requêtes ne faisant pas l'objet de l'analyse et celle provenant des robots Web. La suppression du premier type de requêtes dépend selon [20] de l'intention de l'analyste. En effet, si son objectif est de trouver les failles de la structure du site Web ou d'ouvrir des liens dynamiques personnalisés aux visiteurs du site Web, la suppression des requêtes auxiliaires comme celles pour les images ou les fichiers multimédia est possible. Par contre, quand l'objectif est le " Web pre-fetching ", il ne faut pas supprimer ces requêtes puisque dans certains cas les images ne sont pas incluses dans les fichiers HTML mais accessibles à travers des liens, ainsi l'affichage de ces images indique une action de l'utilisateur. La suppression du second type de requêtes i.e. les entrées dans le fichier Log produites par les robots Web (WR) permet également de supprimer les sessions non intéressantes. En effet, les WRs suivent automatiquement tous les liens d'une page Web. Il en résulte que le nombre de demandes d'un WR dépasse en général le nombre de demandes d'un utilisateur normal. [20] a utilisé trois heuristiques pour identifier les requêtes et les visites issues des WRs :

- Identifier les adresses IP's qui ont formulé une requête à la page "robots.txt".

- Utiliser des listes des "User agents" connus comme étant des WRs.

- Utiliser un seuil pour la vitesse de navigation " BS (Browsing Speed)", qui représente le rapport entre le nombre de pages consultées pendant une visite de l'utilisateur et la durée de la visite. Si BS est supérieure à deux pages par seconde et la visite dépasse 15 pages, alors la visite a été initiée par un WR.

Transformation des données : Cette phase regroupe plusieurs tâches telles que l'identification des utilisateurs et des sessions et l'identification des visites.

2.2.1.3 Identification des utilisateurs et des sessions

Plusieurs méthodes ont été proposées pour identifier les utilisateurs.

	Identification	Durée
Adresse IP	Groupe d'ordinateurs	Session
Identifiant de session	Individu	Session
Cookie	Ordinateur	Permanent/Session
Mot de passe	Individu	Permanent

Tab 2 : Principales méthodes d'identification des internautes

L'adresse IP : les adresses IP toujours disponibles et ne nécessitant aucun traitement préalable peuvent être utilisées pour identifier les internautes. Cependant, leur utilisation présente principalement deux limites. D'une part, les internautes utilisant un serveur Proxy sont identifiés par l'unique adresse IP de ce serveur. Ainsi, le site visité ne peut déceler s'il s'agit d'un ou de plusieurs visiteurs. D'autre part, l'attribution dynamique des adresses IP's ne permet une identification valable que pour une seule session ininterrompue i.e. si l'internaute interrompt sa visite en se déconnectant un bref instant, son adresse IP sera changée bien qu'il s'agit toujours du même utilisateur.

Les cookies (Client Side Storage) : ces fichiers peuvent contenir des informations telles que la date et l'heure de la visite, la page visitée, un code d'identification du client, etc. Chaque fois que l'utilisateur introduit une URL, le navigateur parcourt les cookies. Si l'un d'entre eux contient cette URL, la partie du cookie contenant les données associées est transférée conjointement à la requête afin de permettre au serveur d'identifier la provenance de cette requête. Cette méthode présente plusieurs avantages. En effet, les cookies permettent une identification s'étalant sur plusieurs sessions. Ils permettent également de stocker plus qu'un simple code d'identification et de collecter et d'enregistrer des informations directement exploitables par le serveur (comme le mot de passe) ; Cependant, l'identification par cookies présente des inconvénients. D'une part, les cookies identifient la machine, et non l'utilisateur ; D'autre part, ils nécessitent l'acceptation de l'utilisateur qui peut à tout moment désactiver leur chargement.

Le mot de passe : pour qu'un serveur puisse identifier un visiteur de manière certaine, l'internaute doit s'identifier lui-même à l'aide d'un pseudonyme (Login) et un mot de passe (Password). Ainsi, le serveur est sûr de l'identité de son visiteur. Cette technique permet d'identifier les internautes de façon permanente et fiable mais elle requiert la participation de l'utilisateur et ne peut être réalisée à son insu. Le serveur devra donc attendre que son visiteur s'enregistre et ne pourra profiter des requêtes effectuées en dehors de l'identification. Pour remédier à cet inconvénient, les mots de passe et les pseudonymes sont souvent enregistrés dans un cookie. L'identification établie lors d'une session ultérieure portera alors sur la machine et non plus sur l'utilisateur.

L'identifiant de session : les identifiants de session permettent à un site entièrement dynamique d'identifier les internautes individuellement. Ils reposent sur la technologie PHP. Cette technique permet d'attacher un identifiant à chacun des liens hypertextes présents sur une page. Lors de la première requête émise, le serveur attribue arbi traitement à cette requête un identifiant de session, la réponse du serveur sera une page préparée dynamiquement. Le serveur peut ainsi insérer l'identifiant de session dans tous les liens hypertextes de cette page.

Lorsque l'utilisateur clique sur l'un de ces liens, sa requête contiendra automatiquement l'identifiant qui lui a été attribué au départ. Cette technique est très fiable mais limite l'identification du visiteur à une seule session.

D'autres méthodes ont été proposées afin de résoudre le problème d'identification de l'utilisateur. Dans [23], la méthode proposée combine l'utilisation de la topologie du site et des informations contenues dans le referrer. Si une requête de page provient de la même adresse IP que les requêtes précédentes sans qu'il y'ait d'hyperliens directs entre les pages demandées, alors l'utilisateur n'est plus le même. Cependant cette méthode n'identifie pas complètement l'utilisateur. [24] emploie une technique différente pour identifier l'utilisateur. Cette technique consiste à inclure, pour chaque utilisateur, un identifiant unique généré par le serveur Web dans les URL's des pages Web du site. Cependant, cette technique nécessite l'intervention de l'internaute qui doit créer un signet, qui inclut l'identifiant comme une partie de l'URL dans l'une des pages afin d'identifier l'utilisateur s'il revient au site.

Ainsi, il s'avère que toutes les techniques proposées présentent des inconvénients dont le plus important est l'introduction dans le domaine privé de l'utilisateur. Cependant, dans [22], les auteurs rapportent que la combinaison de l'adresse IP et le User Agent constitue un bon identificateur de l'utilisateur dans 92% des cas.

	Avantages	Inconvénients
Adresse IP	- Toujours disponible - Aucun traitement préalable	- Identifie un groupe d'ordinateurs - Problème d'attribution dynamique
Identifiant de session	- Grande fiabilité	- Limité aux sites Web dynamiques
Cookie	- Simplicité de mise en oeuvre - stocke plus d'un simple code d'identification	- Identifie la machine - Un seul cookie pour plusieurs utilisateurs - Désactivation ou destruction possible - Inadapté à la mobilité - Introduction dans la vie privée
Mot de passe	- Aucune approximation	- Nécessité de l'intervention humaine

Tab 3 : Comparaison des méthodes d'identification des internautes

Identification des visites ou des navigations

Une fois l'utilisateur identifié par l'une de méthodes décrites ci-dessus, il est possible de reconstituer sa session en regroupant les requêtes contenues dans les fichiers Log et émises par cet utilisateur. Les méthodes d'identification des sessions des utilisateurs peuvent être classifiées en méthodes basées sur le contexte (exemple : accès à des pages de types spécifiques) et méthodes basées sur le temps (exemple : limite seuil de temps de consultation d'une page). Les méthodes basées sur le temps sont les plus couramment utilisées. Elles consistent à considérer que l'ensemble des pages visitées par un utilisateur constitue une visite unique si les pages sont consultées pendant un intervalle de temps ne dépassant pas un certain seuil temporel.

Ce " temps de vue de pages " varie de 25,5 minutes à 24 heures. Le temps de vue de pages couramment utilisé est de 30 minutes. Cependant, l'utilisateur peut passer plus de trente minutes à lire la même page ou quitter son poste pendant un moment et retourner pour consulter la même page. De plus, l'utilisateur du cache peut donner l'impression que la session est finie alors qu'il consulte les pages enregistrées par le cache. [25].

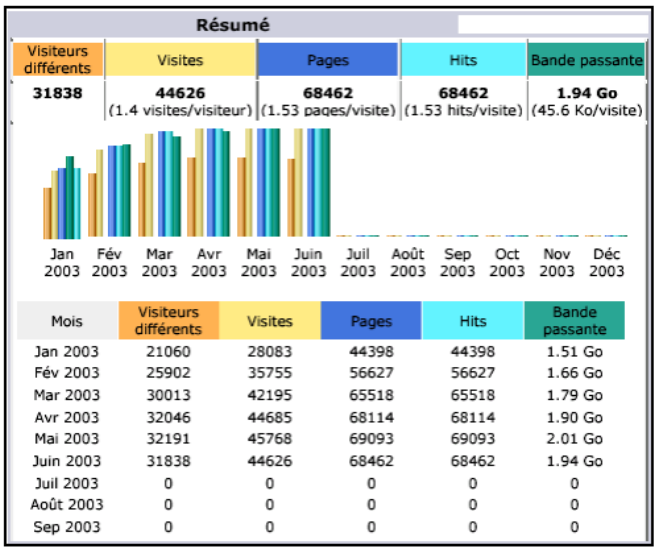
Selon les critères empiriques de Kimball [26], une visite est caractérisée par une série d'enregistrements séquentiellement ordonnés, ayant la même adresse IP et le même nom d'utilisateur, ne présentant pas de rupture de séquence de plus d'une certaine durée.

2.2.2 Fouille de données et analyse des résultats

Cette étape consiste à appliquer des techniques s de fouille des données sur le fichier de sessions ou le fichier de navigations afin d'extraire des patrons d'utilisation du site.il existait plusieurs méthodes de classification utilisées dans la fouille de données, peu de ces méthodes sont appliquées aux données du Web vu la difficulté de les adapter aux particularités de ces données à savoir la taille des tableaux de sessions ou de pages. Ainsi, les méthodes les plus utilisés sont les règles d'association pour la découverte de motifs fréquents de navigation et les motifs séquentiels de navigation. Quant aux méthodes de classification, il est difficile, voire impossible, d'utiliser certaines d'entre elles à cause de la taille gigantesque des données du Web, surtout que la plupart des méthodes de classification retiennent toutes les données en mémoire. Les méthodes de classification les plus utilisées sont les cartes de Kohonen pour la classification des utilisateurs.

2.2.2.1 Analyse statistique

La technique statistique est la méthode la plus commune pour extraire des connaissances à propos des visiteurs d'un site web. En analysant le fichier session, on peut calculer plusieurs types de grandeurs en statistique descriptive (fréquences, moyennes, médiane…) sur les variables telles que les visites des pages, les temps de visites et les longueurs des chemins parcourus. En dépit du manque de «profondeur » de ces analyses, ces types de connaissances peuvent être potentiellement utiles pour l'amélioration de la performance du système par exemple.



Mois	Visiteurs différents	Visites	Pages	Hits	Bande passante
Jan 2003	21060	28083	44398	44398	1.51 Go
Fév 2003	25902	35755	56627	56627	1.66 Go
Mar 2003	30013	42195	65518	65518	1.79 Go
Avr 2003	32046	44685	68114	68114	1.90 Go
Mai 2003	32191	45768	69093	69093	2.01 Go
Juin 2003	31838	44626	68462	68462	1.94 Go
Juil 2003	0	0	0	0	0
Août 2003	0	0	0	0	0
Sep 2003	0	0	0	0	0

Figure 11 : Résumé de statistiques appliquées sur le fichier log du site

2.2.2.2 Règles d'association

La génération de règles d'association peut être utilisée pour renseigner sur les pages qui sont le plus souvent invoquées ensemble dans une même session. Dans le contexte du web usage mining, les règles d'association se réfèrent aux ensembles de pages accédées ensemble avec une valeur support dépassant un certain seuil prédéfini. Il est à noter que ces pages peuvent ne pas être directement reliées par des hyperliens. A côté de l'application dans le marketing, la présence ou l'absence de telles règles peut aider les concepteurs de pages web à restructurer leurs sites.

Rappelons qu'une règle d'association prend la forme d'une implication (si antécédent alors conséquence [Support, Confiance]), par exemple une règle du genre A.html, B.html ⇒ C.html, exprime que si l'utilisateur a visité les pages A et B, alors il est très probable (selon la confiance de la règle) qu'il a visité aussi la page C dans la même session. L'un des problèmes cruciaux dans cette tâche est de garantir une certaine efficacité dans les algorithmes d'extraction.

En effet, le nombre d'items impliqués et les règles à découvrir peut être très élevé, ce qui affectera de manière significative les performances. C'est ainsi que le recours à des techniques de réduction de dimensionnalité a été adopté dans plusieurs travaux. Dans le même ordre d'idées, de multiple projets en WUM ont choisi la structuration des données sous forme d'arbres compacts, pour la représentation et le stockage des itemsets fréquents, afin d'accroître l'efficacité dans le processus de découverte [28].

Par ailleurs, un support minimum global pour l'ensemble de données est toujours fixé au départ dans les algorithmes standard d'extraction de RA. Ceci ne permet de découvrir que les items les plus importants, et ne peut en aucune façon déceler les éléments rares.

Pour résoudre ce problème, une approche de découverte de RA reposant sur l'utilisation de multiples supports minimums a été introduite dans [28]. Elle prend en considération la nature et la fréquence des éléments dans l'ensemble de données, qui sont souvent différents dans beaucoup d'applications réelles.

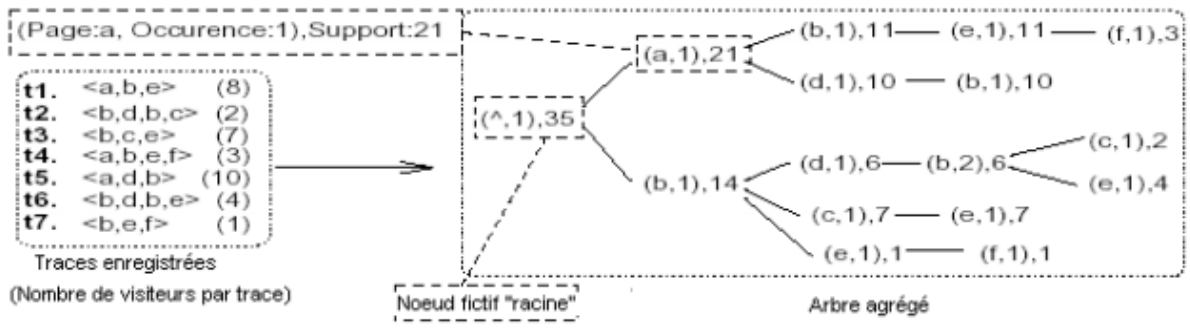


Figure 12 : Un log sous forme d'un arbre agrégé

2.2.2.3 Clustering

Le clustering est une technique regroupant des items ayant des caractéristiques similaires. Dans le domaine du web usage mining, il y a deux types de clusters à découvrir : les clusters d'utilisateurs et les clusters de pages. Le clustering des utilisateurs tend à trouver des groupes d'internautes exhibant des modèles de navigation similaires. D'autre part, le clustering des pages regroupera les pages dont les contenus sont sémantiquement proches.

Figure 13 : L'opération du clustering

2.2.2.4 Classification

La classification est la tâche consistant à mapper un item parmi une ou plusieurs classes prédéfinies. La classification peut être faite en utilisant des algorithmes d'apprentissage inductif comme les arbres de décision. Par exemple, la classification au niveau d'un certain site peut amener à découvrir d'intéressantes règles du genre : 30% des utilisateurs achetant en ligne un CD R&B appartiennent au groupe 18-25 ans et habitent la Côte Ouest.

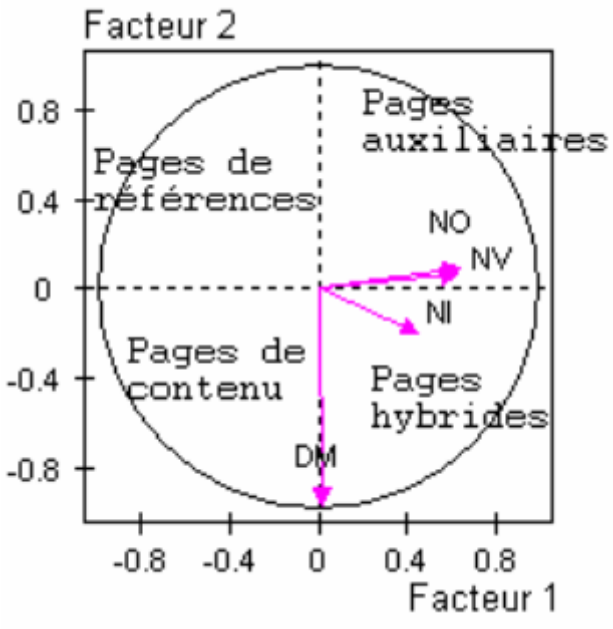


Figure 14 : Classification des Pages.

2.2.2.5 Motifs séquentiels

La technique sur la découverte des motifs séquentiels consiste à trouver des modèles de sessions tels que la présence d'un ensemble d'items soit suivie par un autre item dans un ensemble ordonné de sessions ou d'épisodes. En utilisant cette approche, les webmarketer peuvent prédire les modèles des visites futures qui permettront par exemple de mettre des avertissements visant un certain groupe d'utilisateurs.

Transaction	Fréquence
A, B, E	10
B, D, B, C	4
B, C, E	10
A, B, E, F	6
A, D, B	12
B, D, B, E	8

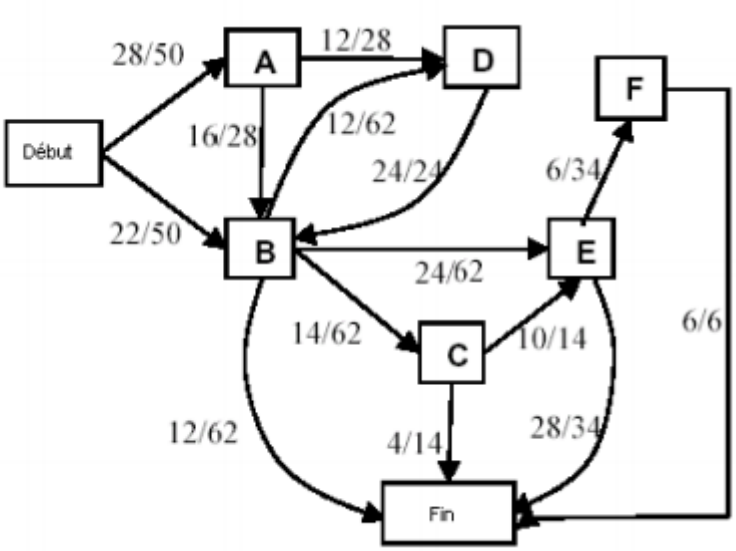


Figure 15 : Exemple d'un motif séquentiel

Web Utilization Miner est un outil de spécification, de découverte et de visualisation de motifs séquentiels d'usage intéressants d'un point vue statistique ou structurel. Le log est stocké, après transformation, dans un arbre agrégé afin d'améliorer le temps de réponse du système, qui implémente MINT un langage de requête de type SQL, afin de découvrir les motifs séquentiels. L'utilisateur spécifiera, comme illustré dans l'exemple de la requête suivante, le format des motifs selon des templates, peuvent inclure des jokers (wildcards), et éventuellement autres contraintes statistiques.

```
Select T From node as XY, template X*Y* as T Where X.url
!="/balk.html" and X.support>40 and Y.url="/kontakt.html"
```

Figure 16 : Requête d'extraction de motifs par MINT dans WUM

2.2.2.6 OLAP

OnLine Analytical Processing dont la source de données est toujours un entrepôt de données (un cube multidimensionnel) est une autre forme d'exploration de données d'usage du web offrant un cadre d'analyse plus intégré et flexible. Dans la majorité des cas, les entrepôts de données en WUM intègrent en plus des données d'usage des données de contenu pour chaque dimension. Les outils OLAP offrent l'avantage de permettre le changement dans le niveau d'agrégation le long de chaque dimension pendant l'analyse. De plus, les résultats de leurs requêtes peuvent constitués l'entrée à une variété d'outils de fouille ou de visualisation de données [28]. WebLogMiner présenté dans [29] est un système de cette famille d'outil. Il permet l'analyse, par les opérations drilldown, rollup, slide et dice, sur plusieurs dimensions telles que : l'url, l'utilisateur, le type de ressource, la taille de ressource, le temps de requête, la durée de consultation, le logiciel client, l'état du serveur, définies dans des hiérarchies de concepts afin de faciliter la spécialisation et la généralisation. Bien que, dans ce papier, les

auteurs se focalisent sur l'analyse des séries temporelles, les données d'usage collectées et fouillées par WebLogMiner ont servi à contrôler la charge et les performances d'un système de e-learning, et d'examiner comment il est exploité par un groupe d'apprenants.

2.2.3 Analyse des modèles

L'analyse des modèles de navigation est la dernière étape dans tout le processus du web usage mining. Il s'agit ici de voir comment exploiter toutes les informations qui ont été obtenues. Depuis 2002, il y a eu plusieurs projets de recherche et produits commerciaux qui ont eu pour but d'analyser les données d'utilisation du web (voir figure suivante).

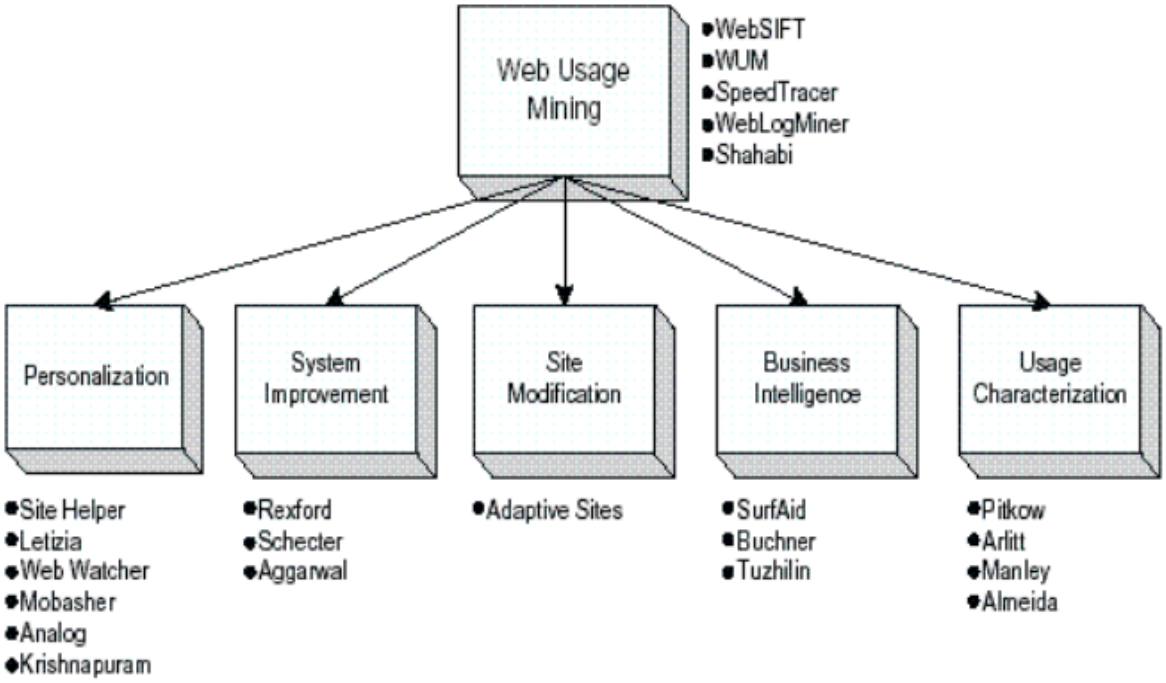


Figure 17 : L'analyse des modèles de navigation.

Cette analyse est une autre tâche non aisée du processus du WUM. En effet, elle repose principalement sur la mesure de l'intérêt des connaissances obtenues, qui n'est pas toujours évidente. La quantification de l'importance d'une règle ou d'un motif trouvé est tributaire de plusieurs facteurs complexes, tels que : l'objectif de l'étude dans un domaine cible (motifs plus populaires pour un site de e-commerce, ceux moins 1 Une sous-classe des grammaires régulières probabilistes, dont les non terminaux correspondent aux pages web et les productions de règles aux hyperliens. 78 normaux dans une analyse de la sécurité d'un système…etc.), l'algorithme d'ECD utilisé, et notamment la composition du groupe chargé de l'analyse [23]. Il est reporté dans [20] que l'étape d'analyse des connaissances découvertes n'a pas toujours été le principal souci des équipes de recherche en WUM, et que le nombre de travaux consacrés à la validation et l'interprétation des résultats reste limité. Toutefois, [23] a tenté de définir les propriétés souhaitables d'un système de mesure de l'importance et de l'intérêt des connaissances en WUM, et a présenté et discuté quelque outils et approches théoriques. Selon cette dernière référence, il existe trois approches relativement simples et communément utilisées dans cette phase : Les langages de requêtes, permettant l'interrogation des motifs extraits.

Dans [25] en est un exemple, Les entrepôts de données, comme le système conçu par [29] permet comme déjà mentionné d'effectuer certain formes d'analyse, Les techniques de visualisation, qui offre un cadre direct et intuitif montrant l'allure des données résultats.

Conclusion

Dans ce chapitre, nous avons présenté une vue détaillé sur fichier log et une méthodologie générale du Web Usage mining pour la division de prétraitement, la division extraction de motifs séquentiels, et l'analyse de résultat.

Chapitre 3: Panorama des travaux Sur le domaine WUM

Introduction

Depuis 1997 il existe plusieurs travaux qui traitent le domaine web mining, dans cette section sont décrits quelques outil et algorithmes et méthodologies de l'état de l'art sur le web usage mining.

3.1 Algorithme Leader (classification des visiteurs)

Dans [30] est décrite une approche pour une classification automatique des visiteurs d'un site Web en fonction de leurs modèles de navigation. Les accès utilisateurs sont examinés afin de découvrir des groupes d'utilisateurs exhibant des besoins similaires en information ; c'est à dire accédant à des pages similaires. Ceci peut aboutir à une meilleure compréhension de la manière dont les utilisateurs visitent le site, et même à une amélioration de l'organisation des liens pour une navigation plus commode. Plus intéressant encore, en fonction de la catégorie à laquelle l'utilisateur appartient, il peut y avoir des suggestions de liens pour lui faciliter sa navigation.

3.1.1 Conception du système :

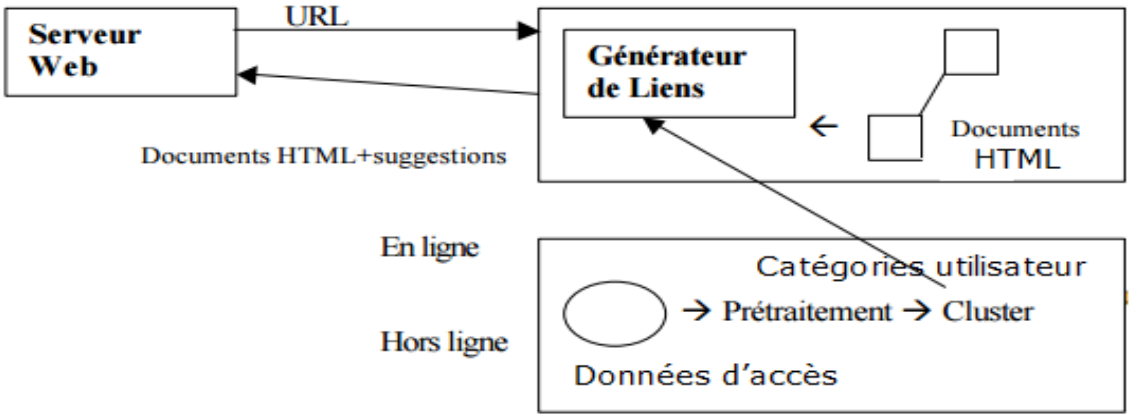


Figure 18 : System basé sur l'algorithme leader

Le système comprend trois principales composantes : un serveur web capable de stocker les informations sur les sessions utilisateurs, un module hors ligne responsable de l'analyse des accès, et un module en ligne se chargeant de la génération de liens dynamiques. Dans le module hors ligne, le préprocesseur extrait périodiquement des informations à partir des fichiers d'accès utilisateurs pour générer des enregistrements de sessions utilisateurs. Un enregistrement est généré pour chaque session. Les enregistrements seront par la suite regroupés en catégories, contenant des sessions similaires. Le module en ligne effectue la génération dynamique de liens. Quand un utilisateur demande une nouvelle page, ce module essaie de classifier la session partielle courante parmi une ou plusieurs catégories déjà obtenues. La catégorie qui correspond le mieux est identifié, et les liens vers les pages inexplorées dans ces catégories sont insérés au début de la page à retourner à l'utilisateur.

3.1.2 Le prétraitement

On peut voir un site web comme étant un ensemble d'items intéressants. Une alternative serait de grouper les pages selon une considération sémantique. Lors d'une session, un utilisateur peut montrer différents degrés d'intérêt dans ces items. S'il y a n items d'intérêts dans un site web, une session utilisateur pourra être représentée comme un vecteur en n dimensions, le i ème

élément étant le poids, ou le degré d'intérêt assigné à l'i ème item. Quand on considère une page web comme un item d'intérêt, on peut lui donner un poids égal au nombre au nombre de fois qu'elle a été visitée, ou à une estimation du temps passé par l'utilisateur sur cette page (peut être normalisé par la longueur de la page), ou au nombre de liens de la page sur lesquels l'utilisateur a cliqué. La principale tâche du prétraitement est de transformer l'information contenue dans les fichiers d'accès en une représentation vectorielle. L'ordre d'accès aux pages est une partie très importante de l'information, mais n'est pas capturé par cette représentation vectorielle.

3.1.3. Le regroupement

Une fois que les sessions sont représentées dans un format vectoriel, nous sommes prêts à appliquer un algorithme de regroupement sur elles. Le but dans ce traitement est de découvrir les groupes de sessions exhibant des intérêts similaires. Cela revient au même en trouvant les groupes de vecteurs « similaires ». La similarité peut être définie de plusieurs sortes. Par exemple deux vecteurs sont similaires si la distance euclidienne entre eux est suffisamment petite, ou si l'angle entre eux est suffisamment petit. Les techniques de regroupement (clustering) sont très étudiées et il y a beaucoup d'algorithmes très connus (leader, k-means, ..). Dans certains algorithmes, un vecteur peut appartenir à plus d'un groupe (cluster). On peut imposer quelques contraintes par soucis de performance (temps de regroupement). La première est que nous devons nous intéresser qu'aux seules sessions accédant à plus d'un certain nombre de pages appelé MinNumPages. Par exemple, il n'est pas utile de regrouper des utilisateurs qui ont juste visité la page d'accueil. Avec cette contrainte, le nombre de sessions pour l'analyse est réduit. En second lieu, nous ne devons considérer que les clusters dépassant une certaine taille appelée MinClusterSize. Ceci élimine les clusters insignifiants et peut améliorer la performance. Cette discussion peut être illustrée avec un algorithme simple, l'algorithme Leader. En entrée nous avons un ensemble V de vecteurs. En sortie nous avons un ensemble C de clusters (un cluster est un groupe de vecteurs). Nous commençons sans cluster et parcourons un à un les vecteurs en entrée. Pour chaque vecteur nous essayons de l'ajouter au cluster le plus connexe tel que la médiane sera inférieure à une distance euclidienne MaxDistance. Si ce cluster n'existe pas, le vecteur forme un nouveau cluster.

```
    Mettre C à vide
  Pour chaque v
    Si (Le cardinal de v est supérieur à MinNumPages) Alors
        Trouver le cluster c de C tel que la distance entre la médiane de c et
v est le minimum (soit d ce minimum) parmi tous les clusters dans C
    Si (La distance d est inférieure à MaxDistance) Alors
        Ajouter v à c
    Sinon
        Ajouter {v} à C
  Pour (chaque c dans C Si la taille de c est inférieure à MinClusterSize) Alors
        Enlever c de C
        Retourner C
```

Cet algorithme a plusieurs inconvénients. Le plus visible est le fait qu'il ne soit pas invariant aux permutations dans les vecteurs. En plus la distance entre le vecteur et la médiane du cluster auquel il appartient est illimitée. Toutefois une force importante de cet algorithme est sa rapidité. Après la découverte des clusters, on peut calculer la médiane de chaque cluster et

caractériser ce qu'il représente. Les pages dominantes sont celles auxquelles sont associés les plus grands poids, et on peut donc dire quelles pages caractérisent le cluster.

3.1.4. Génération dynamiques de liens

Puisqu'un utilisateur navigue à travers les pages d'un site web, nous aurons besoin de suivre ses traces sur les pages visitées, et de le mettre si possible dans une ou plusieurs catégories connues. Nous pourrons donc lui insérer des liens vers des pages intéressantes à suivre. Pour maintenir l'information sur la session utilisateur active, les accès sont temporairement stockés dans un buffer. Quand un utilisateur en ligne demande un nouvel URL, le vecteur est mis à jour. Il faut noter qu'à ce stade, seul le vecteur représente un enregistrement partiel de la session en cours. En classifiant le vecteur de la session partielle, la distance entre la médiane d'un cluster et le vecteur partiel peut ne pas être la bonne mesure, puisque le vecteur partiel a beaucoup plus d'élément nul. Une alternative est de compter le nombre de pages que l'utilisateur a accédé dans chaque catégorie ; si le décompte est supérieur à un certain seuil prédéfini, la catégorie correspondante est trouvée. Après que les catégories correspondantes sont identifiées, on va maintenant s'intéresser aux pages dans cette catégorie. Les pages auxquelles l'utilisateur n'a pas encore accédé, et qui sont inaccessibles à partir de l'URL demandée, sont incluses comme suggestion au début du document html retourné au navigateur.

3.2 Algorithme page gather : génération de la page index

Dans [31] est présentée une approche utilisant l'algorithme page gather (qui met à profit l'expérience d'un groupe d'utilisateurs pour guider le processus d'adaptation), pour tenter de résoudre une partie du problème « index page synthesis ». De quoi s'agit-t-il ? « Page synthesis » est la création automatique d'une page web. Un index page est une page contenant des liens sur un ensemble de pages sémantiquement proches. Etant donné un site web et un ensemble d'accès utilisateurs ; il s'agira dans le problème « index page synthesis »de créer de nouvelles pages index contenant des liens vers des pages apparentées, mais qui n'ont été liées en aucun moment dans le site. Le fichier des logs est un document contenant une entrée par page demandée au serveur web. Chaque requête renseigne au minimum sur son origine (adresse IP), l'URL demandée et l'heure à laquelle la demande a été faite. Deux pages sont dites liées s'il existe un lien de l'une vers l'autre ou s'il existe une autre page se liant au deux. Le problème peut être décomposé en plusieurs sous problèmes :

- Quel sera le contenu de la page index? (en termes de liens)
- Comment ces liens seront-t-ils ordonnés ?
- Comment seront-t-ils étiquetés?
- Quel sera le titre de la page ? Correspondra-t-il à un concept cohérent ?
- Sera-t-il adéquat d'ajouter la page au site? Si oui, où l'insérer ?

L'attention n'est ici portée que sur le premier sous problème, à savoir la génération du contenu de la nouvelle page.

3.2.1 Page Gather

Etant donné un grand nombre d'accès, la tâche est de trouver des ensembles de pages qui tendent à être cooccurrentes. Le clustering est une technique à considérer pour ce genre de problème. Dans le clustering(ou regroupement), les données sont représentées dans un espace à N dimensions (comme des vecteurs de mots par exemple). Techniquement parlant, un cluster

(groupe) est un ensemble connexe de données relativement « distant » des autres clusters. Les algorithmes standards de regroupement partitionnent les données en un ensemble de clusters qui s'excluent l'un à l'autre. Cluster mining(ou analyse de clusters) est une variante du clustering traditionnel qui convient bien dans ce problème. En effet, il s'est avéré plus judicieux d'essayer de trouver un petit nombre de groupements très significatifs, plutôt que de tenter de partitionner la totalité des données. En outre, là où le clustering traditionnel tente de placer chaque donnée (vecteur) dans exactement un cluster, il se peut qu'il y ait dans le cluster mining, un même vecteur appartenant à plusieurs clusters. L'algorithme page gather utilise « cluster mining » pour trouver des ensembles de pages apparentées d'un site web. Par essence, page gather prend en entrée les accès et les transforme en une structure de données convenable pour le cluetering ; il lui applique ensuite cluster mining et donne en sortie des pages index candidates. L'algorithme a cinq principales étapes : Ë traitement des fichiers logs des visiteurs Ë calculer les fréquences de cooccurrence entre les pages et créer une matrice de similitude. Ë créer le graphe correspondant à la matrice, et trouver les cliques maximales(ou les composants connexes) dans le graphe Ë ranger les groupes trouvés et en choisir les meilleurs Ë pour chaque groupe, créer une page web contenant des liens vers les éléments du groupe, la présenter au webmaster pour évaluation.

3.2.2 Détail des étapes

Traitement des fichiers log : Une visite est une séquence ordonnée de pages accédées par un menu utilisateur au cours d'une même session. Cependant, le fichier log est une séquence de pages visitées ou demandées au serveur web. Chaque requête contient l'heure à laquelle elle a été faite, l'URL demandé et l'IP de la machine cliente. On suppose que chaque machine cliente correspond à un seul visiteur.

Calcul des fréquences de cooccurrence entre les pages : Pour chaque couple de pages p1 et p2, on calcule P (p1/p2), la probabilité pour qu'un visiteur visite p1 sachant qu'il a déjà visité p2. La fréquence de cooccurrence entre p1 et p2 est F = min {P (p1/p2), P (p2/p1)}. Le choix du minimum des probabilités conditionnelles a pour but d'éviter de faire une erreur sur une relation asymétrique pour un vrai cas de similitude. Par exemple une page p1 très sollicitée peut se trouver sur le même chemin d'accès qu'une page inconnue p2. Dans ce cas P (p1/p2) sera très grand, nous amenant à penser que ces pages sont très liées. Par contre P (p2/p1) peut être très petit. En rappel il a été dit que le but est de trouver des regroupements de pages apparentées, mais n'ayant en aucun moment été liées dans le site. Pour cette raison, il est à éviter des groupements de pages ayant déjà été liées, en mettant un zéro dans chaque case de la matrice correspondant à un lien existant. Il faut remarquer que cette matrice de similarité peut être vue comme un graphe, permettant d'appliquer des algorithmes de graphes pour la tâche d'identification d'ensembles de pages apparentées. Toutefois, le graphe correspondant à la matrice de similarité peut être complètement(ou presque) connexe. Afin de réduire le nombre de parasites, on met un certain seuil puis on supprime les arêtes ayant des fréquences très petites.

Création du graphe : Un graphe est créé où chaque page est un nœud et chaque cellule non nulle de la matrice est un arc. Ensuite on applique des algorithmes fournissant des informations sur la connexité du graphe (par exemple l'algorithme en temps linéaire pour l'identification des composants connexes). Les informations sur les évaluations des arcs sont ignorées dans cette étape pour des raisons d'efficacité. En créant un graphe clairsemé, et en utilisant un algorithme efficace pour l'analyse de groupe, on peut identifier des clusters très significatifs plus

rapidement que quand on utilise les méthodes traditionnelles de regroupement. Nous pouvons extraire deux genres de sous-graphes menant à deux variantes de l'algorithme page gather.

- PGclique trouve les cliques maximales du graphe. Une clique est un sous graphe dans lequel chaque couple de sommets est lié par un arc. Une clique maximale n'est jamais un sous ensemble d'un autre clique. Pour des raisons d'efficacité, la taille des cliques découvertes est bornée.

- PGcc trouve tous les composants connexes, sous graphes dans lesquels chaque paire de pages est reliée par un chemin(suite d'arcs) entre elles.

Dans la deuxième étape de l'algorithme, on avait mis un seuil à la matrice de similarité. Si ce seuil est élevé, le graphe sera clairsemé, nous trouverons peu de clusters, qui seront de petites tailles et de grande qualité. Si le seuil est bas, nous trouverons des clusters très grands. Il faut noter que PGclique et PGcc ne seront pas utilisés avec les mêmes seuils. Un graphe clairsemé contenant beaucoup de composants connexes peut être trop clairsemé pour contenir des cliques.

Choix de cluster : L'étape précédente peut faire sortir plusieurs clusters, et donc l'idéal serait d'en utiliser que quelques-uns uns. Les clusters trouvés sont évalués et triés par rapport aux moyennes des fréquences de cooccurrence entre les paires de pages dans les clusters. On trouve que PGclique a tendance à découvrir plusieurs clusters similaires. Donc deux techniques sont utilisées pour réduire le nombre de clusters similaires dans le résultat final :

- la réduction des chevauchements : qui consiste à se déplacer sur la liste rangée de cluster en supprimant ceux qui empiètent le plus sur un cluster déjà rencontré.
- la fusion : qui consiste à parcourir la liste rangée de cluster et à fusionner les paires de cluster qui se chevauchent suffisamment.

Toutes ces deux approches requièrent une mesure sur le chevauchement. On utilise le rapport entre les tailles de l'intersection et de l'union de deux clusters. Il faut noter que puisque les composants connexes ne se chevauchent jamais, ni la réduction, sur la fusion n'auront un effet sur PGcc. La réduction et la fusion ont tous deux des avantages et des inconvénients. En général la réduction préserve la cohérence des clusters trouvés (puisqu'elle ne change pas leurs contenus) mais peut rater des pages qui pouvaient être mises dans le même cluster. La fusion, de son côté peut regrouper toutes les pages apparentées, mais au coût de réduire la cohérence des clusters. Par défaut, la réduction est utilisée pour préserver la cohérence des clusters.

Création des pages web : L'algorithme page gather trouve les ensembles de liens candidats, et les présente au webmaster. L'administrateur est incité à accepter ou rejeter le cluster, à le nommer, à effacer tout lien qui lui semble inapproprié. Les liens sont étiquetés avec les titres des pages cibles et ordonnés alphabétiquement. Ceci n'est qu'une vision globale de l'algorithme, on pourrait avoir plusieurs variantes puisque les spécificités des sites peuvent varier d'un site à l'autre.

3.3 Méthodologie CBR

Raisonnement à partir de cas. « To solve a new problem by remembering a previous similar situation and by reusing information and knowledge of that situation » [32].

Le Case-Based Reasoning ou raisonnement à partir de cas est une méthodologie de résolution de problèmes fondée sur la réutilisation des expériences passées, appelées Cas pour la résolution

de nouveaux problèmes. Un cas est composé de deux parties : la partie problème et la partie solution. La partie problème est composée d'un ensemble d'indices qui déterminent dans quelle situation un cas est applicable et utile. Les problèmes résolus sont stockés dans une base d'expériences dite base de cas. Lorsqu'un nouveau problème se présente ce problème est décrit par un cas dit cas cible où seule la partie problème est connue. La méthodologie du RàPC opère alors selon quatre phases séquentielles :

3.3.1 La phase de remémoration (case retrieval)

L'objectif est d'extraire de la base de cas des anciens cas dont la partie problème est similaire au problème à résoudre. Des mesures de similarités sont alors à définir sur les indices constituant la partie problème d'un cas. Les cas extraits de la base sont appelés les cas sources. Les cas sources sont alors passés à la phase d'adaptation.

3.3.2. La phase de réutilisation (case reuse)

L'objectif de cette phase est de proposer une solution au problème courant(le cas cible) en adaptant les solutions proposées par les cas sources. L'adaptation repose souvent sur l'utilisation de connaissances dans le domaine de l'application. A l'issue de cette phase, une ou plusieurs solutions seront proposées pour le cas cible.

3.3.3 La phase de révision (case revision)

L'objectif de cette phase est de réviser les solutions proposées par la phase précédente en fonction de certaines règles et /ou heuristiques, qui dépendent du domaine de l'application. La phase de révision peut être faite par des experts dans le domaine de l'application ou d'une manière automatique.

3.3.4 La phase d'apprentissage (case retainment-learning)

Cette phase a la charge d'enrichir l'expérience du sytème RàPC en enrichissant la base de cas par les nouveaux problèmes résolus (cas cible auquel on a apporté une solution). En effet les cas résolus peuvent être ajoutés à la base de cas et être utilisés dans des cycles de raisonnement futurs. Cependant, avant d'ajouter ces cas, il faut juger la pertinence de cet ajout. Il faut éviter par exemple d'ajouter des cas redondants ce qui affecter les performances du système en termes de temps et de traitement sans pour autant améliorer la qualité des solutions apportées. Une caractéristique très importante du CBR est son côté apprentissage.

La notion de raisonnement à partir de cas ne dénote pas seulement une méthode particulière de raisonnement, elle dénote aussi une « machine learning » qui permet un apprentissage par mise à jour de la base de cas après qu'un certain problème a été résolu. Quand un problème est bien résolu, l'expérience est retenue afin de résoudre des problèmes similaires dans le futur.

Quand une tentative de résolution d'un problème échoue, les causes de l'échec sont identifiées et mémorisées dans le but d'éviter la même erreur dans le futur. Le raisonnement à partir de cas favorise l'apprentissage à partir d'expériences. Toutefois, l'apprentissage efficace avec cette méthodologie nécessite un ensemble de « bonnes » stratégies afin d'extraire des connaissances pertinentes à partir des expériences, d'intégrer un cas dans une structure de connaissance existante, et d'indexer le cas pour l'égaler à d'autres cas similaires.

3.4 Adaptation structurelle des sites web

COBRA. Un site web est dit site adaptatif s'il peut évoluer automatiquement en fonction de l'intérêt d'un ou plusieurs acteurs concernés par le site. Dans [33] est décrite une approche d'auto adaptation des sites web. L'approche est appelée COBRA, une abréviation de Cbr-based COllaborative Browsing Assistant. Cobra est une approche d'auto adaptation qui applique à la fois une adaptation structurelle et une adaptation de la présentation des sites. En effet l'approche permet de guider les utilisateurs à naviguer dans un site en annotant les liens existant par rapport à leur intérêt pour l'utilisateur courant et en ajoutant des liens de raccourcis d'une manière adaptative. L'approche est orientée utilisateur. En effet, l'approche Cobra consiste à observer le comportement de la navigation d'un utilisateur dans le site. Elle essaie de prédire les prochains mouvements de cet utilisateur en réutilisant des épisodes de navigations passées qui sont similaires au comportement de navigation de l'utilisateur actuel. La méthodologie RàPC est utilisée à cet effet. A l'issue du cycle de raisonnement à partir de cas, l'approche prédit un ensemble de pages susceptibles d'intéresser l'utilisateur. Certaines de ces pages sont directement liées par les liens hypertextes à la page courante ; dans ce cas ces liens seront mis en relief. D'autres pages prédites peuvent ne pas être connectées à la page courante. Dans ce cas des liens dynamiques seront ajoutés à la page courante. La prédiction des pages étant fondée sur l'analyse de similarité entre la navigation courante et les navigations passées des autres utilisateurs, l'approche applique alors une stratégie d'apprentissage à partir de groupe d'utilisateur. L'adaptation est faite en ligne et d'une manière personnelle. Les navigations des utilisateurs dans le site seront enregistrées et stockées dans une base dite base de navigations. Un cas source décrit une expérience de navigation dans une navigation donnée. Avant de décrire le cycle de raisonnement appliqué par l 'approche Cobra il convient de décrire d'abord la structure de sauvegarde de navigations utilisateurs et la structure des cas.

3.4.1 Structure d'une navigation

Une navigation N est décrite par une structure qui comporte les champs suivants :

- Nid : un identificateur qui identifie chaque navigation d'une manière unique.
- ADMIN : un champ qui porte des informations relatives à l'administration de la navigation comme par exemple la date de l'enregistrement de la navigation, la longueur de la navigation, l'adresse IP du client, la date de la dernière utilisation de la navigation par le mécanisme du raisonnement (voir ci-après), etc.
- LIST : est une liste chaînée qui décrit la séquence des pages visitées pendant la navigation. Cette liste est composée d'une séquence de nœuds. Chaque nœud représente une page visitée. Un nœud porte les informations suivantes : 1) l'adresse (l'URL) de la page et 2) une description du contenu de la page. Différentes manières peuvent être utilisées pour décrire le contenu d'une page. A titre d'exemple, le contenu d'une page peut être décrit par un vecteur de mots clés. Deux nœuds (pages) successifs sont connectés par un ou plusieurs arcs. Chaque arc correspond à une hypothèse d'action qui peut justifier la transition entre les deux pages. Quatre types d'action (ou hypothèse d'actions) sont définis :

    1. Suivre le lien de rang i.
    2. Suivre le lien dont l'ancre est A
    3. Revenir n pas dans la navigation courante.

4. Aller vers l'URL U La construction d'une navigation est faite comme suit : l'approche Cobra observe la succession des pages (ou URL) demandées par un utilisateur. Chaque page demandée est représentée par un nœud. Pour chaque page visitée, l'ensemble des liens hypertextes qu'elle contient est extrait. Ces liens sont stockés temporairement dans un tableau de liens qui donne pour chaque lien les informations suivantes : 1) l'ancre du lien et 2) l'URL destination de ce lien. Les liens sont stockés dans le tableau dans l'ordre de leur extraction de la page. Maintenant, lorsque l'utilisateur change de page, l'URL de la page demandée et le contenu du tableau de liens de la page précédente sont comparés. Les deux nœuds qui représentent les deux pages successives P1 et P2 seront reliées par :

- Des actions de types «suivre le lien du rang i » si l'URL du rang i dans le tableau de liens de P1 est égale à l'URL de la page P2. Il est à remarquer que plusieurs actions de ce même type peuvent lier deux pages. Afin de rendre l'approche plus souple face aux modifications des pages du site le rang de liens est décrit d'une manière relative au nombre total de liens dans la page. Ainsi si l'URL demandé correspond à l'URL de lien de rang i et au dernier lien de la page courante ; deux actions seront déduites : « suivre le lien i/n » et suivre le lien n/n » où n est le nombre total de liens contenus dans la page P1.
- Des actions de type « suivre le lien dont l'ancre est A » si l'URL de la page P2 est égal à un URL contenu dans le tableau de liens de la page P1 et dont l'ancre est A. Encore une fois plusieurs actions de ce type peuvent connecter deux nœuds successifs.
- Des actions de type « revenir n pas » si l'URL de la page P2 est la même que l'URL d'une page Pn visitée n pas plus tôt dans la même navigation.
- Une action de type « aller vers l'URL u» où u est l'URL de la page P2. C'est l'action par défaut.

A chaque arc (action) est attribué un facteur de confiance qui mesure le degré de confiance du système dans la capacité à expliquer la transition de P1 à P2. A la construction d'une navigation, à chaque action est associé un degré de confiance égal à 1/NA où NA est le nombre d'actions déduites pour relier les deux nœuds P1 et P2.

3.4.2 Structure d'un cas

Un cas référence une expérience ou comportement dans une navigation donnée. Une structure composée des informations suivantes est proposée :

- **CID :** c'est l'identificateur NID de la navigation à partir de laquelle le cas est extrait.
- **Index :** c'est le rang de la page dans la navigation NID à laquelle le cas réfère. Autrement dit c'est l'instant dans la navigation qui détermine l'expérience de la navigation capturée par le cas.
- **Historique :** c'est une séquence d'au plus p pages (nœuds) qui précèdent immédiatement la page du rang INDEX dans la navigation NID.
- **ACTIONS :** c'est l'ensemble des actions qui relient la page du rang INDEX à la page qui la succède dans la navigation NID. Chaque action est associée à son degré de confiance.

Le couple INDEX, HISTORIQUE constitue la partie problème du cas. Le champ ACTIONS représente la partie solution.

3.4.3 Phases du raisonnement

Une première phase consiste à élaborer le cas à partir d'une navigation courante, Nc, dans le site. L'élaboration de cas cible est faite comme suite :

- Le CID a la valeur de l'identificateur de la navigation Nc.
- L'index est le rang de la dernière page visitée dans la navigation Nc.
- L'HISTORIQUE est composé de la séquence de p pages (nœuds) visitées avant la page INDEX.
- La partie solution est vide.

Une fois le cas cible élaboré, le cycle de raisonnement peut commencer en appliquant les quatre phases du cycle RàPC :

3.4.4 Phase de remémoration

Une opération de recherche est lancée dans la base de cas pour retrouver des cas dont le champ HISTORIQUE est similaire à l'historique du cas cible. La similarité est mesurée par une fonction d'agrégation de similarité entre les pages (nœuds) qui constituent les deux historiques. L'ordre des pages est pris en compte. La similarité entre deux pages est elle-même une agrégation entre la similarité sur le contenu des pages et la similarité entre les URL des pages. A l'issue de cette phase le cas dont la similarité avec le cas cible dépasse un certain seuil Sr est retenu. Les K cas les plus similaires sont passés à la phase de la réutilisation.

3.4.5 Phase de réutilisation

Chaque cas source, Csi, retenu par la phase précédente propose un ensemble d'actions A(Csi). L'approche Cobra utilise ces ensembles d'actions de la manière suivante pour prédire le prochain mouvement de l'utilisateur : chacune des actions proposées est évaluée dans le contexte de la navigation courante Nc. L'évaluation d'une action permet de désigner une URL. Par exemple étant donné une action a de type « suivre le lien n/n », l'évaluation de cette action dans le contexte de la navigation Nc donne l'URL destination du dernier lien hypertexte contenu dans la page courante dans la navigation Nc. Remarquer que l'évaluation des actions différentes peut donner la même URL en résultat. A chaque URL ainsi calculé est associé un facteur de confiance qui est fonction du nombre des actions dont l'évaluation a donné comme résultat cet URL, le degré de similarité entre le cas cible, le degré de similarité entre le cas cible et le cas source qui propose l'action et le degré de confiance dans l'action. Les différents URL obtenus après évaluation de toutes les actions proposées par les cas sélectionnés par la phase de remémoration seront triés en fonction de la valeur de leur facteur de confiance. Les URL dont le facteur de confiance dépasse un certain seuil F sont retenues. Ces URL constituent l'ensemble des URL prédits par le système. Si un URL prédit désigne une page liée par un lien hypertexte à la page courante dans Nc, alors ce lien va être mis en relief, sinon, l'URL est ajoutée dans une section spéciale de liens recommandés.

3.4.6 Phase de révision

Après avoir calculé les URL prédits, l'approche observe si l'utilisateur suit l'une des prédictions ou non. Dans le cas où une prédiction aurait été confirmée par l'utilisateur le degré de confiance dans les actions proposées par les cas sources retenus par la phase de remémoration sera augmenté. Les confiances dans les actions qui ont donné de fausses prédictions seront diminuées.

3.4.7 Phase de l'apprentissage

A la fin de chaque navigation l'approche ajoute la navigation courante à la base de navigation rendant ainsi possible de réutiliser les expériences de navigation contenues dans Nc pour des problèmes futurs. Des mesures d'évaluation de l'intérêt de l'ajout d'une navigation sont utilisées par l'approche mais elles ne sont pas présentées ici.

3.5 Autres Travaux

3.5.1 Identification de l'utilisateur (sessionizing)

Dans [23] méthodes présentées pour identification de l'utilisateur, sessionizing (c.-à-construire ou reconstruire des séances), la page vue d'identification, complétion du chemin, et l'identification épisode. Toutefois, certaines des heuristiques proposées ne sont pas appropriées pour les grandes, les sites Web les plus complexes.

Par exemple, ils proposent d'utiliser la topologie du site en collaboration avec le fichier ECLF pour ce qu'ils appellent l'identification de l'utilisateur "(dans notre recherche, cette question est traitée en vertu de la session utilisateur terme"). L'heuristique proposée vise à établir une distinction entre les utilisateurs avec la même adresse IP en vérifiant chaque page demandée dans un ordre chronologique. Si une page demandée n'est pas visée par n'importe quelle page précédente demandé, alors il appartient à une nouvelle session utilisateur. L'inconvénient de cette approche est qu'elle ne considère que d'une façon de naviguer dans un site Web, en suivant les liens. Toutefois, afin de changer la page en cours, les utilisateurs peuvent, par exemple, tapez la nouvelle adresse URL dans la barre d'adresse (la plupart des navigateurs ont une fonction de saisie semi-automatique qui facilite cette fonction), ou ils peuvent choisir parmi leurs favoris.

En outre, l'heuristique sessionizing est complexe car elle définit les limites de session selon à cinq catégories de pages Web (par exemple \ tête ", \ media", \ navigation ", \ look-up", et la saisie de données \ "). Classification des pages du site Web n'est pas une tâche simple, en raison de leur grand nombre et parce que parfois une page ne rentre pas dans un seul de ces cinq catégories.

3.5.2 SchulWeb

Dans [22], les auteurs utilisent la hiérarchie basée sur les services conceptuels pour la modélisation des fonctions d'interrogation d'un catalogue en ligne (SchulWeb). Ce site Web offre la navigation et des services de recherche pour les écoles à travers le monde. Les auteurs ont voulu évaluer les requêtes plutôt que le contenu des pages Web! Générés, par conséquent, leur service axé sur la hiérarchie conceptuelle. En ce qui concerne l'étape de prétraitement, les auteurs proposent trois heuristiques dont un semblable à la vitesse de navigation". Les deux autres heuristiques se réfèrent au nombre de demandes sans le domaine des référents et aux demandes répétées pour la même ressource à partir du même hôte. Toutefois, nous croyons que le plus tard doivent être testé contre d'autres heuristiques pour valider son efficacité.

3.5.3 Entrepôt de données Log

Dans [34] a également développé un entrepôt de données pour stocker les fichiers journaux sur le Web. Contrairement à notre modèle relationnel, leur modèle ne contient pas d'informations structurées sur l'utilisation (sessions, visites, etc), les utilisateurs ou les variables agrégées. Les objectifs de leur travail était la mise en cache Web, par conséquent, l'entrepôt de données, mis

en œuvre dans Microsoft SQL Server, a été rempli avec des journaux en ligne à partir de plusieurs serveurs Web. Pour les applications Web mise en cache, toutes les demandes présentes dans les journaux Web sont importants et doivent être stockés, par conséquent, l'étape de prétraitement est différent à partir d'outils tressés autres parce que presque toutes les demandes sont conservés. En ce qui concerne l'utilisateur et l'identifiant de session, les auteurs uniquement utiliser la notion de \ utilisateur ", identifiés à l'aide de l'heuristique IP.

3.5.4 Web Miner (WUM)

Dans [25] L'utilisation Web Miner (WUM) outil vise à découvrir des motifs séquentiels qui sont considérés comme «intéressant» à partir d'un point de vue statistique. WUM propose d'extraire des motifs séquentiels ayant un minimum de soutien et correspondant à un modèle défini par l'utilisateur. Pour ce faire, les séances sont transformées en un arbre agrégé. Chaque nœud de cet arbre est lié à une page de la voie session. Les auteurs proposent ensuite un arbre de préfixes correspondant aux sessions telles que détaillées dans la Figure suivent.

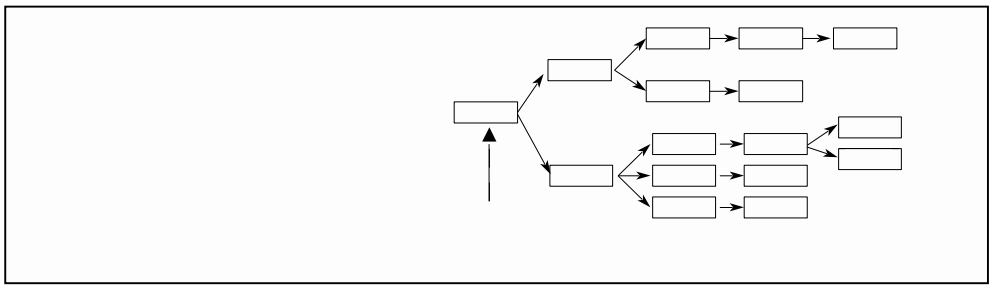


Figure 19 : Exemple d'un arbre agrégé pour sept sessions

L'étape d'extraction est basé sur les modèles demandes composées de pages et des caractères génériques, écrites dans la langue MINT (une langue de type SQL). L'analyste d'exploration de données demandera par exemple, le nombre et les motifs de la forme "a * c", où a et c sont deux pages Web. Pour cette demande, que le début modèles avec un c et se terminant par vous sera retourné. Ainsi, cette méthode appartient à des méthodes basées sur des contraintes et il ne permet pas la découverte de «inconnu» de nouveaux modèles comme nos approches faire.

3.5.5 WebTool

Dans [35], les auteurs proposent un nouveau système de WUM, le WebTool. Ce système prend en compte tous les étapes d'un processus WUM, de la sélection des données à l'affichage des résultats, via la transformation de données et l'extraction des motifs. Le WebTool est basé sur un arbre de préfixes (la PSP, proposé par les auteurs) pour extraire des motifs séquentiels. L'objectif est d'obtenir tous les motifs fréquents en s'appuyant sur un élagage de production méthode (principe Apriori). PSP, comme d'autres méthodes basées sur ce principe, devient moins efficace lorsque le support minimum ou la représentativité de motifs séquentiels est très faible.

L'algorithme de PSP : Le WebTool utilise l'algorithme PSP introduite dans [35]. L'algorithme est basé sur le même algorithme général comme SPG [36], mais il utilise une structure de données améliorée en forme d'arbre pour stocker des séquences candidats. La structure de données de préfixe d'arbre utilisé dans la PSP contient tous les candidats de la manière suivante: n'importe quelle succursale allant de la racine à une feuille représente une séquence candidate,

et en considérant une seule branche, chaque nœud à la profondeur i capte le membre i de la séquence. En outre, avec chaque élément, le support de la séquence de la racine à ce noeud feuille est également stockée. Par exemple, le soutien de la séquence candidate <(10), (20)> est de 2. L'algorithme PSP considère deux cas pour des éléments consécutifs: ils peuvent soit appartenir à la même transaction (ligne en pointillés dans la figure 3.18) ou à des transactions différentes. Unike la PSP, notre algorithme, l'Apriori-TPS, estime que la transaction ne peut avoir qu'un seul élément.

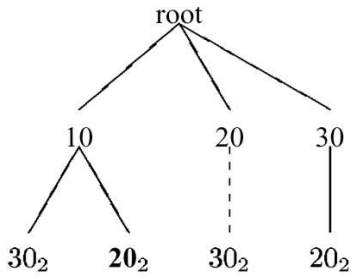


Figure 20 : Arbre PSP Structure des données

3.6 Quelques logiciels pour l'analyse des fichiers log

Il existe plusieurs logiciels pour l'analyse des fichiers log sur le marché :

3.6.1 AWStats

C'est un outil d'analyse statistique d'un site Web. Programmé en perl il y a juste à placer le script dans le répertoire cgi de votre serveur, changer le chemin, et le tour est joué.

Voici quelques-unes de ses fonctionnalités :

- Quand : historique mensuel, hebdomadaire, journalier, horaire ;
- Qui : pays, hôte, dernière visite, robots ;
- Navigation : durée des visites, fichiers vus, page d'entrée et de sortie, système d'exploitation, navigateurs (dont la version) ;
- Provenances : origine, moteurs de recherche, sites, mots recherchés.

Jour	Visites	Pages	Hits	Bande passante
01 Avr 2015	106	272	800	13.34 Mo
02 Avr 2015	98	224	929	9.56 Mo
03 Avr 2015	135	351	668	10.66 Mo
04 Avr 2015	79	228	636	9.16 Mo
05 Avr 2015	24	202	219	2.16 Mo
06 Avr 2015	0	0	0	0
07 Avr 2015	0	0	0	0
Moyenne	73	212	542	7.48 Mo
Total	442	1 277	3 252	44.87 Mo

Figure 21 : Exemple de statistiques Awstats

3.6.2 Webalizer

C'est un logiciel permettant d'analyser l'utilisation des serveurs web en générant, à partir de leurs journaux d'accès (log), des comptes rendus sous forme de pages web. Diffusé sous licence GPL.

Summary by Month

Month	Daily Avg				Monthly Totals					
	Hits	Files	Pages	Visits	Sites	KBytes	Visits	Pages	Files	Hits
May 1999	6377	5570	903	455	10484	884568	14119	28004	172671	197696
Apr 1999	6216	5394	858	419	10087	821968	12594	25758	161844	186504
Mar 1999	7530	6582	1046	499	12128	1052978	15480	32432	204059	233445
Feb 1999	4712	4128	656	321	6629	511793	8048	16419	103203	117816
Jan 1999	4470	3934	607	284	8079	605694	8808	18844	121980	138571
Dec 1998	2998	2673	411	197	6524	410110	6120	12769	82875	92951
Nov 1998	2910	2567	400	192	4260	346705	5588	11627	74468	84403
Oct 1998	3052	2668	457	202	2203	189253	2839	6399	37360	42738
Sep 1998	2072	1826	345	169	3475	314492	5075	10376	54807	62165
Aug 1998	1014	901	211	125	2693	196560	3890	6571	27958	31455
Jul 1998	1484	1325	302	184	4041	298225	5716	9383	41102	46019
Jun 1998	1707	1502	322	222	4809	251502	6675	9687	45077	51227
Totals						5883848	94952	188269	1127404	1284990

Figure 22 : Exemple de statistiques Webalizer

3.6.3 Google Analytics

C'est un service gratuit d'analyse d'audience d'un site Web ou d'applications utilisé par plus de 10 millions de sites. L'outil représente plus de 80 % du marché mondial1. C'est ainsi le service d'analyse de visites de sites web le plus utilisé au monde.

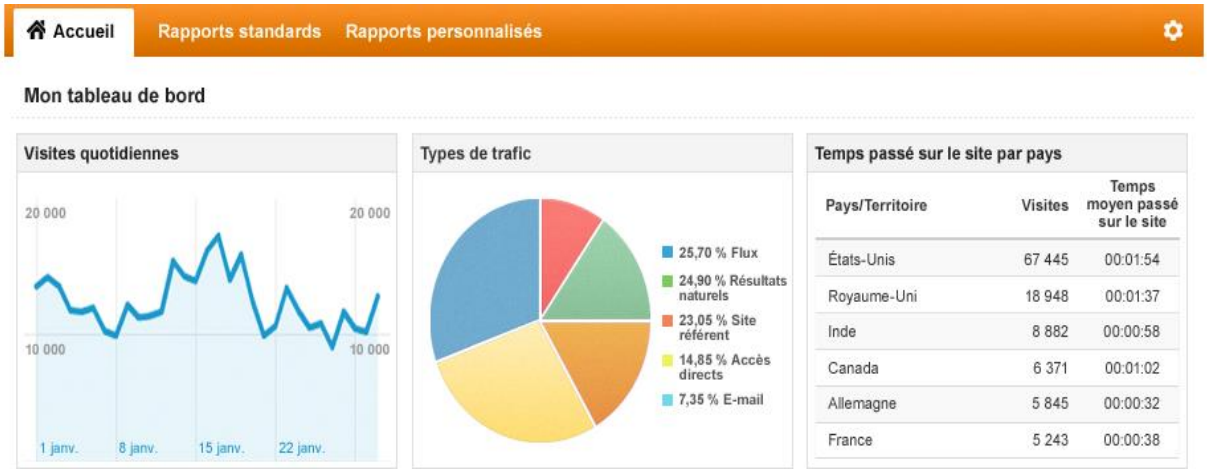


Figure 23 : Exemple de statistiques Google Analytics

Conclusion :

À travers l'étude réalisée dans ce chapitre Nous avons présenté une vue global sur les fichiers log, ainsi des méthodologies et algorithmes relié avec le domaine WUM.
Le choix d'un algorithme ou d'une méthodologie pour l'analyse des utilisations d'un site dépend principalement des spécificités du site et des besoins du webdesigner.

Chapitre 4:Conception Et Réalisation

Introduction

Dans ce chapitre, nous nous intéressons à l'analyse des fichiers « Log » **[13]**, afin de comprendre le comportement des internautes sur un site Web (Site de l'université kasdi Merbah Ouargla www.univ-ouargla.dz).

Le présent chapitre est donc subdivisé en trois sections distinctes :

- La première section présente la conception, par la méthode UML, de la solution mise en place.
- Dans la deuxième section, on trouve les différentes étapes du prétraitement et nettoyage du fichier Log.
- La dernière est consacrée à l'exploration et l'analyse du fichier Log.

4.1 L'apport principal du travail

L'apport de ce travail réside principalement dans les points suivants :

Connaissances sur les visiteurs

Détenir le nombre des visiteurs par heurs des jours ou par un intervalle du temps donné par l'administrateur de site web.

Connaissances sur les pages

- Reconnaître les pages Web les plus et les moins consultées (pages populaires et pages non populaires),
- Reconnaître les pages inexistantes (Page Not Found).

Connaissances sur les fichiers téléchargés :

Reconnaître les fichiers les plus téléchargés par les visiteurs du site web avec le nombre de téléchargement de chaque fichier.

Connaissances sur le trafic internet de site web (bandwidth):

Connaître le trafic Internet de site web par jour ou par un intervalle du temps donné par l'administrateur de site web.

Connaissances sur les attaques des hackers:

Reconnaître les attaques des pirates avec le lien utilisé et avec le type de chaque attaque

Connaissances sur les navigateurs :

Connaître le pourcentage des navigateurs les plus utilisés (Google Chrome, Mozilla Firefox, Safari, Opera).

4.2 Analyse du problème et conception de la solution méthode UML

4.2.1 Le Processus Unifié et UML

Pendant plusieurs décennies, le monde informatique a toujours rêvé d'un processus qui puisse garantir le développement efficace de logiciels de qualité, valable quel que soit la grandeur et la complexité du projet, et présentant de bonnes pratiques adaptées à la méthode en question, surtout que, de nos jours, les logiciels demandés sont de plus en plus imposants et exigeants qu'auparavant.
Le processus unifié semble être la solution idéale pour remédier à l'éternel le problème des développeurs. En effet, il regroupe les activités à mener pour transformer les besoins d'un utilisateur en un système logiciel quel que soit la classe, la taille et le domaine d'application de ce système. Le processus unifié utilise le langage UML (Unified Modeling Language). Ce langage de modélisation est une partie intégrante du processus unifié, ils ont été d'ailleurs développés de concret.

Essayons tout d'abord de présenter UML, car ses diagrammes sont utilisés dans chaque phase et activité du processus unifié, ensuite nous reviendrons sur la présentation du processus unifié.

4.2.2 Présentation d'UML

UML (Unified Modeling Language), se définit comme un langage de modélisation graphique et textuel destiné à comprendre et à définir des besoins, spécifier et documenter des systèmes, esquisser des architectures logicielles, concevoir des solutions et communiquer des points de vue. UML modélise l'ensemble des données et des traitements en élaborant des différents diagrammes. En clair, il ne faut pas designer UML en tant que méthode (Il y manque la démarche) mais plutôt comme une boite d'outils qui sert à améliorer les méthodes de travail.

4.2.3 Présentation d'Edraw Max:

Edraw Max Logiciels scientifiques permettent à des étudiants, des enseignants et des professionnels de créer de façon fiable et de publier les types de diagrammes pour représenter des idées.

C'est un logiciel graphique tout-en-un qui rend simple la création des diagrammes d'aspect professionnel, organigrammes, diagrammes de réseau, les présentations d'affaires, les plans de construction, des dessins de mode, des diagrammes UML, programme de structures, conception de diagrammes de sites Web, diagrammes de l'ingénierie électrique, des cartes de direction, base de données des diagrammes et plus.

Avec les larges bibliothèques de pré-dessiné et plus de 4600 symboles de vecteur, le dessin ne pouvait pas être plus facile! Edraw Max vous permet de créer un large éventail de schémas à l'aide de modèles, des formes et des outils de dessin, tout en travaillant dans une interface intuitive et familière.

4.3 Modalisation de l'application

4.3.1 Diagramme de cas d'utilisation

Dans cette étape, il s'agira de structurer les besoins des utilisateurs et les objectifs correspondants.

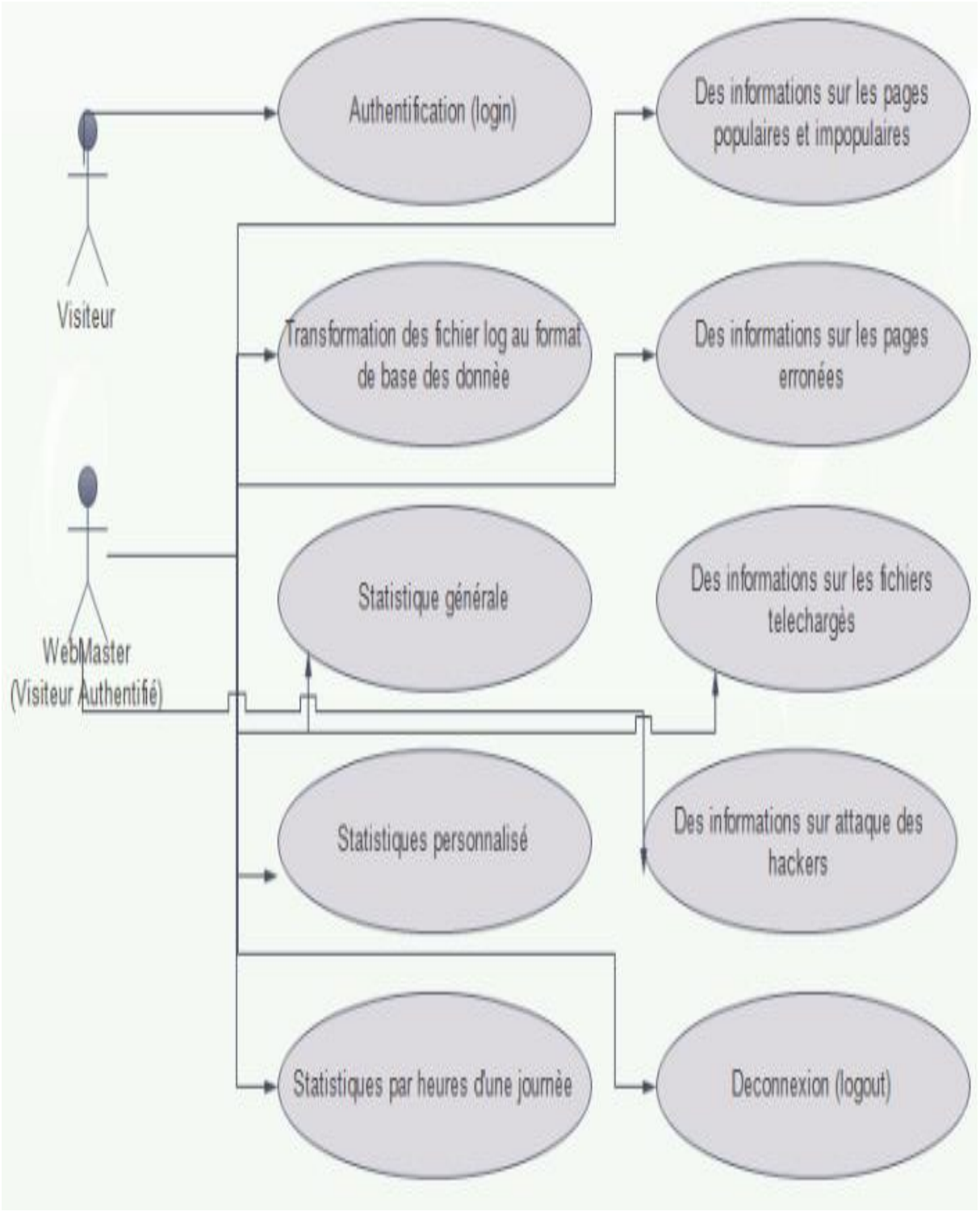


Figure 24 : Digramme de cas d'utilisation

4.3.2 Diagramme de classe

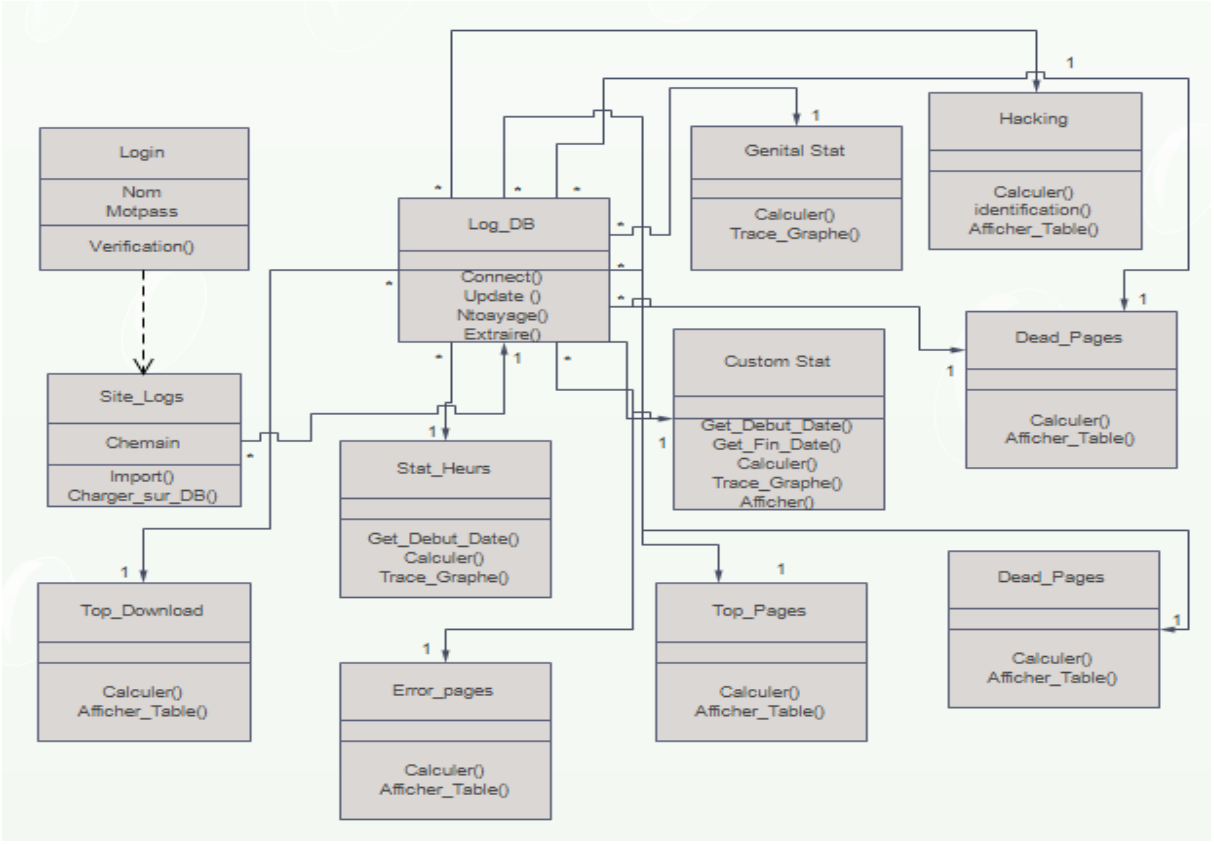


Figure 25 : Diagramme des classes

4.3.3 Diagramme d'état de transition

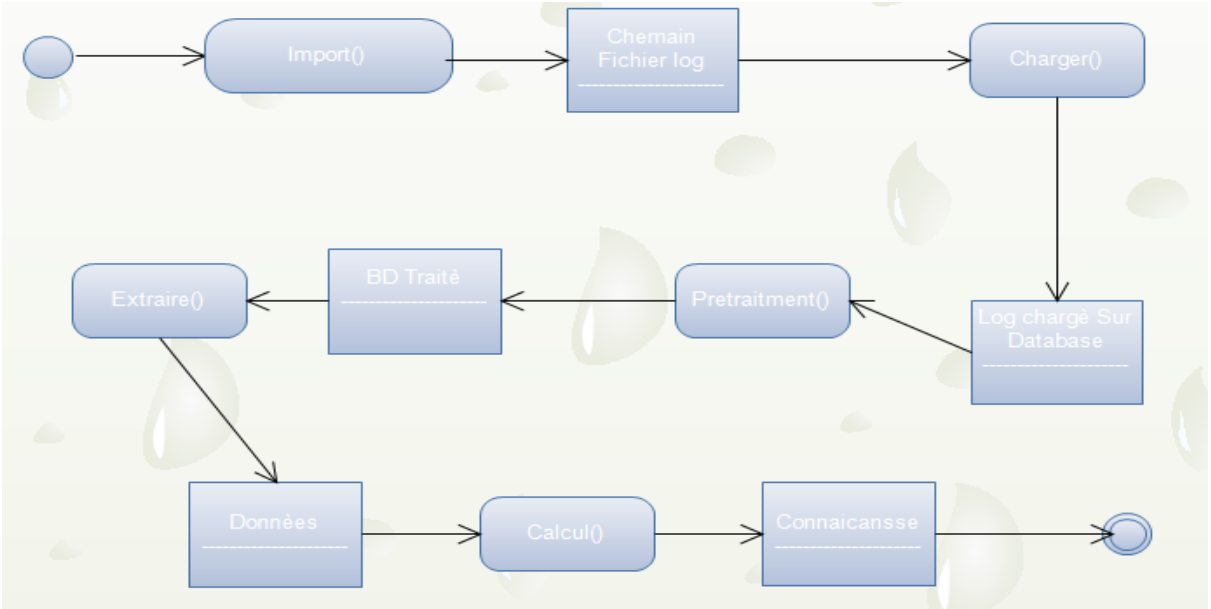


Figure 26 : Diagramme d'état de transition

4.3.4 Diagramme de séquence

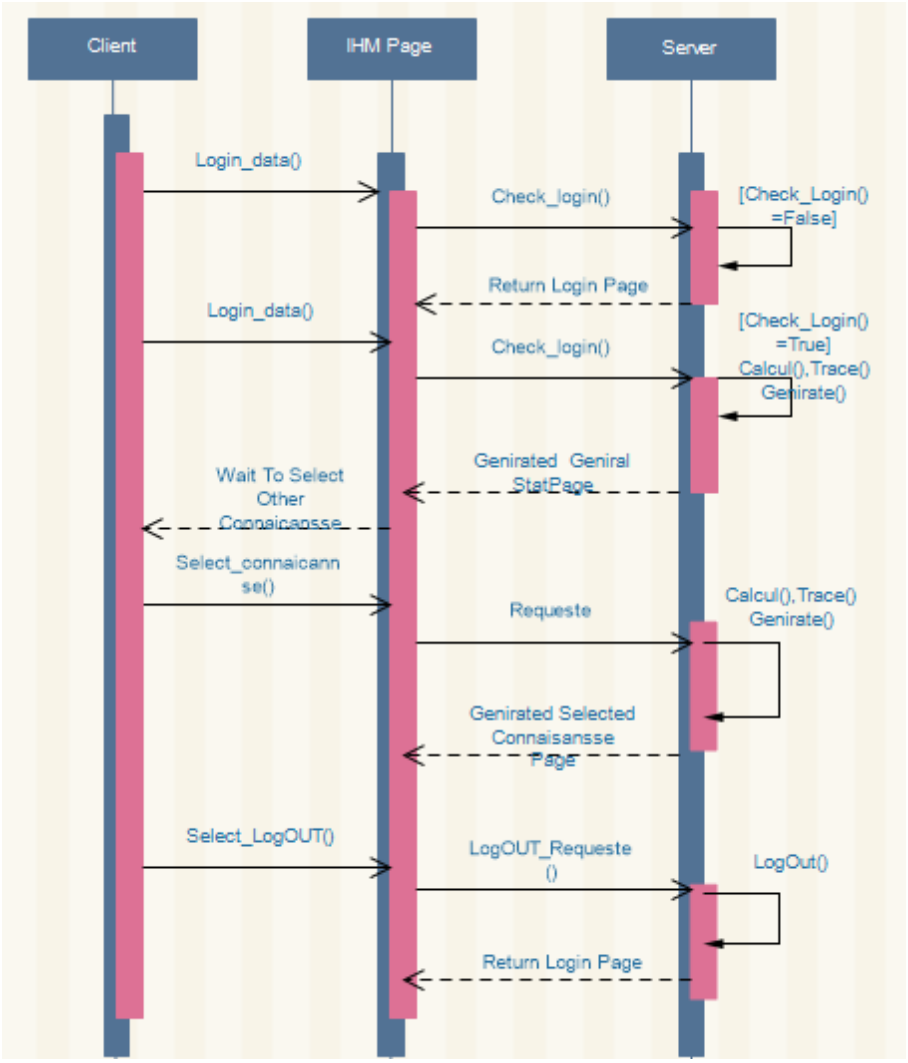


Figure 27 : Diagramme de séquence

4.4 Prétraitement et nettoyage du fichier Log

Les différents champs de ce fichier seront importés dans une base de données

4.4.1 Chargement du fichier Log et transformation en une table d'une BDD

Le fichier Log se transforme en une table composée de plusieurs colonnes. Chaque colonne correspond à un champ spécifique du fichier Log.

- La colonne « hote_client » correspond aux adresses IP des visiteurs.
- La colonne « login_client » correspond au Nom du serveur utilisé par le visiteur.
- La colonne « utilisateur_client » correspond au Nom de l'utilisateur (en cas d'accès par mot de passe).
- La colonne « date_et_heure » correspond à la date d'accès.
- La colonne « methode » correspond à la méthode utilisée (GET/POST).
- La colonne « url_des_pages » correspond au URL demandé.
- La colonne « protocole » correspond au protocole utilisé,

- La colonne « code_de_retour »,
- La colonne « taille_chargé » correspond à la taille chargée, ce fichier doit

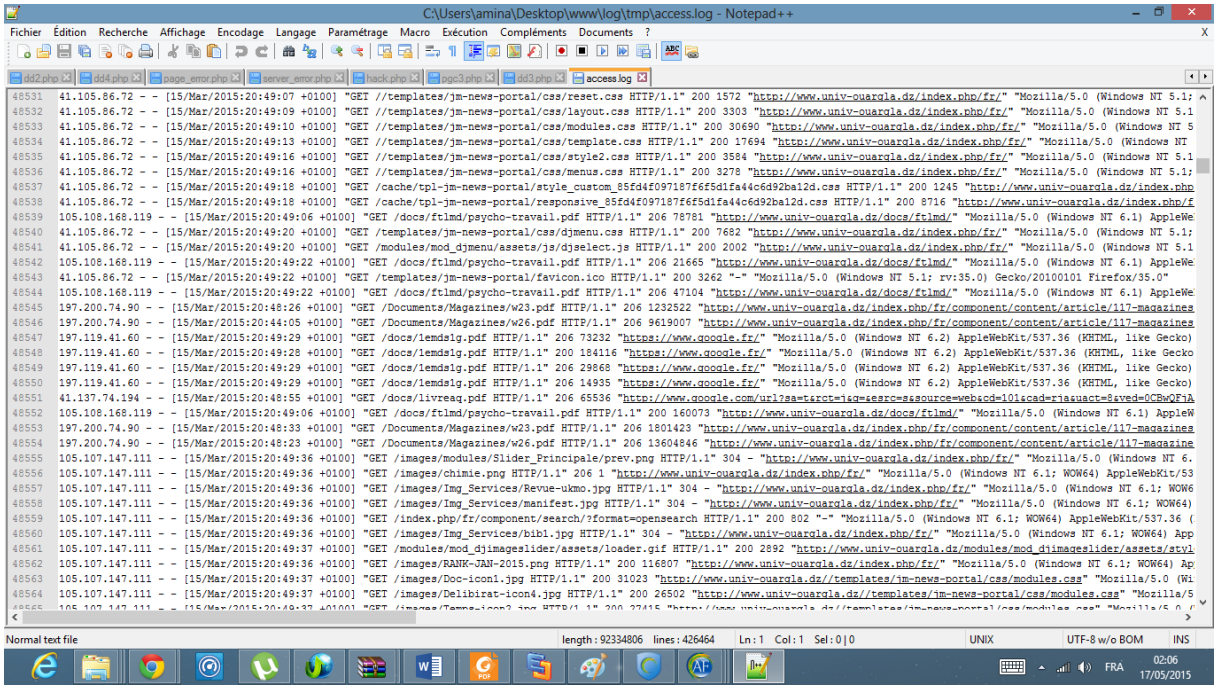


Figure 28 : Fichier Log Brut

Algorithme de transformation :

TQ (!feof(logFile)) {

line = Return_ligne(logFile);

Si(Est_vide(line))

Continue;

Tab_separation = Tab[n];

// Remplire la ligne au tableau avec cette expression régulière

*Remplire_exp('/^(\S+) (\S+) (\S+) \[([^:]+):(\d+:\d+:\d+) ([^\]]+)\] "(\S+) (.+?) (\S+)" (\S+) (\S+) "([^"]+)" "([^"]+)"$/', line, Tab_separation)***[39]**

//Remplire la ligne dans la base des donnèe

Remplire_database(Tab_separation)

---

	remote_host	ident_user	auth_user	time_stamp	request_method	request_url	request_protocol	status	bytes	referer	user_agent	id
Modifier Copier Effacer	5.255.253.227	-	-	2015-03-15 03:16:12	GET	/index.php/fr/accueil/annonces/item/220-fiche-tech...	HTTP/1.1	200	26268	-	Mozilla/5.0 (compatible; YandexBot/3.0; +http://ya...	1
Modifier Copier Effacer	66.249.75.207	-	-	2015-03-15 03:16:16	GET	/Pagesweb/PressUniversitaire/doc/06%20El%20Athar/T...	HTTP/1.1	302	0	-	Mozilla/5.0 (compatible; Googlebot/2.1; +http://ww...	2
Modifier Copier Effacer	157.55.39.191	-	-	2015-03-15 03:16:28	GET	/index.php/home/actualites/307-2014-01-26-08-38-17	HTTP/1.1	303	0	-	Mozilla/5.0 (compatible; bingbot/2.0; +http://www....	3
Modifier Copier Effacer	157.55.39.191	-	-	2015-03-15 03:16:28	GET	/index.php/fr/home/actualites/307-2014-01-26-08-38...	HTTP/1.1	200	69065	-	Mozilla/5.0 (compatible; bingbot/2.0; +http://www....	4
Modifier Copier Effacer	83.229.114.19	-	-	2015-03-15 03:16:32	GET	/index.php/fr/component/search/?format=opensearch	HTTP/1.1	200	802	-	Mozilla/5.0 (Windows NT 6.1; WOW64) AppleWebKit/53...	6
Modifier Copier Effacer	5.255.253.227	-	-	2015-03-15 03:16:32	GET	/index.php/fr/component/k2/item/138-prix-de-la-ban...	HTTP/1.1	200	23783	-	Mozilla/5.0 (compatible; YandexBot/3.0; +http://ya...	7
Modifier Copier Effacer	66.249.75.239	-	-	2015-03-15 03:16:33	GET	/index.php/accueil/presentation-de-l-universite/l-...	HTTP/1.1	303	0	-	Mozilla/5.0 (iPhone; CPU iPhone OS 6_0 like Mac OS...	8
Modifier Copier Effacer	66.249.75.239	-	-	2015-03-15 03:16:33	GET	/index.php/fr/accueil/presentation-de-l-universite...	HTTP/1.1	200	103317	-	Mozilla/5.0 (iPhone; CPU iPhone OS 6_0 like Mac OS...	9
Modifier Copier Effacer	5.255.253.227	-	-	2015-03-15 03:16:52	GET	/index.php/fr/accueil/actualites/634-2014-11-04-11...	HTTP/1.1	200	21564	-	Mozilla/5.0 (compatible; YandexBot/3.0; +http://ya...	11
Modifier Copier Effacer	68.180.228.151	-	-	2015-03-15 03:17:02	GET	/index.php/fr/accueil/annonces/item/55-fiche-de-su...	HTTP/1.1	200	27251	-	Mozilla/5.0 (compatible; Yahoo! Slurp; http://help...	12
Modifier Copier Effacer	5.255.253.227	-	-	2015-03-15 03:17:12	GET	/index.php/fr/component/k2/item/169-ouverture-de-l...	HTTP/1.1	200	27164	-	Mozilla/5.0 (compatible; YandexBot/3.0; +http://ya...	13
Modifier Copier Effacer	157.55.39.190	-	-	2015-03-15 03:17:19	GET	/Pagesweb/siteweb_asi?C=D;O=D	HTTP/1.1	200	3027	-	Mozilla/5.0 (compatible; bingbot/2.0; +http://www....	14
Modifier Copier Effacer	66.249.75.32	-	-	2015-03-15 03:17:29	GET	/index.php/home/actualites/295-2014-01-21-15-09-35...	HTTP/1.1	303	0	-	Mozilla/5.0 (compatible; Googlebot/2.1; +http://ww...	15
Modifier Copier Effacer	66.249.75.32	-	-	2015-03-15 03:17:29	GET	/index.php/fr/home/actualites/295-2014-01-21-15-09...	HTTP/1.1	200	68790	-	Mozilla/5.0 (compatible; Googlebot/2.1; +http://ww...	16
Modifier Copier Effacer	5.255.253.227	-	-	2015-03-15 03:17:32	GET	/index.php/fr/accueil/presentation-de-l-universite...	HTTP/1.1	200	32785	-	Mozilla/5.0 (compatible; YandexBot/3.0; +http://ya...	17
Modifier Copier Effacer	5.255.253.227	-	-	2015-03-15 03:17:53	GET	/pagesweb/PressUniversitaire/doc/05%20Sciences%20s...	HTTP/1.1	302	0	-	Mozilla/5.0 (compatible; YandexBot/3.0; +http://ya...	18
Modifier Copier Effacer	157.55.39.69	-	-	2015-03-15 03:18:11	GET	/index.php/fr/rectorat/pedagogie/95-annonces/269-b...	HTTP/1.1	200	4768	-	Mozilla/5.0 (compatible; bingbot/2.0; +http://www....	19
Modifier Copier Effacer	5.255.253.227	-	-	2015-03-15 03:18:12	GET	/pagesweb/PressUniversitaire/doc/Pubs/Chehma_Synth...	HTTP/1.1	302	0	-	Mozilla/5.0 (compatible; YandexBot/3.0; +http://ya...	20
Modifier Copier Effacer	5.255.253.227	-	-	2015-03-15 03:18:32	GET	/index.php/fr/accueil/actualites/toutes-les-annonc...	HTTP/1.1	200	19705	-	Mozilla/5.0 (compatible; YandexBot/3.0; +http://ya...	21
Modifier Copier Effacer	68.180.228.151	-	-	2015-03-15 03:18:43	GET	/index.php/fr/component/search/?Itemid=1020&task=5...	HTTP/1.1	200	920	-	Mozilla/5.0 (compatible; Yahoo! Slurp; http://help...	22
Modifier Copier Effacer	66.249.75.239	-	-	2015-03-15 03:18:43	GET	/audio_visuel/index.php/photos-universite/image/92...	HTTP/1.1	200	2738376	-	Googlebot-Image/1.0	23
Modifier Copier Effacer	5.255.253.227	-	-	2015-03-15 03:18:52	GET	/index.php/fr/accueil/actualites/531-2014-05-06-13...	HTTP/1.1	200	21408	-	Mozilla/5.0 (compatible; YandexBot/3.0; +http://ya...	24
Modifier Copier Effacer	5.255.253.227	-	-	2015-03-15 03:19:12	GET	/index.php/fr/planning-des-examens/Documents/Annon...	HTTP/1.1	302	0	-	Mozilla/5.0 (compatible; YandexBot/3.0; +http://ya...	25
Modifier Copier Effacer	5.255.253.227	-	-	2015-03-15 03:19:32	GET	/index.php/fr/accueil/actualites/749-2015-03-05-09...	HTTP/1.1	200	21693	-	Mozilla/5.0 (compatible; YandexBot/3.0; +http://ya...	26
Modifier Copier Effacer	157.55.39.36	-	-	2015-03-15 03:19:44	GET	/pagesweb/siteweb_afssi-arc/Siteweb_AFSSI_UOGX_fic...	HTTP/1.1	200	1981	-	Mozilla/5.0 (compatible; bingbot/2.0; +http://www....	27

Figure 29 : Tableau du Base de données après la transformation du fichier log

4.4.2 Nettoyage des données

Les données concernant les pages possédant des graphiques, des images ou des scripts, n'apportent rien à l'analyse. Elles seront donc filtrées et éliminées.

Pour cela on est amené à supprimer de notre base de données les URLs qui ont les formes suivantes :

("delete from TABLOG where url_des_pages like '%.gif%")

("delete from TABLOG where url_des_pages like '%.jpg%")

("delete from TABLOG where url_des_pages like '%.png")

("delete from TABLOG where url_des_pages like '%.ico")

("delete from TABLOG where url_des_pages like '%.css")

("delete from TABLOG where url_des_pages like '%.js")

4.5 Réalisation

Pour l'exploration et l'analyse du fichier Log, une application web responsive **[40]** a été conçue et réalisé «WuStat », dont l'interface est comme suit :

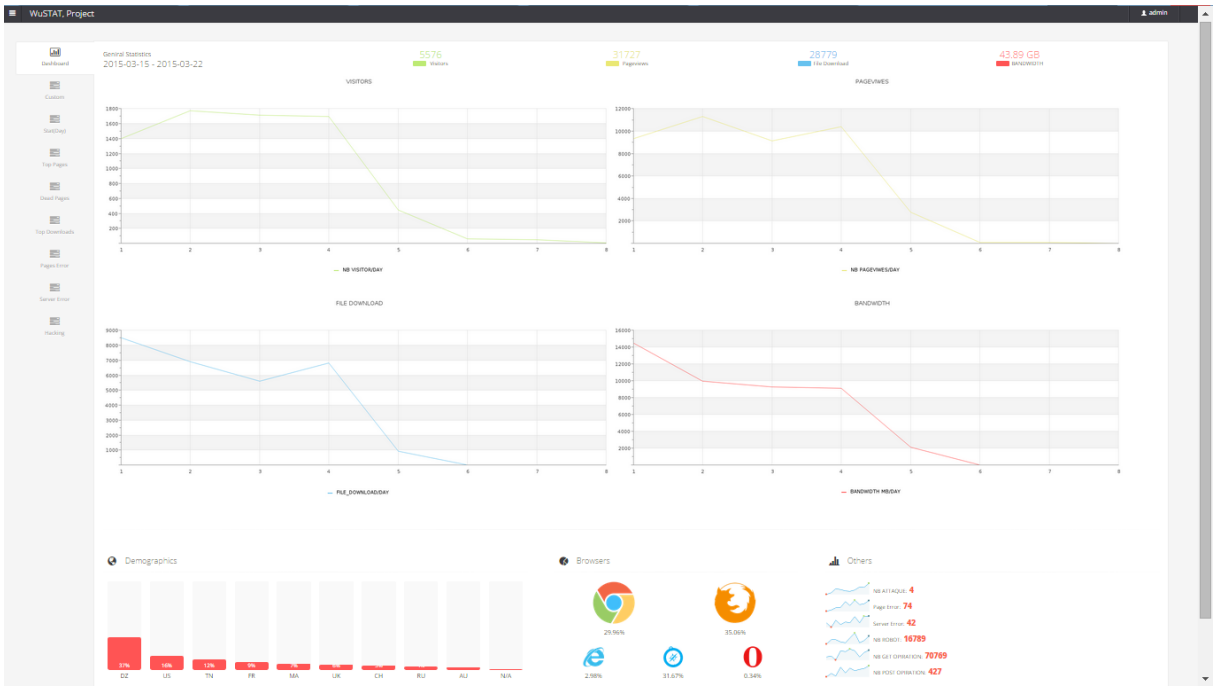


Figure 30 : La page accueille du WuStat

4.5.1 L'accueil de l'application

L'accueil de l'application est composés par :

- Nombre visiteur total
- Nombre des pages total consultés par les visiteurs
- Nombre des fichiers téléchargés
- Trafic internet de site web
- Des graphes totaux sur les (Visiteurs / Jour), (Pages Consultés / Jour), (Fichiers Téléchargés / jour), (Trafic/jours).
- Les payes populaires des visiteurs
- Les pourcentages des navigateurs utilisés par les visiteurs du site web
- Nombre des attaques
- Nombre erreurs des pages
- Nombre erreur du serveur
- Nombre de l'opération GET
- Nombre de l'opération POST

4.5.2 Analyse personnalisé (Custom Stat)

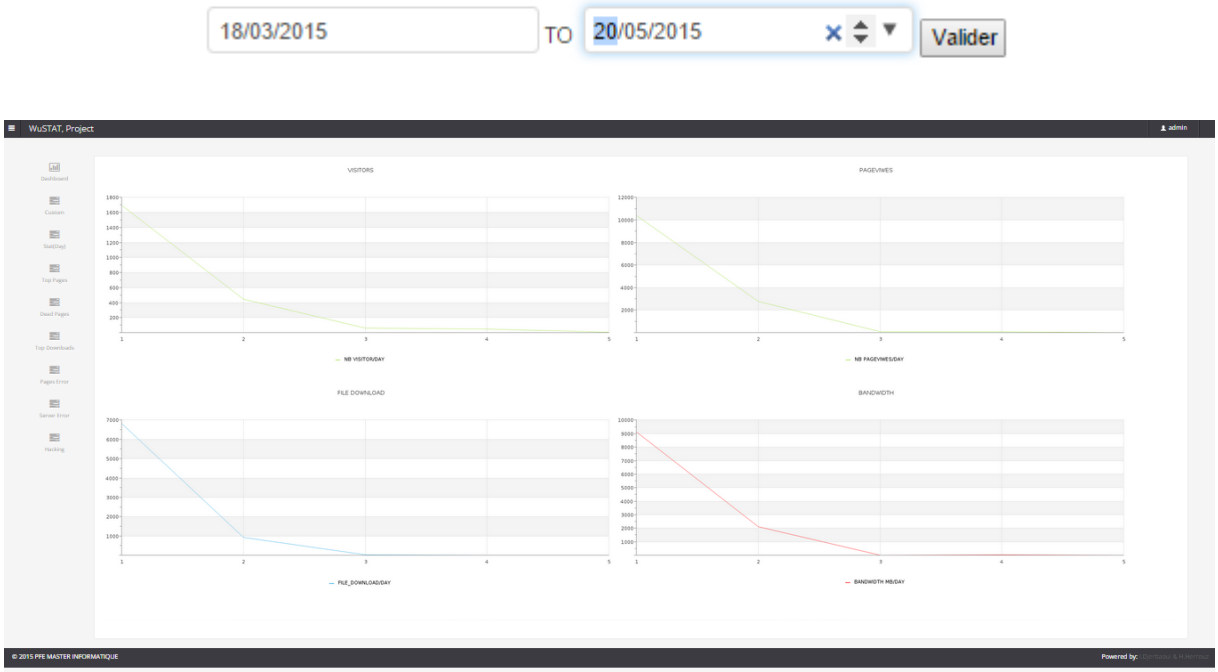


Figure 31 : La page des analyses personnalisé du WuStat

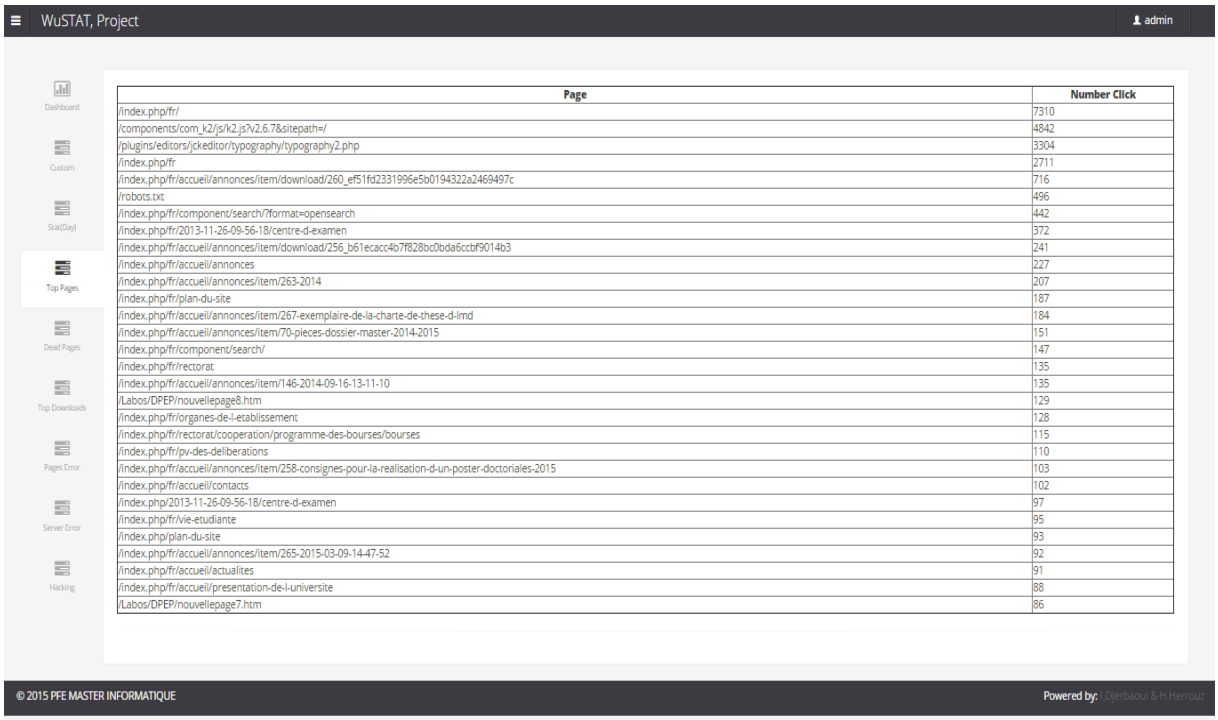
Dans cette page l'utilisateur entrer une date début et une date fin et l'application WuStat traces les graphes des : (Visiteurs / Jour), (Pages Consultés / Jour), (Fichiers Téléchargés / jour), (Trafic/jours), sur l'intervalle du temps entre les deux dates.

4.5.3 Analyse par heurs des jours

Figure 32 : La page des analyses par heurs des jours du WuStat

Dans cette page, l'utilisateur a entré une date et l'application WuStat trace les graphes des : (Visiteurs / heurs), (Pages Consultés / heure), (Fichiers Téléchargés / heure), (Trafic/herue), sur les 24 heures du jour sélectionné.

4.5.4 Les pages populaires



Page	Number Click
/index.php/fr/	7310
/components/com_k2/js/k2.js?v2.6.7&sitepath=/	4842
/plugins/editors/jckeditor/typography/typography2.php	3304
/index.php/fr	2711
/index.php/fr/accueil/annonces/item/download/260_ef51fd2331996e5b0194322a2469497c	716
/robots.txt	496
/index.php/fr/component/search/?format=opensearch	442
/index.php/fr/2013-11-26-09-56-18/centre-d-examen	372
/index.php/fr/accueil/annonces/item/download/256_b61ecacc4b7f828bc0bda6ccbf9014b3	241
/index.php/fr/accueil/annonces	227
/index.php/fr/accueil/annonces/item/263-2014	207
/index.php/fr/plan-du-site	187
/index.php/fr/accueil/annonces/item/267-exemplaire-de-la-charte-de-these-d-lmd	184
/index.php/fr/accueil/annonces/item/70-pieces-dossier-master-2014-2015	151
/index.php/fr/component/search/	147
/index.php/fr/rectorat	135
/index.php/fr/accueil/annonces/item/146-2014-09-16-13-11-10	135
/Labos/DPEP/nouvellepage8.htm	129
/index.php/fr/organes-de-l-etablissement	128
/index.php/fr/rectorat/cooperation/programme-des-bourses/bourses	115
/index.php/fr/pv-des-deliberations	110
/index.php/fr/accueil/annonces/item/258-consignes-pour-la-realisation-d-un-poster-doctoriales-2015	103
/index.php/fr/accueil/contacts	102
/index.php/2013-11-26-09-56-18/centre-d-examen	97
/index.php/fr/vie-etudiante	95
/index.php/plan-du-site	93
/index.php/fr/accueil/annonces/item/265-2015-03-09-14-47-52	92
/index.php/fr/accueil/actualites	91
/index.php/fr/accueil/presentation-de-l-universite	88
/Labos/DPEP/nouvellepage7.htm	86

Figure 33 : La liste des pages populaires

Dans cette page l'application WuStat affiche un tableau des pages populaires avec le nombre des clicks de chaque page.

4.5.5 Les pages impopulaires

Figure 34 : La liste des pages populaires

Dans cette page l'application WuStat affiche un tableau des pages impopulaires avec le nombre des clicks de chaque page.

4.5.6 Les téléchargements populaires

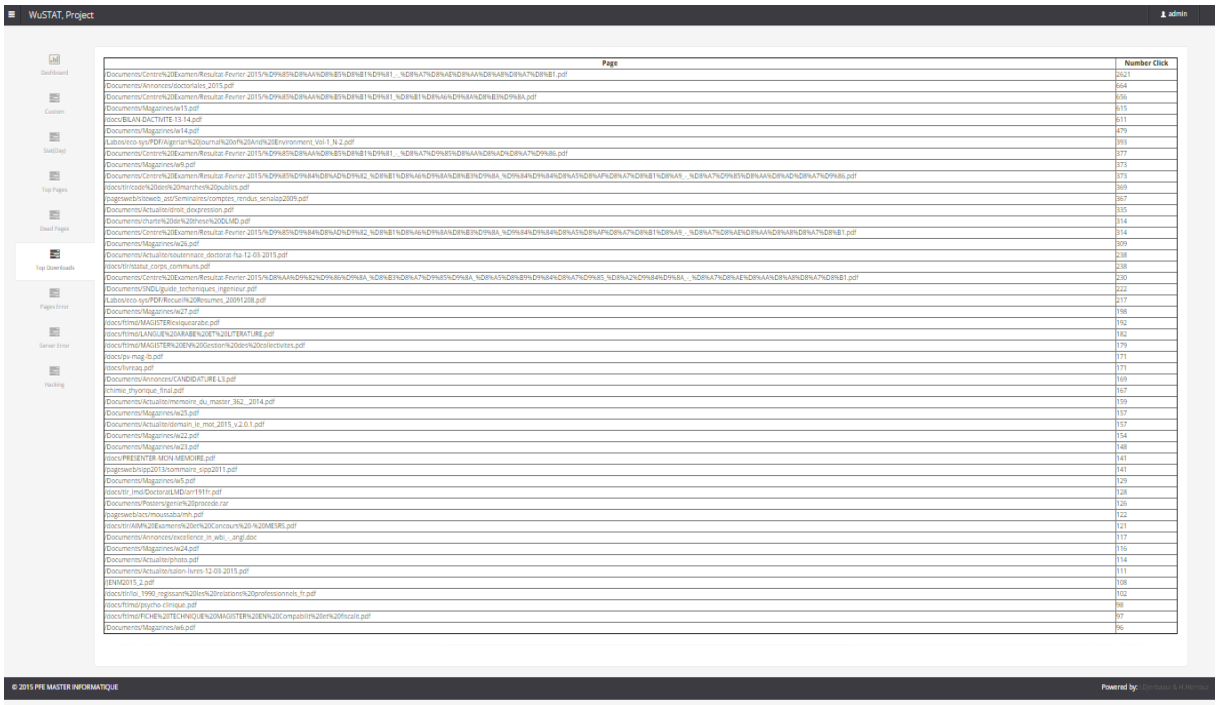


Figure 35 : La liste des téléchargements populaires

Dans cette page l'application WuStat affiche un tableau des téléchargements populaires avec le nombre de téléchargements de chaque fichier.

4.5.7 Les pages erronées

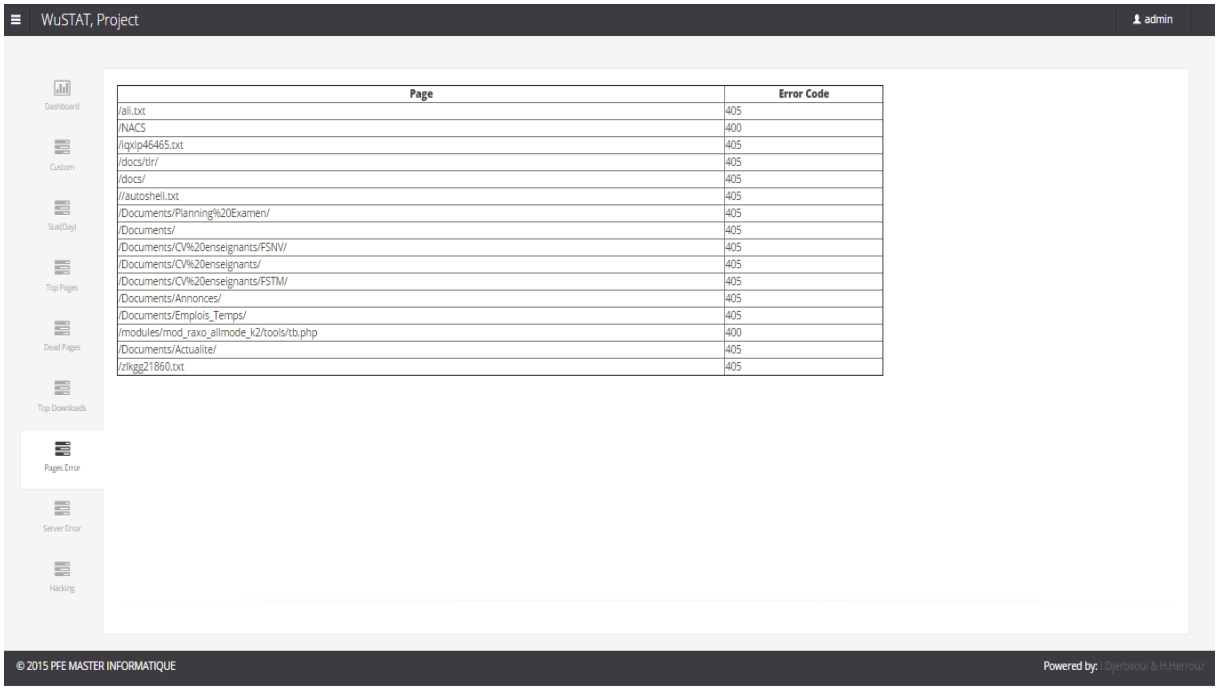
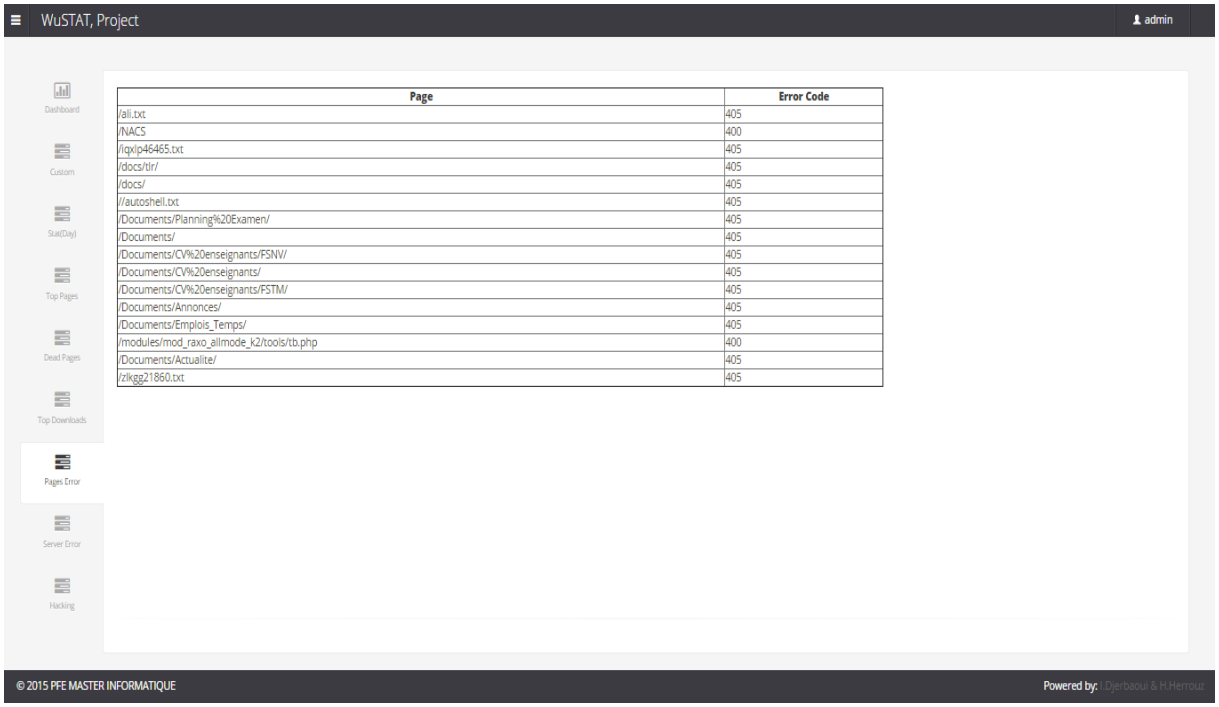


Figure 36 : La liste des pages erronées

Dans cette page l'application WuStat affiche un tableau des pages erronées avec le code erreur de chaque page (Par ex : Code erreur 404, la page n'existe pas sur le serveur).

4.5.8 La liste des erreurs du serveur

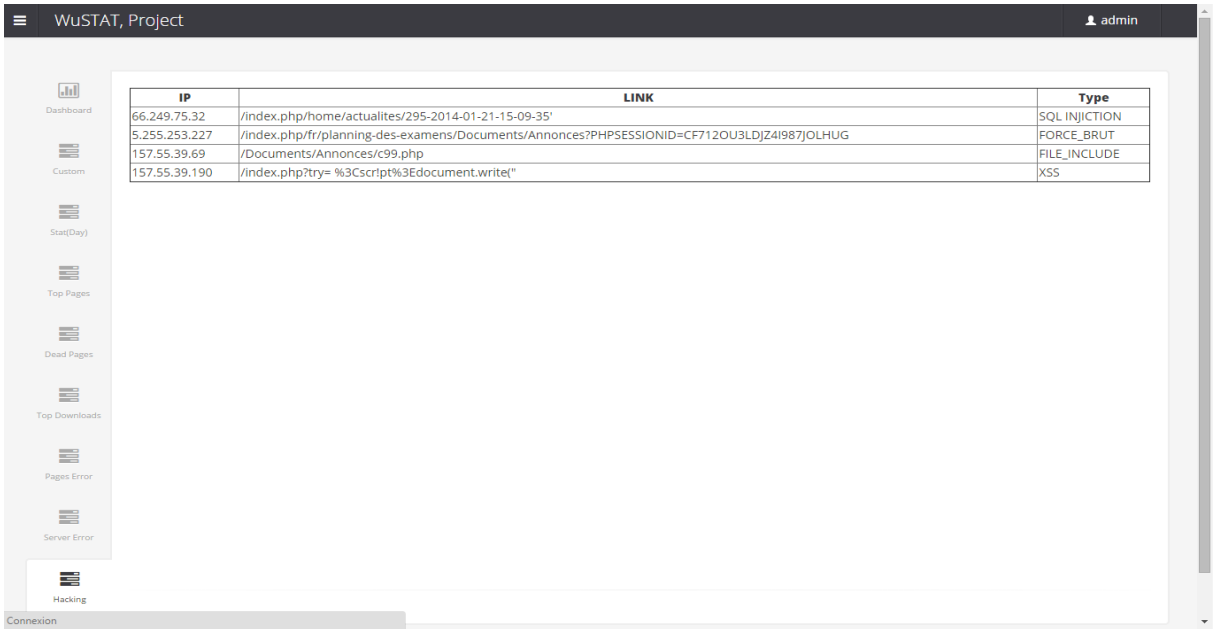


Page	Error Code
/ali.txt	405
/NACS	400
/iqxlp46465.txt	405
/docs/tlr/	405
/docs/	405
//autoshell.txt	405
/Documents/Planning%20Examen/	405
/Documents/	405
/Documents/CV%20enseignants/FSNV/	405
/Documents/CV%20enseignants/	405
/Documents/CV%20enseignants/FSTM/	405
/Documents/Annonces/	405
/Documents/Emplois_Temps/	405
/modules/mod_raxo_allmode_k2/tools/tb.php	400
/Documents/Actualite/	405
/zlkgg21860.txt	405

Figure 37: La liste des erreurs du serveur

Dans cette page l'application WuStat affiche un tableau des erreurs du serveur avec le code erreur de chaque action.

4.5.9 Les attaque



IP	LINK	Type
66.249.75.32	/index.php/home/actualites/295-2014-01-21-15-09-35'	SQL INJICTION
5.255.253.227	/index.php/fr/planning-des-examens/Documents/Annonces?PHPSESSIONID=CF712OU3LDJZ4I987JOLHUG	FORCE_BRUT
157.55.39.69	/Documents/Annonces/c99.php	FILE_INCLUDE
157.55.39.190	/index.php?try= %3Cscr!pt%3Edocument.write("	XSS

Figure 38 : La liste des erreurs du serveur

Dans cette page l'application WuStat affiche un tableau des attaques avec l'adresse IP de chaque attaque plus le type d'attaque.

4.5.10 La sécurité de l'application WuStat

L'application WuStat a été protégée par une page d'authentification composé par un couple (Username, Password) qui sont enregistrés dans une table sur la base des données de l'application.

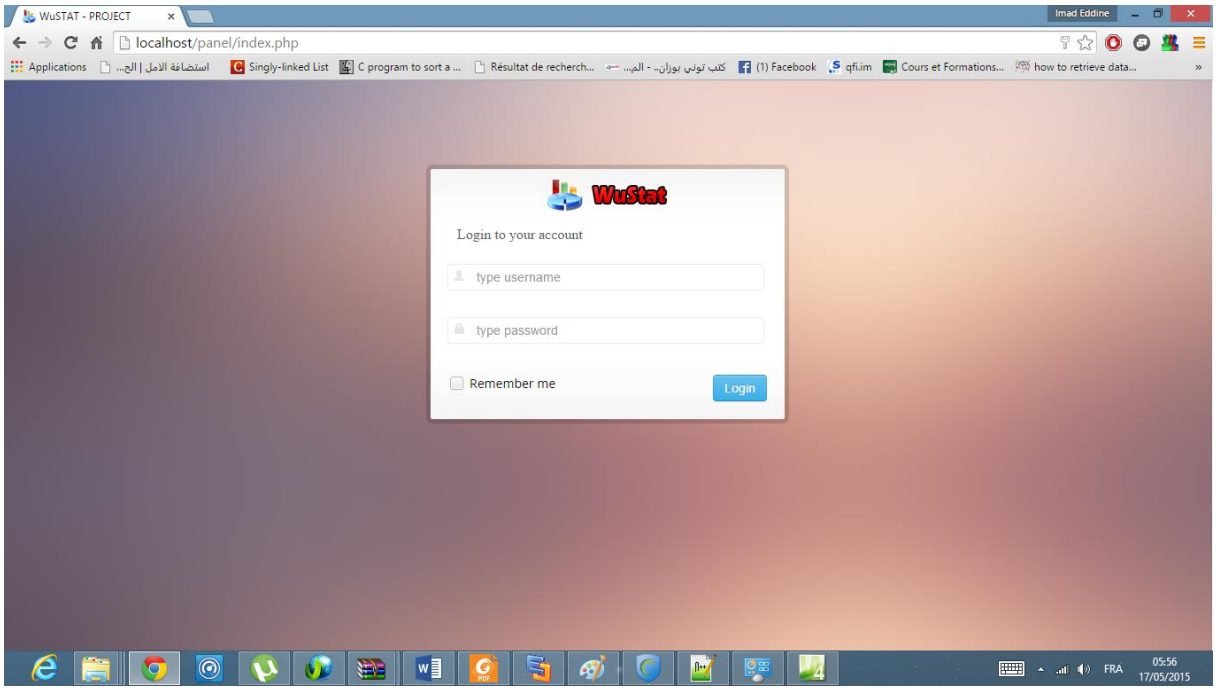


Figure 39: La page login de l'application WuStat

Le mot de passe enregistrer sur la table au format MD5 **[41]**, l'application WuStat à chaque login convertir le mot passe au format MD5 et comparer ce dernier avec le mot passe enregistrer sur la table de base des données.

Figure 40 : La table login

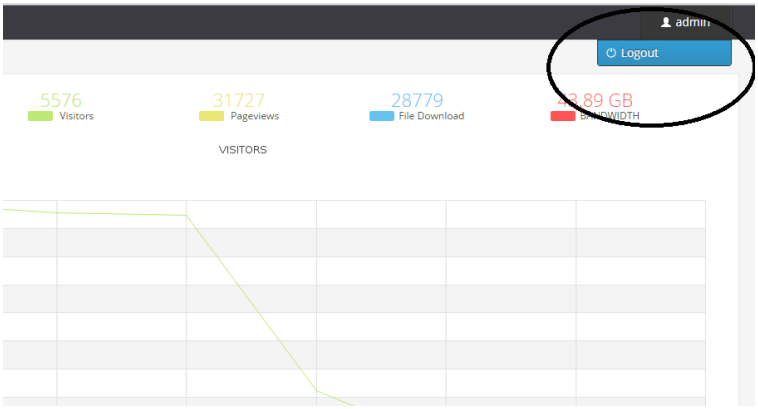


Figure 41 : Le champ de déconnection (logout)

Conclusion :

Dans ce travail, nous nous avons utilisé certaines techniques de Web Usage Mining pour la réalisation d'un outil qui peut contribuer à l'amélioration et au le diagnostic de site Web, sur un cas concret : le site de l'Université Kasdi Merbah Ouargla : www.univ-ouargla.dz , en explorant le fichier Log du serveur Web de l'université.

Ce travail est toujours en cours d'amélioration, en ce qui concerne les thèmes abordés et les fonctionnalités. On focalisera notre intérêt sur le « profil du visiteur » et de « groupe de visiteurs », et on enrichira le travail par l'association d'une « cartographie » des visiteurs.

Chapitre 5 : Expérimentation

Introduction

Ce chapitre décrit l'évaluation et l'expérimentation réalisées sur notre application (outil logiciel), qui consistait à analyser des fichiers log pour extraire des connaissances, et ce, en vue d'une amélioration et une personnalisation du site de l'UKMO.

5.1 Conditions d'expérimentation

L'outil développé a été conçu pour répondre aux besoins des webmasters pour bien comprendre le comportement de leurs visiteurs. Il est nécessaire de le tester avec des webmasters potentiels.

5.1.1 Lieu de l'expérimentation

Notre système est hébergée sur un service d'hébergement (eb2a) donc on a partagé son lien avec des webmasters via les réseaux sociaux (Linkedin, Twitter & Facebook) afin de tester notre application web.

5.1.2 Représentations initiales des apprenants

Les apprenants –une dizaine de webmasters (compétences informatiques et techniques), âgés de 19 à 35 ans – qui sont actuellement responsables d'un site web ou plus.

Pour en savoir un peu plus sur l'intérêt des webmasters des sites web qu'ils possèdent, nous leur avons demandé « Quel est l'objectif de votre site web ? ».

- 50 % vendre des produits enlignes ;
- 30 % site à but éducatif ;
- 15% site personnel ;
- 5% Autres objectifs.

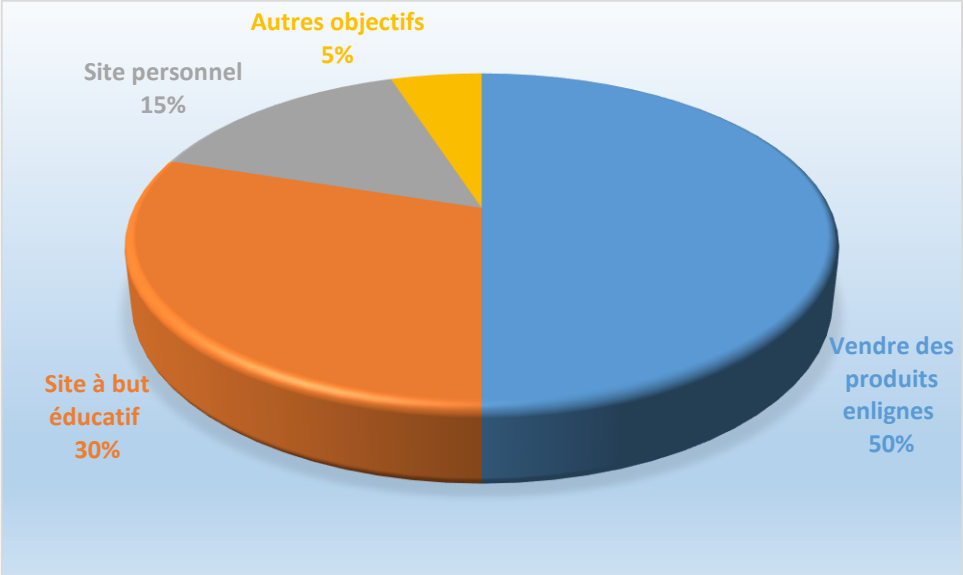


Figure 42 : Le but des sites web potentiels.

Une autre question les été posée « Consultez-vous les fichiers logs de votre site web pour comprendre le comportement de vos visiteurs ? ».

- 60% pas du tout
- 35% Oui, mais je ne comprends pas le comportement des visiteurs;
- 5% autre réponses ;

On peut constater que les webmasters n'arrivent pas à comprendre le comportement des visiteurs de leurs sites web.

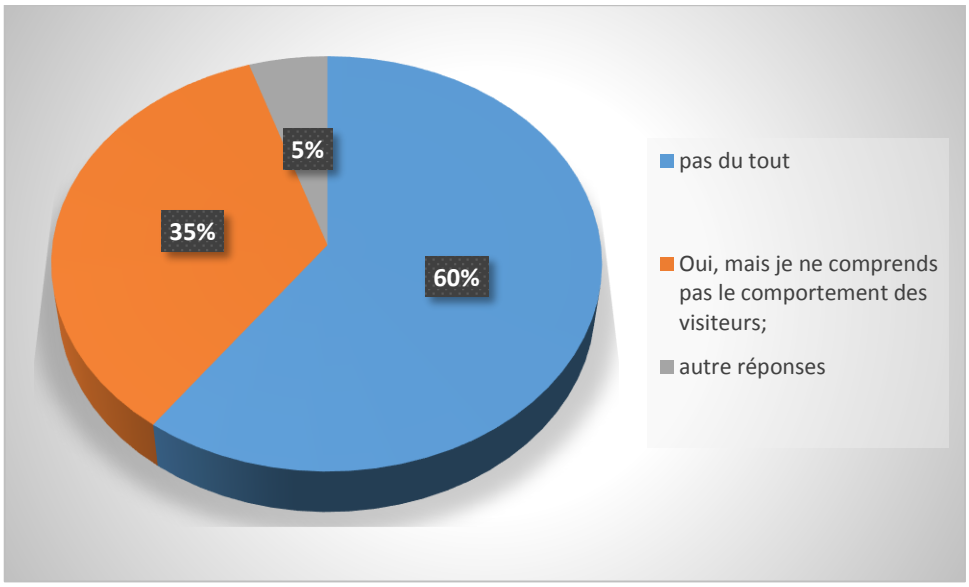


Figure 43 : Les réponses des webmasters

5.2 Mise en œuvre de l'expérimentation

Nous avons organisé la progression de notre expérimentation en deux (02) parties distinctes : une phase de découverte et une phase de production.

5.2.1 Phase de découverte du système WuStat

WuStat – a été hébergé avec succès sur un service d'hébergement gratuit intitulé eb2a.

5.2.2 Phase de production

Le protocole expérimental comporte :

- Un fichier log donné pour tester l'application avant d'Analyzer les fichiers logs des sites web concernés.
- l'accès libre aux toutes les fonctionnalités offertes par notre application (toutes les pages & sous-pages de site).
- Les webmasters doivent nous rendre au moins une capture écran après l'utilisation de WuStat en uploadant leurs fichiers logs.

5.3 Limites de l'expérimentation

Etant donné la courte durée de l'expérimentation, il nous est difficile de tirer des conclusions définitives sur les compétences développées par les webmasters.

5.4 Bilan de l'expérimentation

Grace à notre outil les créateurs des sites web concernés arrivent à bien comprendre le comportement de leurs visiteurs, par cela ils arrivent à fidéliser les internautes fréquentant leurs sites web et à attirer des nouveaux visiteurs en améliorant et personnalisant l'utilisation de leurs sites ainsi ils arrivent à sécuriser et à protéger leurs sites web des attaques des pirates.

D'autres expérimentations sont toujours en cours. Des analyses d'observations d'usages, d'entretiens et de traces informatiques vont permettre d'étudier (plus amplement) le fonctionnement du système WuStat sur Internet. Les résultats attendus seront pris en compte pour améliorer l'ergonomie des fonctionnalités proposées et en concevoir d'autres. Le temps est donc un élément indispensable pour obtenir des résultats généralisables. C'est pourquoi, des expérimentations d'usage sur une période relativement longue sont nécessaires. Toutefois, les premiers résultats sont encourageants.

Conclusion et perspective

Compte tenu du succès flagrant du web, de gigantesques fichiers de traces de navigation laissées par les internautes durant leurs surfs sont continuellement générés. L'analyse de ces fichiers volumineux par les techniques de la fouille de données, appelée Web Usage Mining, peut fournir des connaissances très utiles. Ces dernières peuvent servir dans divers domaines, tels que la personnalisation de sites web, l'amélioration du trafic dans les réseaux, l'analyse de sécurité…etc. Les objectifs que nous nous sommes fixés en menant cette étude consistent tout d'abord à explorer le domaine du Web Usage Mining à travers ses principaux travaux, ses applications, et ses liens avec les domaines connexes, et ensuite de forger une approche un peu différente. A l'opposé des méthodes de fouille de logs centrés serveur, nous avons pu monter une application web de WUM en adoptant une approche centré à l'usage d'utilisateur sur un site web. Cette application inclut trois composants : un transformateur de fichier log vers un format d'une base des données relationnelles, et une procédure de prétraitement de données que nous avons développées, et des fonctions d'extraction de la connaissance a partir les donnée de BDD.

Dans le cadre des perspectives, nous soulevons une des limites de capacité du stockage à cause de la gigantesque quantité d'information contenu dans la base des données qui se reflet sur le cout du stockage, Dans le future nous espérons avoir des nouvelle méthodes d'optimisation de volume de la base des données afin de minimiser le cout de ce dernier.

Bibliographie

[1] Netcraft http://news.netcraft.com/ vu le 31/03/2015.

[2] Blog du modérateur http://www.blogdumoderateur.com/chiffres-internet/ Consulté le : 31/03/2015.

[3] Simon Florentin Adjatan, *La recherche d'information sur internet*, Creative Commons, Avril 2007.

[4] Stéphane Salès, *formation apache le serveur web leader du marché, Maîtriser les points essentiels du serveur web Apache.*

[5] Pierre Yger, *Introduction au world wide web et à xhtml,* IFIPS, Septembre 2009.

[6] Extraction des informations et des connaissances, https://touriaelouahabi.wordpress.com/web-mining/663-2/ Consulté le : 01/04/2015.

[7] Agence Web de Stratégie Internet & Communication Visuelle, http://www.amoks.com/rep-lexique/ido-138/requete_http.html Consulté le : le 01/04/2015.

[8] National instruments, http://zone.ni.com/reference/fr- XX/help/371361-L0114/ivconcepts/ws_http_sessions/ Consulté le : 01/04/2015.

[9] Georges Vignaux, *L'hypertexte Qu'est-ce que l'hypertexte. Origines et histoire,* HAL Archive-Ouvert, Juin 2013.

[10] Olivier GLÜCK, *Architecture et communications Client/serveur*, Université LYON 1, Janvier 2014.

[11] Abdelkrim Jebbour, Fawaz Tairou, Jibril Touzi, Abdourahmane Mbengue, *Initiation à l'informatique et à l'internet*, Creative Commons, Première édition 2009.

[12] M.Jambu, *Introduction au Data Mining*, Eyrolles 1998.

[13] eMarketing.fr, http://www.e-marketing.fr/Definitions-Glossaire/Log-fichier--238242.htm Consulté le : 01/04/2015..

[14] Julien Pauli, *HTTP : le protocole du Web passé en revue, developpez.com,* Avril 2012 .

[15] J. Borges, M. Levene. *Data Mining of User Navigation Patterns* Department of Computer Science, University College London.

[16] L. Tauscher and S. Greenberg. *How people revisit web pages : Empirical findings and implications for the design of history systems*. International Journal of Human Computer Studies, Special issue on World Wide Web Usability 47.

[17] J. Srivastava, R. Cooley, M. Deshpande, and P.-N Tan. *Web usage mining : Discovery and applications of usage patterns from web data*. SIGKKD Explorations, 2000.

[18] D. Pierrakos, G. Paliouras, C. Papatheodorou, and C.D. Spyropoulos. *Web usage mining as a tool for personalization : A survey.User Modeling and User-Adapted Interaction* Vol. 13, 2003.

[19] F.M. Facca and P.L. Lanzi. *Mining interesting knowledge from we blogs : a survey. Data and Knowledge engineering*, Vol. 53, 2005.

[20] D. Tanasa. Web usage mining *: Contributions to intersites logs preprocessing and sequential pattern extraction with low support*. Ph. D. Thesis, University of Nice Sophia Antipolis, 2005.

[21] R. Cooley. *Web usage mining : Discovery and application of interesting patterns from web data*. Phd thesis, University of Minnesota, 2000.

[22] B. Berendt, B. Mobasher, M. Spiliopoulou, and M. Nakagawa. *The impact of site structure and user environment on session reconstruction in web usage analysis*. Proceedings of the 4th WebKDD 2002 Workshop, at the ACM SIGKDD Conference (KDD'2002) on Knowledge Discovery in Databases, 2002.

[23] R. Cooley, B. Mobasher, and J. Sirvastava. *Data preparation for miningworldwide web browsing patterns*. Journal of Knowledge and Information Systems, 1999.

[24] E. Schwarzkopf. *An adaptive web site for the u.m*. conference. In Proceedings of the U.M.2001 Workshop on Machine Learning for User Modeling.

[25] M. Spiliopoulou, L.C. Faulstich, and K. Winkler. *A data miner analysing the navigational behaviour of web users.* In Proc. of the workshop on Machine Learning in User Modeling of the ACAI'99 Int. Conf., Creta, Greece, 1999.

[26] R. Kimball and R. Merz. *Le data webhouse : Analyser des comportements clients sur le web*. Editions Eyrolles, Paris, 2000.

[27] Liu B., Hsu W., Ma Y, *Mining association rules with multiple minimum supports. Proc.* Of the Int. Conf. on Knowledge Discovery and Data Mining. 1999.

[28] Mobasher B. Web Usage Mining, Chapitre 12, dans *Web Data Mining: Exploring Hyperlinks, Contents, and Usage Data.* De Bing Liu, Springer Verlag. 2006.

[29] O. R. Zaïane, M. Xin, J. Han. *Discovering Web Access Patterns and Trends by Applying OLAP and Data Mining Technology on Web Logs*, Proc. Advances in Digital Libraries Conf., 1998.

[30] T. W. Yan, M. Jacobson, H. Garcia-Molina, U. Dayal « *From User Access Patterns to Dynamic Hypertext Linking* » First International World Wide Web Conference May 6-10, 1996, Paris, France.

[31] M. Perkowitz, O. Etzioni. *Towards Adaptative Web Sites : Conceptual Framework and Case Study*, Department of Computer Science and Engineering, University of Washington, Seattle, Jan 1998.

[32] A. Aamodt, E. Plaza. *Case-Based Reasoning: Foundational Issues, Methodological Variations, and System Approaches*, Published in: AI Communication, Vol.7 Nr. 1, March 1994.

[33] R. Kanawati, M. Malek,COBRA: *Une approche d'adaptation structurelle des sites Webfondée sur une technique d'apprentissage à partir des traces d'accès utilisateur et utilisant la méthodologie du raisonnement à partir de cas*, Université Paris Nord.

[34] Bonchi et al. *Taxonomy-driven lumping for sequence mining Data Mining and Knowledge Discovery*, Journal Springer, Oct 2009.

[35] F. Masseglia, P. Poncelet, and R. Cicchetti. *An Efficient Algorithm for Web Usage Mining.* Networking and Information Systems, 1999.

[36] R. Srikant and R. Agrawal. Mining Sequential Patterns: *Generalizations and Performance Improvements*. In Proceedings of the Fifth International Conference on Extending Database Technology (EDBT'96), Avignon, France, September 1996.

[37] AWStats, http://awstats.sourceforge.net/docs/awstats_glossary.html, Consulté le : 06/04/2015.

[38] Webalizer, http://www.webalizer.org/ Consulté le : 06/04/2015.

[39] John Callender, *Perl for Web Site Management*, O'Reilly Media, Octobre 2001.

[40] Mathieu Nebra, *Qu'est-ce que le Responsive Web Design ?*, openclassrooms, 2013.

[41] Pierre-Alain, *Fonctions de hachage*, Ecole normale supérieure, 2009.