République Algérienne Démocratique et Populaire
Ministère de l'Enseignement Supérieur et de la Recherche Scientifique

UNIVERSITE AMAR TELIDJI LAGHOUAT

FACULTE DES SCIENCES ET DE L'INGENIERIE

DEPARTEMENT DE GENIE INFORMATIQUE

PROJET DE FIN D'ETUDES

Pour l'Obtention Du Diplôme

D'INGENIEUR D'ETAT EN INFORMATIQUE

Option : Systèmes Parallèles et Distribués

Thème :

Web Usage Mining : Conception et réalisation d'un outil d'analyse des fichiers logs des serveurs

Réalisé par:

Hiba Souad.
Recioui Amina.

Encadré par:

Mr : B.ziani.

N° d'ordre: ……../2009-PFE/DGI

Remerciements

Au terme de ce travail, nous tenons à exprimer nos vifs remerciements :

En premier lieu au **DIEU TOUT-PUISSANT** qui nous a donné la patience et la sante afin d'accomplir ce travail.

Au Professeur Monsieur **Ziani Benameur**, de l'Université de Laghouat, pour nous avoir honorées en acceptant de diriger, avoir guidé, encourager et conseiller pendant toute la période de ce travail. Nous tenons à mentionner le plaisir qu'on a eu à travailler avec lui.

Nous remercions tous ceux sans qui ce mémoire ne serait pas ce qu'il est, aussi bien par les discussions qu'on a eu la chance d'avoir avec eux, leurs suggestions ou contributions. Nous pensons ici en particulier **M. Ben Saad**.

Nous tenons fermement à mentionner le plaisir qu'on a eu à étudier à l'université de Laghouat.

Nous tenons également à associer cette œuvre à tous nos collègues de promotion qu'on a eu le plaisir de côtoyer pendant cette période des études. Une pensée va particulièrement à tous ceux d'entre nous qui n'ont pas eu la possibilité d'aller jusqu'au bout de leurs études.

Nous tenons également à remercier tous nos frères, sœurs et amis qui ont cru en nous, nous ont encouragé et nous ont donné la force d'aller jusqu'au bout.
Nous pensons enfin fortement à tous ceux qui ont contribué de près ou de loin à la réalisation de ce travail.

Dédicace

Je dédie ce mémoire

A ma grande mère qui ma toujours poussé et motivé dans mes études. Sans elle, je n'aurais certainement pas fait d'études Supérieurs.

A ceux qui sont eux rien de tous cela n'aurait été possible

Mon père et ma précieuse mère.

A mes chères sœurs : Nawel, Iman, Wafa, Chahinéz.

A mon grand père : Remma Mohamed.

A mes tantes et à mes oncles : Bachir, Mourad, Malika, Zhor, Souad.

A l'adorable Assia

A chaque cousins et cousines.

A mon cher et adorable ami Kheirddine.

A mes meilleures amies : Hadjer, Karima, Fatima, karima, fattoum, Hayat, Leila, Hadja, Soumia.

A la plus adorable et la plus sympa de toutes les filles du monde : Amina

Sawcen

Dédicace

A l'Eternel mon berger qui me protège, me conduit et à qui je dois tout.

A mes très chers parents qui ont toujours été là pour moi, et qui m'ont donné un magnifique modèle de labeur et de persévérance. J'espère qu'ils trouveront dans ce travail toute ma reconnaissance et tout mon amour.

A ma chère sœur ***Kawther*** *et son mari.*

A mon adorable nièce ***Nahla****.*

A mes chers frères et sœurs : ***Taha, Soumia, Mouadh****.*

A ma grande mère et mes tantes et à mon oncle.

A ma tante et à touts ces fils et ces filles.

A chaque cousins et cousines.

A ma collègue et ma meilleure amie ***Sawcen*** *pour les multiples discussions constructives tout au long de la rédaction de ce mémoire,*

A mes meilleures amies : ***Hadja, Hadjer, Karima ,Fatima, Fattome ,Nacira, Hayet, Leila, Meriem, Halima, Atika, Zineb,*** *Je dédie ce mémoire*

Amina

Sommaire

La liste des tableaux

La liste des figures

Résumé

L'objectif de ce travail est la conception et la réalisation d'un outil logiciel, 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 visiteurs, et de répondre à leurs besoins en adaptant le contenu, la forme et l'agencement des pages web.

Mots-clés : Web usage mining , fichier Log.

Abstract

The objective of this work is the design and implementation of a software tool, using the concepts of "Web usage mining" which will allow the" Webmaster "to have all the knowledge of the website that manages for improving and customization.
It will extract information from the log file of Web server hosting the website, and decide by discovering habits of visitors and meet their needs by adapting the content, the shape and layout of web pages.

Keywords : Web usage mining, Log file.

تلخيص

الهدف من هذا العمل هو تصميم وتنفيذ أداة البرمجيات ، وذلك باستخدام مفاهيم "التنقيب في استخدامات الويب" ، والتي سوف تسمح لمدير الموقع بالحصول على جميع المعارف على موقع الانترنت الذي يديره. و دلك لأجل التحسين والتخصيص.
ويتم في الواقع باستخراج معلومات من ملف السجل من خادم ويب التي تستضيف الموقع ، وتقرر لاكتشاف عادات مستخدمي الإنترنت ، والاستجابة لاحتياجاتهم من خلال تكييف المحتوى ، وشكل وتصميم صفحات الويب.

المفاتيح التنقيب في استخدامات الويب, ملف السجل.

Chapitre 1

Introduction générale

Sommaire

Introduction générale

On ne saurait parler de fouille de données sans comprendre toutes les notions de ce domaine. Ainsi, il est important de faire une présentation de la fouille de données.

Ce chapitre est structuré en quatre sections. Nous présentons dans la première section, la fouille de données et sa place dans le processus complet d'extraction des connaissances à partir des données (*ECD*). La section suivante présente le web mining et ces types. Ensuite, nous décrivons le travail demandé, et dans La dernière section nous finissons par donner un aperçu sur l'organisation du rapport.

1.1 Des données aux connaissances

Parmi les approches se situant dans une problématique générale d'aide à la décision, des démarches récentes tentent de tirer les leçons des situations passées pour lesquelles d'importantes données sont stockées dans grandes bases.

Motivés par ces problèmes d'aide à la décision, les chercheurs de différentes communautés (intelligence artificielle, statistique, bases de données, interaction homme-machine) se sont intéressés à la conception et au développement de la connaissance de ces grandes bases. Ces techniques sont le sujet d'un thème de théorie et d'outils permettant d'extraire automatiquement de la connaissance de ces grandes bases. Ces techniques sont le sujet d'un thème de recherche connu d'ECD (Extraction des Connaissances à parti des données) ou KDD (Knowledge Discovery in Data mining).

Les volumes des bases de données se sont multipliés d'une manière vertigineuse.les architectures deviennent plus complexes en passant d'un fichier unique à des bases de données réparties dans des environnements hétérogènes. Les procédures de traitement deviennent ainsi plus complexes, les informations recherchées le deviennent aussi.

Après la simple interrogation *Quelles sont les ventes de tels produits à telles périodes?* On cherche plutôt à connaître les caractéristiques des clients, leurs autres préférences d'achat ou plut généralement rechercher des associations client-produits en émergence. Cette évolution nécessite une adaptation des méthodes utilisées jusqu'à présent. Du langage d'interrogation logique (SQL) on passe aux techniques de recherche d'associations, de classification et de modélisation. Les différentes stratégies mises en œuvre pour répondre à ces nouvelles exigences consistent à intégrer des outils statistiques et d'intelligence artificielle aux systèmes de gestion de bases de données.

Il est important de noter la différence entre les trois termes suivants :

- **Donnée** : valeur d'une variable pour un objet (comme le montant d'un retrait d'argent par exemple).
- **Information** : résultat d'analyse sur les données.
- **Connaissance** : information utile pour l'organisme (comme la découverte du mauvais emplacement de certains distributions).

1.1.1 Définition

- L'ECD est le procédé non trivial d'identification de connaissances valides ; nouvelles ; potentiellement utiles et compréhensibles de données.
- L'ECD est un processus itératif et interactif d'analyse d'un grand ensemble de données brutes afin d'en extraire des connaissances exploitables par un utilisateur-analyste qui y joue un rôle centrale.

  Ce processus est itératif car les résultats d'une étape peuvent remettre en cause les traitements effectués durant les étapes précédentes, et il est interactif car la qualité des résultats obtenus dépend en grande parties de l'intervention des utilisateurs finaux. [1]

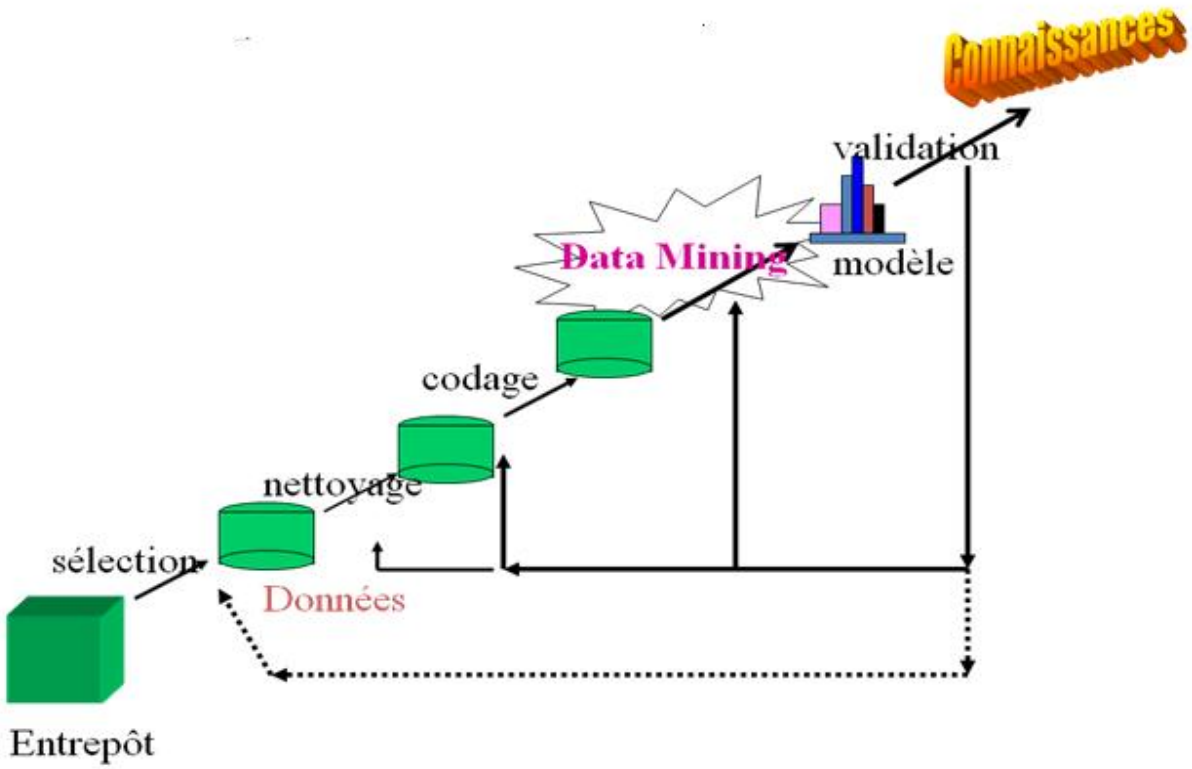


Figure 1.1 : Les étapes de processus d'ECD

Comme nous pouvons le voir sur la figure 1.1, l'ECD est un processus essentiellement interactif et itératif. Sous cette forme, nous pourrions dire que l'ECD est un véhicule dont la fouille de données est le moteur.

1.2 Fouille de données (data mining)

1.2.1 Définition

Le terme de fouille de données signifie littéralement forage de données. Comme dans tout forage, son but est de pouvoir extraire un élément : la connaissance.

De manière générale, on peut la définir comme l'extraction d'informations ou de connaissances originales, auparavant inconnues, potentiellement utiles à partir de gros volumes de données.

La fouille de données correspond donc à l'ensemble des techniques et des méthodes qui à partir de données, permettent d'obtenir des connaissances exploitables. Son utilité est grande dès lors que l'entreprise possède un grand nombre d'informations stockées sous forme de bases de données. [2]

1.2.2 Les technique de fouille de données

La fouille de donnes consiste à utiliser un ensemble de techniques statistiques qui, en fouillant un grand nombre de données structurées, permettent de découvrir et de présenter des informations à valeur ajoutée dans une forme interprétable facilement par un individu.

1.2.2.1 Méthodes statistiques unidimensionnelles

Ces méthodes permettent une analyse exploratoire des données à travers les indicateurs statistiques (moyenne, écart-type,...) et la représentation graphique (histogrammes, boîte de dispersion,...).

1.2.2.2 Méthodes statistiques multidimensionnelles

Ces méthodes (Factorisation, segmentation, classification,) permettent, en réduisant l'espace et en fournissant des représentations graphiques, d'exploiter, de fouiller et de représenter des grands ensembles de données.

- **Méthodes factorielles**

  Les méthodes factorielles se proposent de fournir des représentations synthétiques, souvent sous forme graphique, de vastes ensembles de valeurs numériques.

  - L'analyse en composantes principales (ACP) est une technique permettant de réduire un système complexe de corrélations en un nombre inférieur de dimensions.
  - L'analyse factorielle des correspondances (AFC) traite des variables qualitatives.
  - L'analyse factorielle des correspondances multiples (AFCM) s'applique sur des grands tableaux de variables nominales où les lignes sont les individus et les colonnes des variables descriptives.

- **Méthodes de classification automatique**

  La classification automatique, appelée aussi classification non supervisée, ou segmentation consiste à rechercher des groupes homogènes inconnues au départ dans une population d'individus représentés par une ou plusieurs variables. La fouille de données propose plusieurs méthodes de classification automatique telle que la classification ascendante hiérarchique, la classification descendante hiérarchique, ...etc.

- **Méthodes de classification supervisée**

  La classification supervisée cherche à déterminer l'appartenance d'un événement à des classes préalablement identifiées par segmentation. Pour ce faire, de nombreuses méthodes de classification sont utilisées telles que les arbres de décision, les réseaux de neurones.

1.2.2.3 Méthodes d'association

Les règles d'association est une méthode de fouille des données non supervisée qui consiste à déterminer les valeurs associées parmi les données. Une règle d'association est une règle de la forme : Si condition (prémisse) alors résultat. Par exemple, Si X et Y alors Z.

Le choix d'une règle d'association nécessite de définir des indicateurs servant à sa validation, à savoir le support, la confiance et l'amélioration de la règle.

1.2.2.4 Méthodes basées sur l'intelligence artificielle (réseaux de neurones)

Par analogie avec les neurones biologiques, les réseaux de neurones sont des ensembles de calculateurs numériques (neurones formels) agissant comme des unités élémentaires,

reliés entre eux par des interconnexions pondérées qui transmettent des informations numériques d'un neurone formel à un autre. [3]

1.3 Le Web mining

L'application des techniques de fouille de données à l'Internet est connue sous le nom de Web mining.

D'une manière générale, le Web mining peut être défini comme la découverte et l'analyse d'informations utiles à travers le World Wide Web. On distingue ainsi trois types de découvertes possibles :

1.3.1 Le web content mining

C'est le processus de découvrir des informations utiles à partir de texte, image, audio ou vidéo sur le web.web content mining est parfois appelé text mining, parce que le contenu de texte est le domaine le plus largement étudié.

1.3.2 Le web structure mining

Se focalise sur la structure de l'internet et des liens hypertextes.

1.3.3 Le Web usage mining

C'est le processus d'analyse des accès web des utilisateurs. L'analyse des fichiers logs offre des informations très utiles pour améliorer les performances du réseau, restructurer un site ou, d'une manière générale, cibler le comportement des visiteurs d'un site.

1.4 Description du projet

Notre travail correspond à une application du Web Usage Mining qui est la réalisation d'un outil qui permet d'extraire des connaissances à partir de du fichier log. d'en tirer toutes les informations utiles sur le comportement des utilisateurs d'un site web. Nous nous intéressons dans ce qui suit aux déférentes techniques utilisées dans ce domaine.

Ceci va permettre en particulier d'avoir des informations sur :

- Les visiteurs par période (une semaine, un mois…).
- Les pages et les liens exploités sur le site.

Ainsi, en prenant en considération l'ensemble de ces connaissances, le web master pourra répondre aux besoins des navigateurs en adaptant le contenu, et les liens avec leurs attentes.

1.5 Organisation du rapport

Pour conclure cette introduction, nous présentons le plan de ce mémoire.

- **Le chapitre 1** présente l'étude préalable où nous présentons une définition générale de l'ECD et leur objectif, ainsi la fouille de données et son cas d'application (web mining).

- **Le chapitre 2** il contient une explication détaillée sur le web usage mining, après avoir présenté les problèmes posés, les solutions possibles et celles retenues.

- **Le chapitre 3** est relatif à l'analyse et à la spécification des besoins, il développe la conception du logiciel avec la modélisation UML.

- **Le chapitre 4** présente la réalisation qui décrit l'environnement de développement et l'implémentation en utilisant l'outil de développement le langage Delphi.

Chapitre 2

Web usage mining

Sommaire

2.1 Web usage Mining

[5] définit le WUM 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 Motifs du Web en vue de répondre aux besoins des visiteurs de manière spécifique et adaptée (personnaliser les services) et faciliter la navigation.

2.1.1 Définitions des termes du Web Usage Mining

Le tableau 2.1 présente les définitions des principales notions utilisées dans WUM. Afin de mieux se comprendre, avant de commencer, voyons la signification exacte de chacun de ses termes [1]:

Ressources Web	toute ressource (image, fichier html, etc.) accessible via un protocole http.
Serveur Web	Un serveur qui donne accès à des ressources Web.
Requête Web	Une requête pour une ressource web, faite par un client (navigateur web) à un serveur web.
Site Web	Ensembles des informations consistant en une ou plusieurs ressources Web identifiées par un seul URL (Uniform Resource Identifier)
présentation de page **(page view)**	Affichage d'une page web dans l'environnement visuel client à un moment précis dans le temps.
Navigateur Web	Logiciel de type client chargé d'afficher les pages à l'utilisateur et de faire des requêtes http au serveur Web.
Utilisateur	Personne qui accède à des pages Web situées sur un ou plusieurs serveurs Web, en utilisant un navigateur.
Session utilisateur	Un ensemble délimité des clics utilisateurs sur un ou plusieurs serveurs Web.
Navigation ou visite	Sous-ensemble d'une session utilisateur. La distance en temps entre deux requêtes consécutives est inférieure à un seuil prédéfini(en général le seuil est de 30 minutes).
Fichier log	Fichier texte où est enregistré l'historique des communications entre un serveur et des postes clients. On retrouvera en particulier les requêtes demandées au serveur, les messages d'erreurs générés par l'application.

TAB.2.1- Les principaux termes utilisés en WUM

2.1.2 Sources 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 enregistrées 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 serveur**

Chaque demande d'affichage d'une page Web de la part d'un utilisateur, peut générer plusieurs requêtes. Des informations sur ces requêtes (notamment les noms des ressources demandées et les réponses du serveur Web) sont stockées dans les fichiers Log du serveur Web. 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) [5].

- **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'appliquettes 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 [6]. 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'utiliser ce navigateur modifié dans leurs navigations sachant qu'il peut être considéré comme une menace de leur vie privée [5]. 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 des sessions [7].

- **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.

2.1.3 Fichiers Logs http

Les principales données exploitées dans le WUM proviennent des fichiers Logs. Cependant, il existe d'autres sources d'informations qui pourraient être exploitées à savoir les connaissances sur la structure des sites Web et les connaissances sur les visiteurs des sites Web.

Présentation des fichiers Log http

Un Log file (en français, journal de bord des connexions ou fichier journal) est un fichier créé par un logiciel spécifique installé sur le serveur Web permettant de garder une trace des requêtes reçues et des opérations effectuées par le serveur. Il existe plusieurs formats des fichiers Logs Web, mais le format le plus courant est le CLF (Common Log file Format). Selon ce format six informations sont enregistrées:

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

Le format ECLF (Extended Common Log file Format), supporté par certains serveurs Web, représente une version plus complète du CLF. En effet, il indique en plus l'adresse de la page de référence (où se trouvait l'internaute lorsqu'il a lancé la requête (referrer)) et l'identification de l'agent client (le nom du navigateur Web et le système d'exploitation).

FIG. 2.1 : Extrait d'un fichier log

- **Analyse des fichiers Logs**

L'analyse des fichiers Logs consiste à collecter automatiquement à partir des fichiers Logs les navigations de tous les utilisateurs. Cette analyse permet de quantifier la fréquentation des pages d'un site donné, déterminer les parcours de navigation, les motifs de navigation et les profils des usagers du site considéré.[8].

2.1.4 Prétraitement des données

Le prétraitement des données consiste à convertir ces informations en des données structurées adéquates à l'analyse.

Il s'agit principalement de :

- Fusionner les fichiers log et nettoyer les données.
- Stocker les données structurées dans une base de données.

- **Nettoyage des données**

Le nettoyage des données est une étape cruciale dans le processus du WUM en raison du volume important des données enregistrées dans les fichiers Log Web. En effet, la dimension de ces fichiers dans les sites Web et les portails Web très populaires peut atteindre des centaines de giga-octets par heure. 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 des images et des scripts dépend selon [7] de l'intention de l'analyste. En effet, si son objectif est de trouver les failles de la structure du site Web ou d'offrir 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. La suppression des requêtes 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. [5] a utilisé trois heuristiques pour identifier les requêtes et les visites issues des WRs :

1. Identifier les adresses IPs qui ont formulé une requête à la page « robots .txt».

2. Utiliser des listes des « User agents» connus comme étant des WRs.

3. 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.

2.1.5 Application du Web usage Mining

Il existe cinq activités dans lesquelles on trouve des applications du WUM [9] :

1. Évaluation et caractérisation générale de l'activité sur un site Web : l'objectif est l'observation et non pas la modélisation. Les techniques d'analyse utilisées sont souvent simples. Elles relèvent, en effet, du dénombrement et des statistiques simples (moyennes, histogramme, indices, tris croisés).

2. Amélioration des modes d'accès aux informations : le WUM permet de comprendre comment les utilisateurs se servent d'un site, d'identifier les failles dans la sécurité et les accès non autorisés.

3. Modification de la structure : le WUM peut révéler le besoin de restructurer des pages et des liens afin d'améliorer la structure du site Web. En effet,

les pages considérées comme similaires par des techniques de classification peuvent être reliées de manière hypertextuelle.

4. Personnalisation de la consultation : cet enjeu important pour de nombreuses applications Internet ou sites de e-commerce consiste à proposer des recommandations dynamiques à un utilisateur en se basant sur son profil et une base de connaissances d'usages connus.

5. Mise en œuvre de l'intelligence économique: cet objectif concerne en particulier les sites marchands. Il s'agit de comprendre quand, comment et pourquoi l'utilisateur est attiré par ce site, les produits qu'il faut lui proposer à la vente...etc.

2.1.6 Les problèmes spécifiques aux données des fichiers Logs :

Bien que les données fournies par les fichiers Logs soient utiles, il importe de prendre en compte les limites inhérentes à ces données lors de leur analyse et de leur interprétation. Parmi les difficultés qui peuvent survenir:

- **Les requêtes inutiles** : Chaque fois qu'il reçoit une requête, le serveur enregistre une ligne dans le fichier Log. Ainsi, pour charger une page, il y'aura autant de lignes dans le fichier que d'objets contenus sur cette page (les éléments graphiques). Un prétraitement est donc indispensable pour supprimer les requêtes inutiles.

- **Les firewalls** : Ces protections d'accès à un réseau masquent l'adresse IP des utilisateurs. Toute requête de connexion provenant d'un serveur doté d'une telle protection aura la même adresse et ce, quel que soit l'utilisateur. Il est donc impossible, dans ce cas, d'identifier et de distinguer les visiteurs provenant de ce réseau.

- **Le Web caching**: Afin de faciliter le trafic sur le Web, une copie de certaines pages est sauvegardée au niveau du navigateur local de l'utilisateur ou au niveau du serveur proxy afin de ne pas les télécharger chaque fois qu'un utilisateur les demande. Dans ce cas, une page peut être consultée plusieurs fois sans qu'il y' ait autant d'accès au serveur. Il en résulte que les requêtes correspondantes ne sont pas enregistrées dans le fichier Log.

- **L'utilisation des robots** : Les annuaires du Web, connus sous le nom de moteurs de recherche, utilisent des robots qui parcourent tous les sites Web afin de mettre à jour leur index de recherche. Ce faisant, ils déclenchent des requêtes qui sont enregistrées dans tous les fichiers Logs des différents sites, faussant ainsi leurs statistiques.

Chapitre 3

Modélisation et conception du logiciel

Sommaire

3.1 Modélisation du logiciel

3.1.1 Introduction

Dans cette partie, nous abordons la phase de modélisation de logiciel. Ainsi, nous présentons les besoins fonctionnels et non fonctionnels de notre application. Nous utilisons le langage UML comme un moyen simple et compréhensible afin de décrire les principaux cas d'utilisation.

3.1.2 Besoins fonctionnels

Cette partie décrit les exigences que le système doit satisfaire d'une façon informelle.

Les fonctionnalités qu'on se propose de fournir dans notre logiciel sont les suivantes :

- Générer un rapport détaillé contenant toutes les statistiques .Ces statistiques concernent particulièrement :
- pages web les plus visités avec des informations relatives aux nombres de visites.
- les utilisateurs les plus actifs sur le réseau avec une description des activités.
- Les navigateurs et les systèmes d exploitations les plus utilisés.

3.1.3 Besoins non fonctionnels

- L'application doit présenter des interfaces conviviales et ergonomiques afin de faciliter l'utilisation de l'application par un utilisateur qu'il soit spécialiste ou non.
- Seuls les journaux de Serveur seront pris en considération.

Afin de mieux comprendre les fonctionnalités de notre outil nous présentons le diagramme de cas d'utilisations qui nous jugeons le plus représentatif.

3.1.4 Les cas d'utilisation du système

Dans ce qui suit, nous présentons un formalisme semi formel de spécification des besoins de notre système, à l'aide de diagramme de cas d'utilisation accompagné par une explication textuelle de se principal cas d'utilisation.

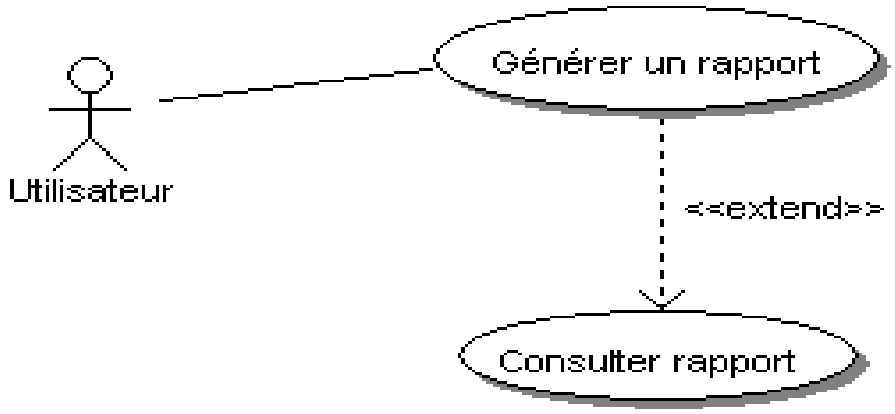


Figure 3.1 : diagramme de cas d'utilisation : Utilisateur

- **Le cas d'utilisation « Générer rapport »**

L'utilisateur peut générer directement des rapports détaillés des activités des visiteurs sous format HTML.

- **Le cas d'utilisation « Consulter statistiques»**

L'utilisateur peut visualiser les statistiques correspondantes sous formes tabulaire et graphiques.

3.2 Conception du logiciel

3.2.1 Introduction

Cette partie comporte une première section relative à la conception de cette application. La deuxième section décrit la base de données utilisée pour le stockage des informations obtenues à partir des fichiers « log » en utilisant le langage UML.

3.2.2 Conception

Notre conception doit prévoir deux transformations de données. La première transformation permet le passage des données du fichier « log » sous leur format brut vers une forme adaptée aux calculs ultérieurs. La deuxième transformation reprend les résultats

de la première transformation et permet de déduire les résultats à afficher sous une forme facile, intelligible et conviviale.

Nous avons envisagé une solution possible. Cette solution consiste à lire le fichier log, stocker les résultats dans une base de données, tous les calculs et la génération des statistiques seront faits à partir de la base.

Cette solution permet d'exploiter le fichier « log » une seule fois et en utilisant des variables intermédiaires permettant d'accélérer certains calculs pour la navigation sans revenir systématiquement à la base de données.

3.2.2.1 Diagramme de classe

Nous avons une classe **FichierLog** qui contient la méthode **AnalyserLog (),** cette méthode permet la lecture du fichier log ligne par ligne.

La classe **BaseDeDonnées** permet d'apporter des modifications sur le contenu de la base de données c'est à dire qu'elle est responsable de la suppression des données concernant les pages possédant des graphiques, Images ou des scripts via la méthode **NettoyerImg ()** ainsi que de la suppression des requêtes des rebots de la base par le billet de la méthode **NettoyerWR ().**

La méthode **insererEnreg ()** qui permet d'apporter des modifications sur le contenu de la base de données c'est à dire qu'elle est responsable de l'insertion dans la base de nouveaux enregistrements. La méthode **Interroger ()** est responsable de l'interrogation de la base.

Une autre classe **Rapport** responsable de la génération de rapports contenant les différentes statistiques. Cette classe contient une méthode **GenererRapport ()** qui permet de générer le rapport final sous le format HTML, et la méthode **GenererGraphe()** qui permet la génération des graphes des différentes statistiques.

La classe **Graphe** pour générer différents graphiques correspondant aux statistiques affichées en utilisant le graphe le mieux adapté pour chaque statistique. Cette classe contient une méthode **GrapheBarChart3D ()** qui génère un histogramme comme c'est le cas pour les statistiques concernant les utilisateurs, les pages et une méthode **GraphePieChart3D ()** qui génère des statistiques concernant les navigateurs, les OS et les pays.

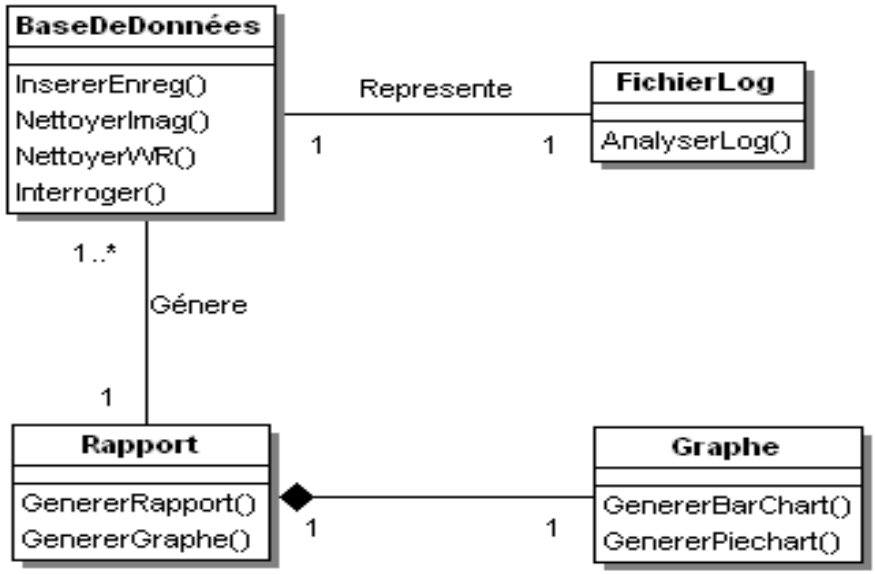


Figure 3.2 : Le diagramme de classe

3.2.2.2 Diagramme de séquence

Le diagramme de séquences permet de représenter des collaborations entre objets selon un point de vue temporel, on y met l'accent sur la chronologie des envois de messages.

- L'utilisateur fait l'initialisation de la base de données.
- Il sélectionne un fichier log pour l'analysé.
- Les données importées de ce fichier seront stoker dans une base de données un nettoyage des requêtes inutiles tel que les images et les robots se fait sur la base de données.
- L'utilisateur peut générer un rapport a partir de cette base el peut consulter ce rapport.

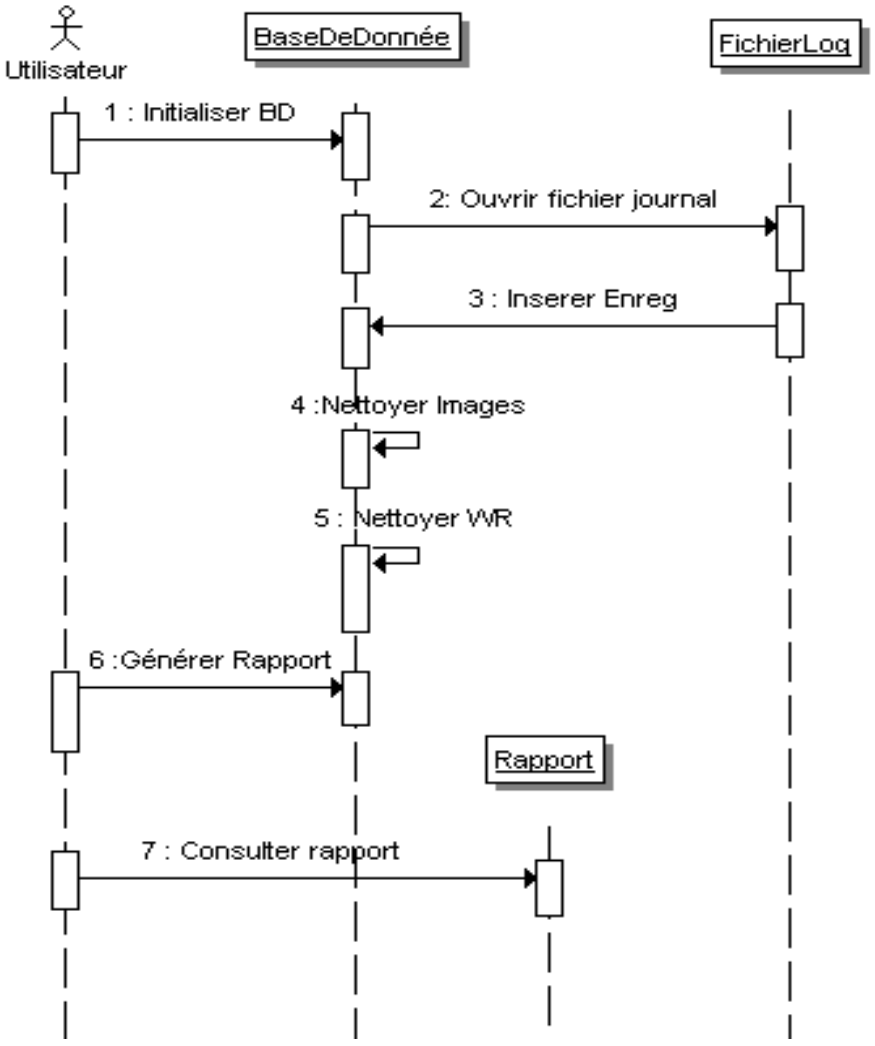


Figure 3.3 : Le diagramme de séquence : Génération de rapport

3.2.2.3 Diagrammes d'activité

- **Interface Homme Machine**

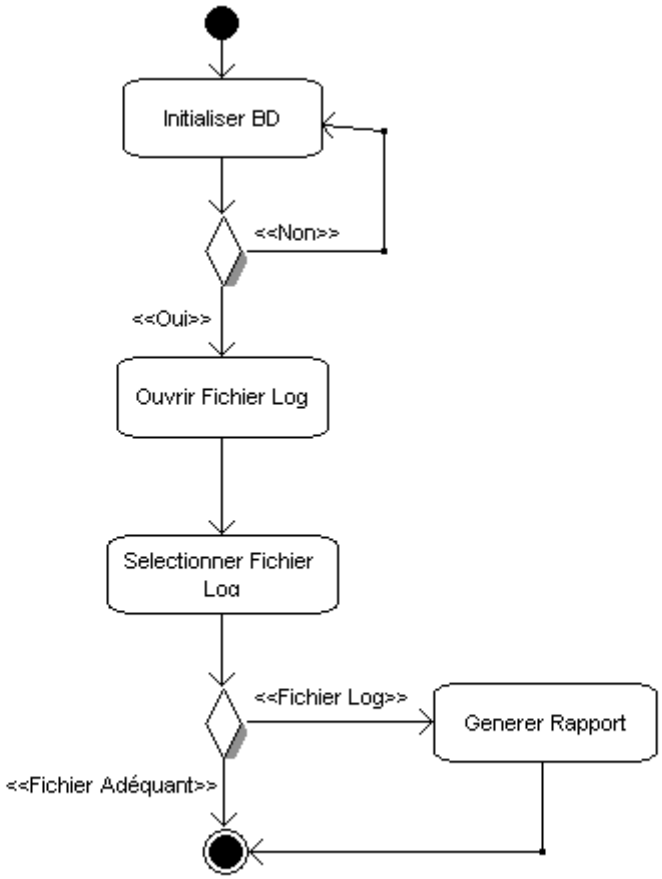


Figure 3.4 : Le diagramme d'activité : IHM

- **Insérer le contenu d'un fichier log dans la base de données**

Lors de l'analyse lexicale du fichier «log», une solution se présente. Chaque ligne lue à partir du fichier est stockée dans une variable intermédiaire comme chaîne de caractères, ensuite elle est découpée champ par champ en tenant compte des délimiteurs qui sont les espaces et dont le nombre diffère d'un champ à un autre. Ces champs seront stockés par suite dans des variables intermédiaires avant qu'ils ne soient passés en paramètre à la méthode **insererEnreg ()** responsable de l'insertion de la ligne dans la classe **BaseDeDonnées**.

En ce qui concerne le stockage dans la base, nous avons opté pour l'insertion des données au fur et à mesure qu'on lit le fichier, au lieu de lire le fichier tout entier et ensuite insérer les données dans la base.

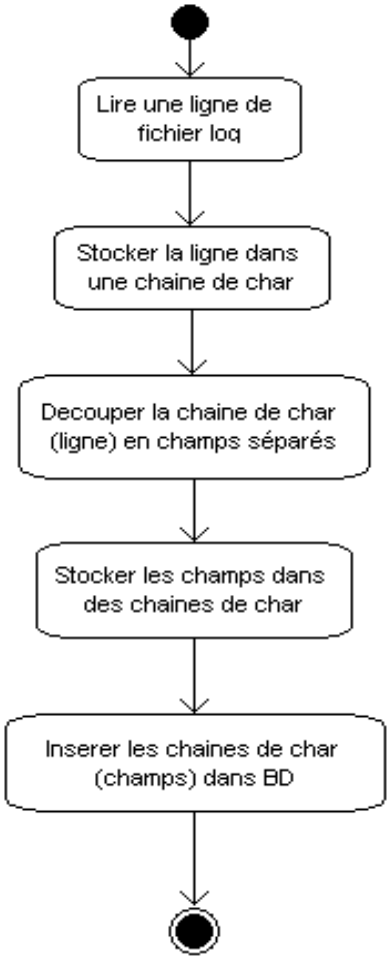


Figure 3.5: Le diagramme d'activité : Log → BD

- **Description de la fonction page_populaire**

  Pour savoir quels sont les pages les plus intéressantes pour les visiteurs :

  - Parcourir la base de données suivant le champ URL1.
  - Enregistrer toutes les urls dans un tableau de chaine de caractères.
  - Eliminer les occurrences de chaque url.
  - Le tableau contient seulement les pages uniques.

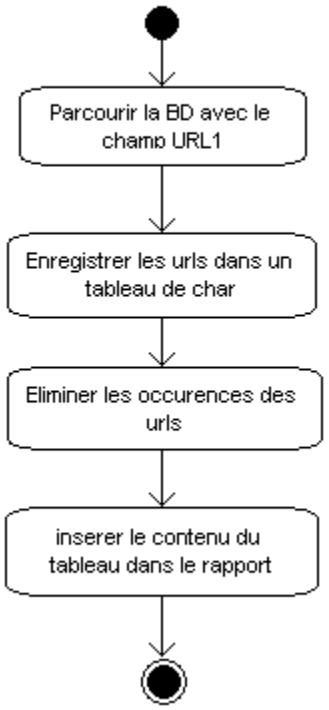


Figure 3.6: Le diagramme d'activité : page_populaire

- **Description de la fonction tous les jours**

Pour savoir l'activité de visiteur pour chaque jour de la période du rapport :

- Parcourir la base de données suivant le champ DATE_HEURE.
- Séparer la date de l'heure.
- Enregistrer toutes les dates dans un tableau de chaine de caractères.
- Eliminer les occurrences de chaque date.
- Le tableau contient seulement les dates uniques.

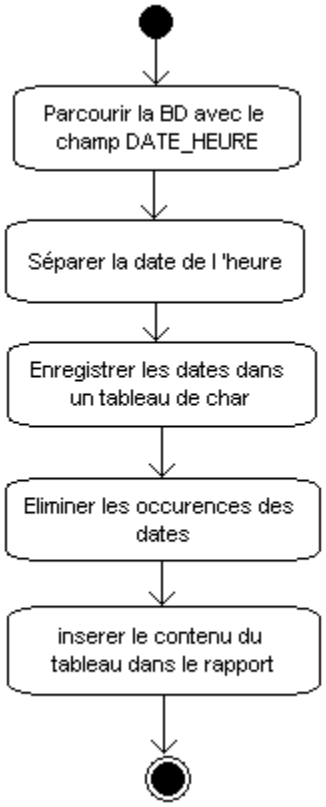


Figure 3.7: Le diagramme d'activité : Tous les jours

3.2.3 La base de données

Nous allons commencer par stocker toutes les informations extraites du fichier log

dans une seule table principale appelée **TableLog** de sorte qu'on ait une représentation fidèle du fichier « log » et en même temps des enregistrements mieux adaptés a nos calculs. Cette table comporte les champs :

• La colonne **« Ip »** correspond aux adresses IP des visiteurs.

• La colonne **« Date_Heure »** correspond à la date d'accès.

• La colonne **« Méthode »** correspond à la méthode utilisée (GET/POST).

• La colonne **« URL1 »** correspond au URL demandé

• La colonne **« Code_De_Re ».**

• La colonne **« VolumeP »** correspond à la taille chargée.

• La colonne **«URL2»** correspond à la page d'origine.

• La colonne **« Navigateur »** correspond au navigateur utilisé.

• La colonne **« OS »** correspond au système d'exploitation utilisé.

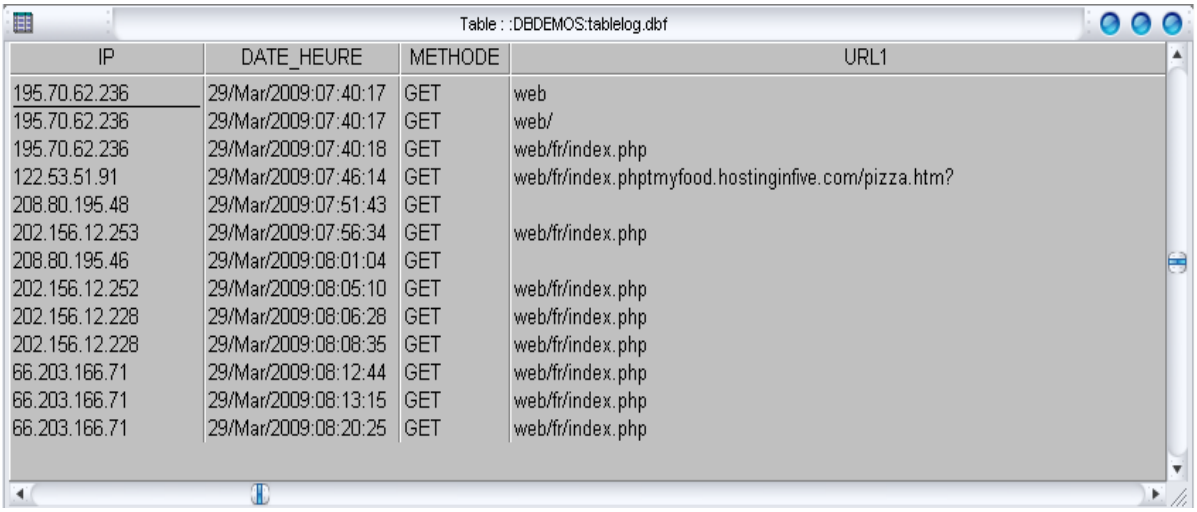


Table : :DBDEMOS:tablelog.dbf

IP	DATE_HEURE	METHODE	URL1
195.70.62.236	29/Mar/2009:07:40:17	GET	web
195.70.62.236	29/Mar/2009:07:40:17	GET	web/
195.70.62.236	29/Mar/2009:07:40:18	GET	web/fr/index.php
122.53.51.91	29/Mar/2009:07:46:14	GET	web/fr/index.phptmyfood.hostinginfive.com/pizza.htm?
208.80.195.48	29/Mar/2009:07:51:43	GET	
202.156.12.253	29/Mar/2009:07:56:34	GET	web/fr/index.php
208.80.195.46	29/Mar/2009:08:01:04	GET	
202.156.12.252	29/Mar/2009:08:05:10	GET	web/fr/index.php
202.156.12.228	29/Mar/2009:08:06:28	GET	web/fr/index.php
202.156.12.228	29/Mar/2009:08:08:35	GET	web/fr/index.php
66.203.166.71	29/Mar/2009:08:12:44	GET	web/fr/index.php
66.203.166.71	29/Mar/2009:08:13:15	GET	web/fr/index.php
66.203.166.71	29/Mar/2009:08:20:25	GET	web/fr/index.php

Figure3.8: Base de données après import

Chapitre 4

Réalisation

Sommaire

Dans ce chapitre, nous présentons le travail réalisé, le choix de la plate forme utilisée ainsi que l'environnement de développement, nous commentons les différentes interfaces graphiques ainsi que quelques courbes et statistiques obtenues.

4.1 L'environnement de développement

4.1.1 Environnement de développement

- **Outil de spécification et conception**

Notre choix s'est porté sur le logiciel **Bouml version 3.4** qui est un produit **Sybase** opérant sur la plateforme Windows. Cet outil supporte tous les modèles du langage unifié de modélisation UML en couvrant toutes les étapes du cycle de développement du logiciel.

- **Langage de programmation**

Nous avons opté pour le langage de programmation **Pascal** sous **Delphi et l'HTML.** Ce choix se justifier par la simplicité avec laquelle il permet d'analyser et de traiter efficacement du texte structuré. En effet, **Delphi** met en œuvre plusieurs astuces de programmation qui facilite l'extraction des informations d'un fichier historique (log) ou d'une base de données. S'ajoute à cet argument, la richesse de ses bibliothèques graphiques et la fiabilité de ce langage.

L' HTML fait davantage parler de lui parce qu'étant plus facilement implémentale et plus pragmatique dans l'utilisation des balises qu'il introduit.

- **Outil de développement**

Le choix de l'outil de développement s'est porté sur le logiciel **Delphi7.**

Delphi7nous permet de concevoir et de développer des applications de haut calibre, en exploitant son éditeur graphique on peut aisément développer les interfaces relatives à nos programmes en leurs donnant une meilleure visibilité et plus de compréhension du coté de l'utilisateur.

- Pour changer l'apparence des formes de notre application, on a utilisé l'outil **SkinEngine.**

4.1.2 Environnement matériel

- Processeur Intel(R) Pentium(R) M processor fréquence d'horloge 1.70 GHz.
- 2 GO de RAM.
- 80 GO de taille de stockage de disque.
- Ecran 15,4 pouces.

4.2 Présentation de LOG+

La plus grande partie de l'application est celle de la navigation entre les différentes Statistiques qu'on arrive à dégager à partir du fichier Access.log. Dans la partie ci-dessous on présente les enchaînements à suivre ainsi que quelques imprimes écrans pour donner un aperçu sur l'utilisation de notre analyseur.

4.2.1 Interface d'accueil

Pour l'exploration et l'analyse du fichier Log, un outil logiciel a été conçu et réalisé : **« LOG+ »**.

L'interface d'accueil est la page de garde de notre outil qui contient son menu principal et qui va donner l'accès aux statistiques. La figure 4.1 illustre l'interface d'accueil de notre application.

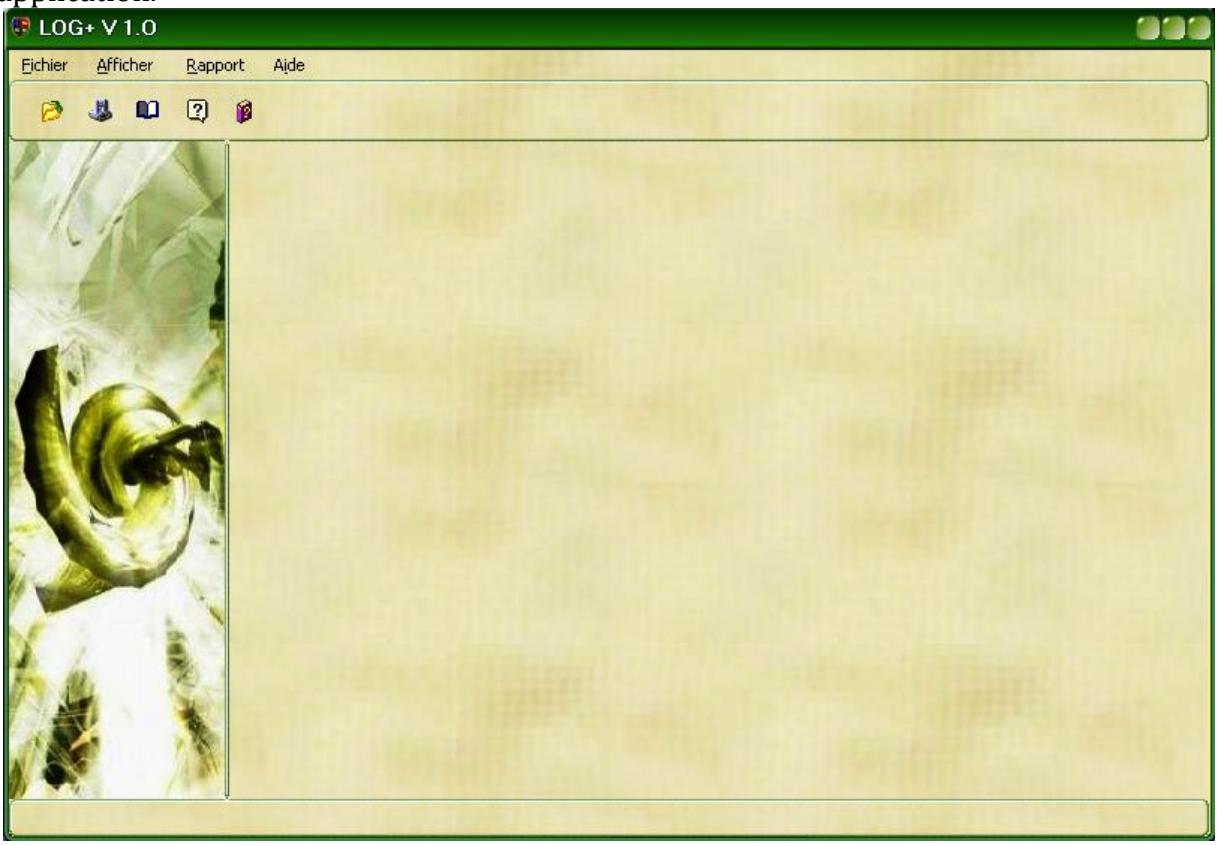


Figure 4.1 : L'interface d'accueil : LOG+

4.2.2 Importation de fichier log

Pour initialiser ou vider la base de données :

- Cliquer sur « **Initialiser la base de données** » dans le menu Fichier.

La figure 4.2 illustre l'initialisation de notre Base de données.

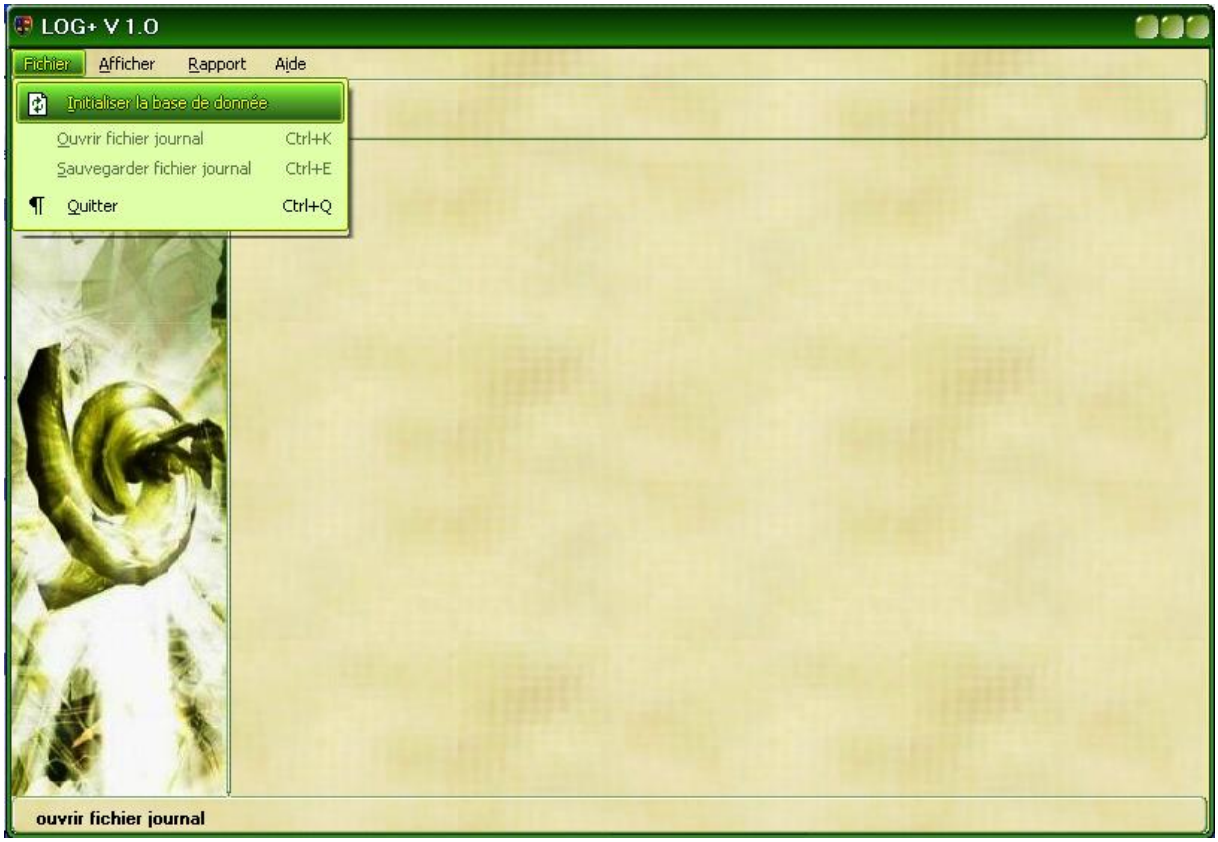


Figure 4.2 : Initialiser la base de données de LOG+.

Pour importer ou ouvrir un fichier log :

- Cliquer sur « **Ouvrir un fichier journal** » dans le menu Fichier.
- Cliquer sur le bouton « **Parcourir** » dans le panneau «**chemin du fichier log**».
- Sélectionner le fichier « **log** » à analyser.

On peut aussi cliquer sur le bouton « **Ouvrir un fichier journal** » .

La figure 4.3 et la figure 4.4 et la figure 4.5 illustrent l'ouverture du fichier journal.

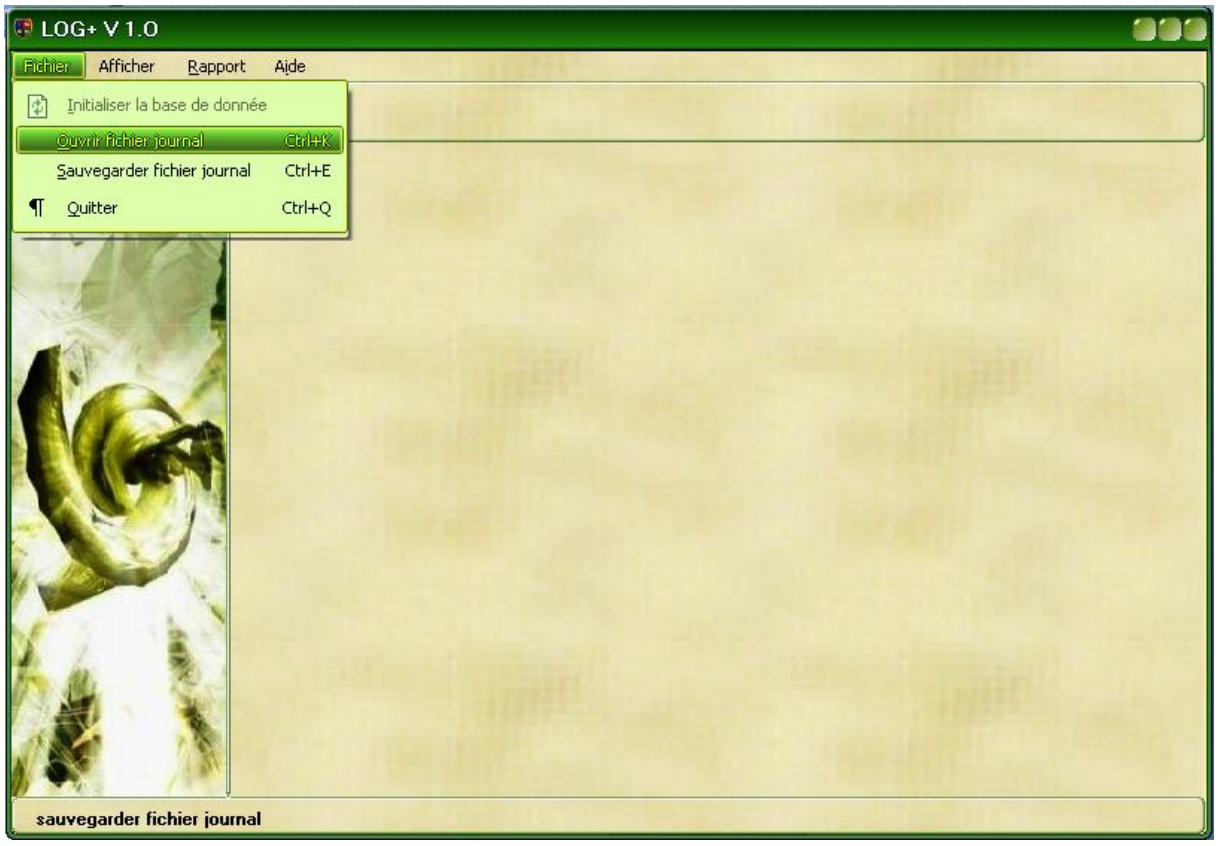


Figure 4.3 : Ouvrir un fichier journal : étape 1

Figure 4.4 : Ouvrir un fichier journal : étape 2

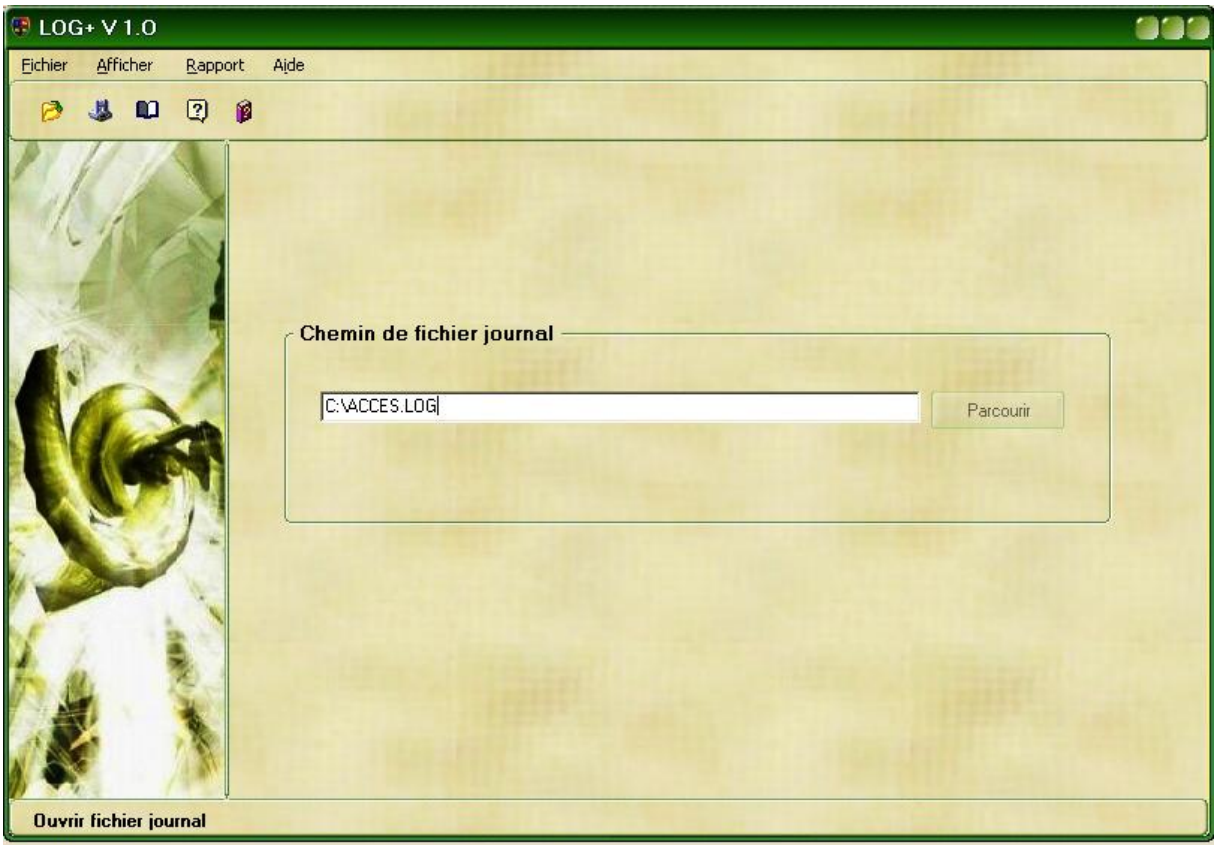


Figure 4.5 : Ouvrir un fichier journal : étape 3

4.2.3 Rapport

Une autre fonctionnalité offerte par notre outil d'analyse est la génération de rapports détaillés des activités des visiteurs.

Après la sélection du fichier journal a analysé il reste à :

- Cliquer sur « **Générer rapport** » du menu rapport pour procéder à la génération des statistiques déduites du fichier « journal » et qui concernent les pages, les visiteurs, les navigateurs, les systèmes d'exploitations.

On peut aussi cliquer sur le bouton «**Générer rapport**».

La figure 4.6 illustre l'ouverture la génération du rapport de notre application.

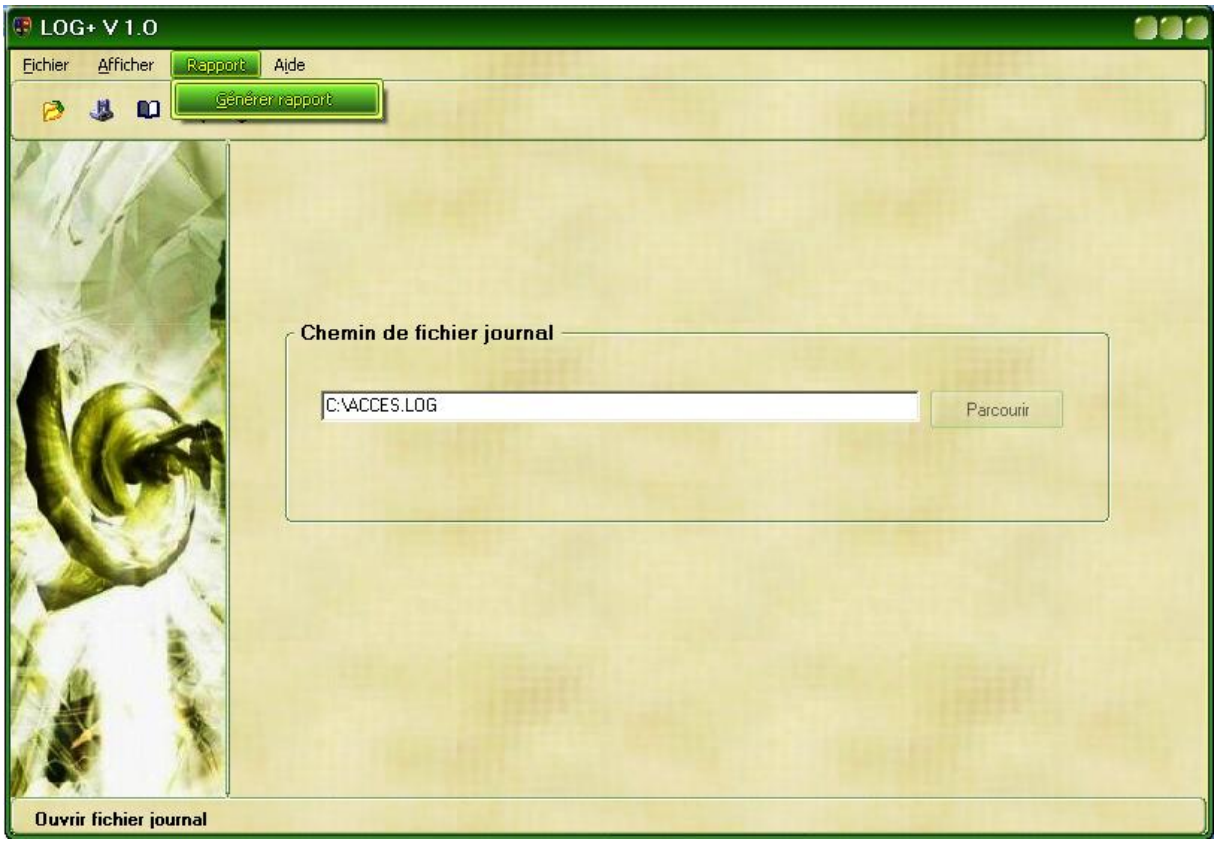


Figure 4.6 : Générer rapport.

La figure 4.7 représente un exemple de rapport sous format HTML de l'activité de l'administrateur.

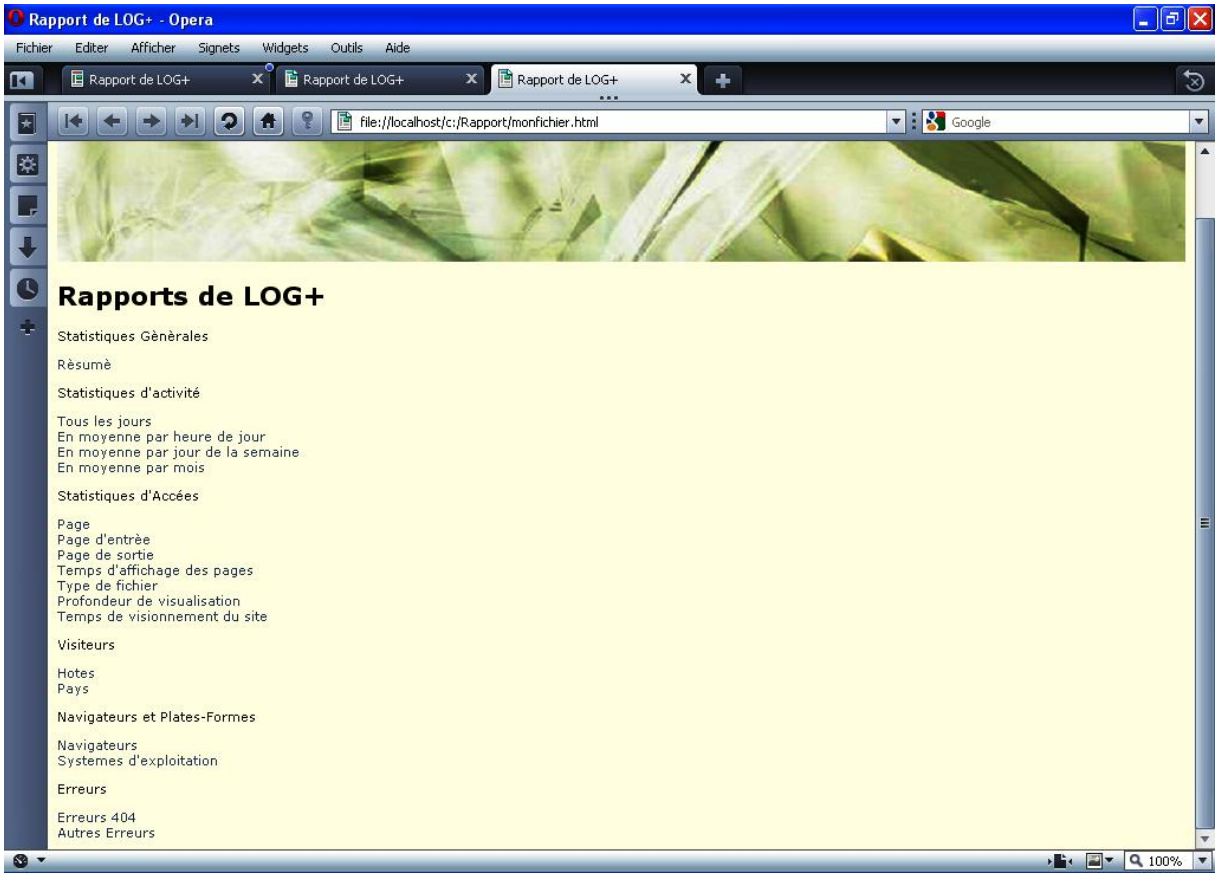


Figure 4.7 : Le rapport d'activités détaillé

- **Les statistiques générales**

Ce rapport contient les statistiques descriptives basées sur les facteurs essentiels (Visiteurs au total, Accés au Total, Visualiations de Page au Total). Ici vous pouvez évaluer l'efficacité générale de votre site web et l'efficacité de vos efforts.

- **Les statistiques d'activités**

  - **Tous les jours**

Ce rapport contient l'information sur l'activité de visiteur pour chaque jour de la période du rapport. La figure 4.8 illustre le cas du choix des statistiques d'activités (tous les jours).

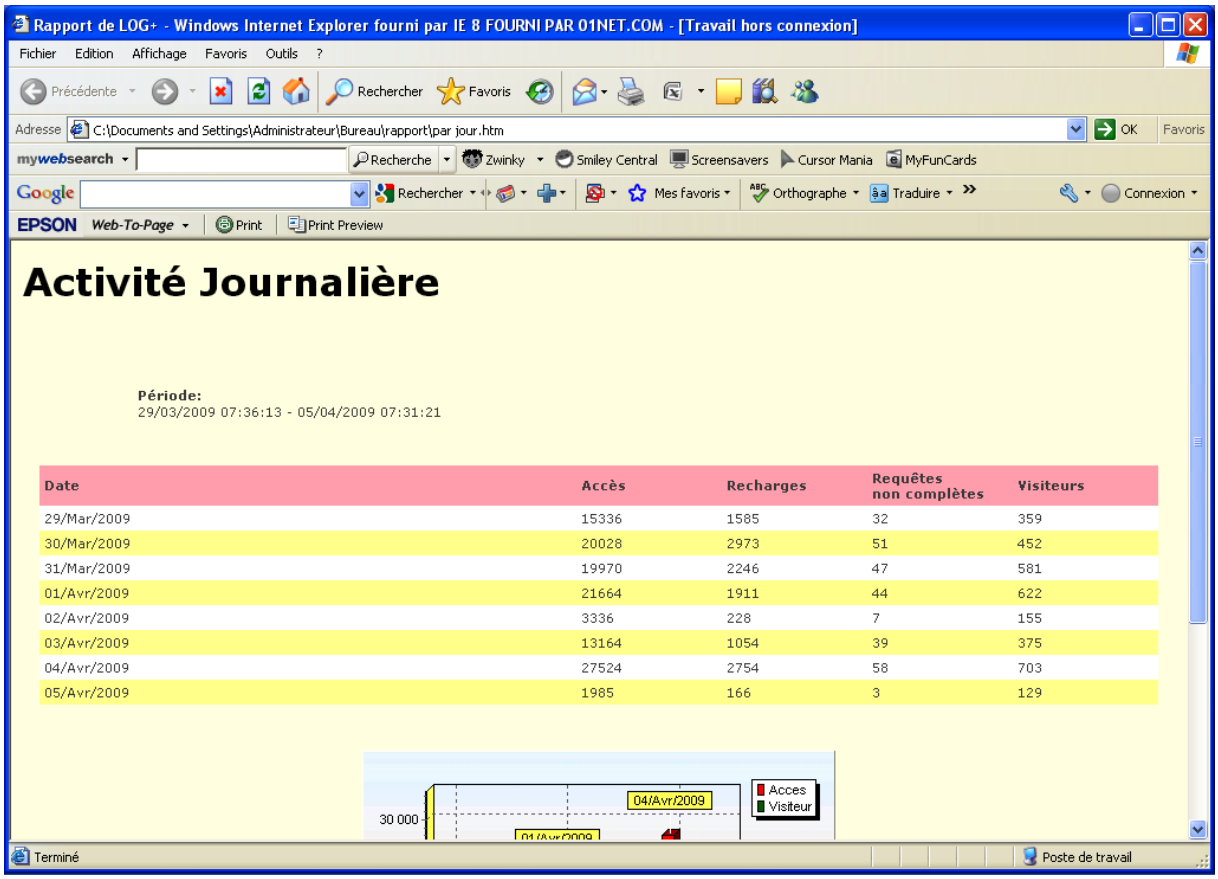


Figure 4.8 : l'activité de visiteur pour chaque jour de la période du rapport.

- **En moyenne par mois**

Ce rapport montre l'activité moyenne de visiteur pour chaque mois.

- **Les statistiques d'accès**

  - **Page**

    Ce rapport contient l'information sur les visites des différentes pages d'un site web. Ici vous pouvez évaluer quels sont les pages les plus intéressantes pour vos visiteurs. Vous devez faire plus attention aux pages qui sont visitées moins souvent. Elles sont peut-être difficiles à accéder ou bien leurs liens ne sont pas intéressants aux visiteurs.

    La figure 4.9 illustre le cas du choix des statistiques propres aux pages.

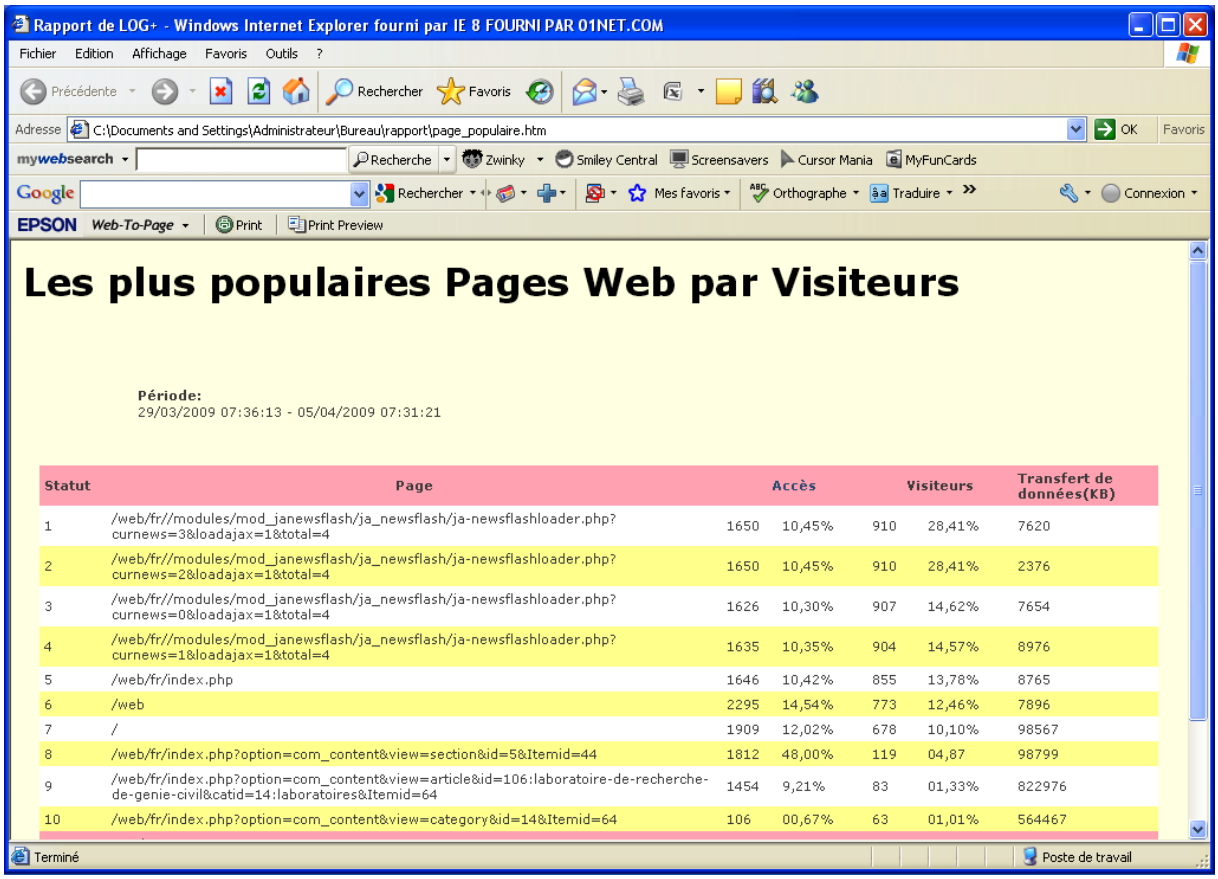


Figure 4.9: Les statistiques sur les pages les plus visités

- **Page d'Entrée**

  Ce rapport contient les pages à partir de laquelle les visiteurs commencent leur visualisation d'un site Web.

- **Page de Sortie**

  Ce rapport contient les pages à partir de laquelle les visiteurs quittent d'un site web. Cette mai se produire en raison de moins intéressant dans la page de contenu ou en raison de gênant la navigation. Analyser les raisons que les visiteurs quittent le site Web vous aidera à accroître l'efficacité de ce site web et d'attirer plus de clients.

- **Temps d'Affichage des Pages**

  Au vu de la minimale, maximale, et le temps moyen qu'il faut un visiteur de visualiser une page, vous pouvez évaluer la façon dont le visiteur est intéressé dans le contenu de la page. La durée de temps est définie comme un intervalle entre la demande pour voir la situation actuelle et la page suivante. L'heure de la dernière page ne peut être déterminée. Par conséquent, le nombre de visites dans le rapport est inférieur ou égal au nombre de visites au total.

- **Type de fichier**

Ce rapport montre les statistiques sur l'accès par extensions de fichiers. La figure 4.10 illustre le cas du choix des statistiques propres aux fichiers.

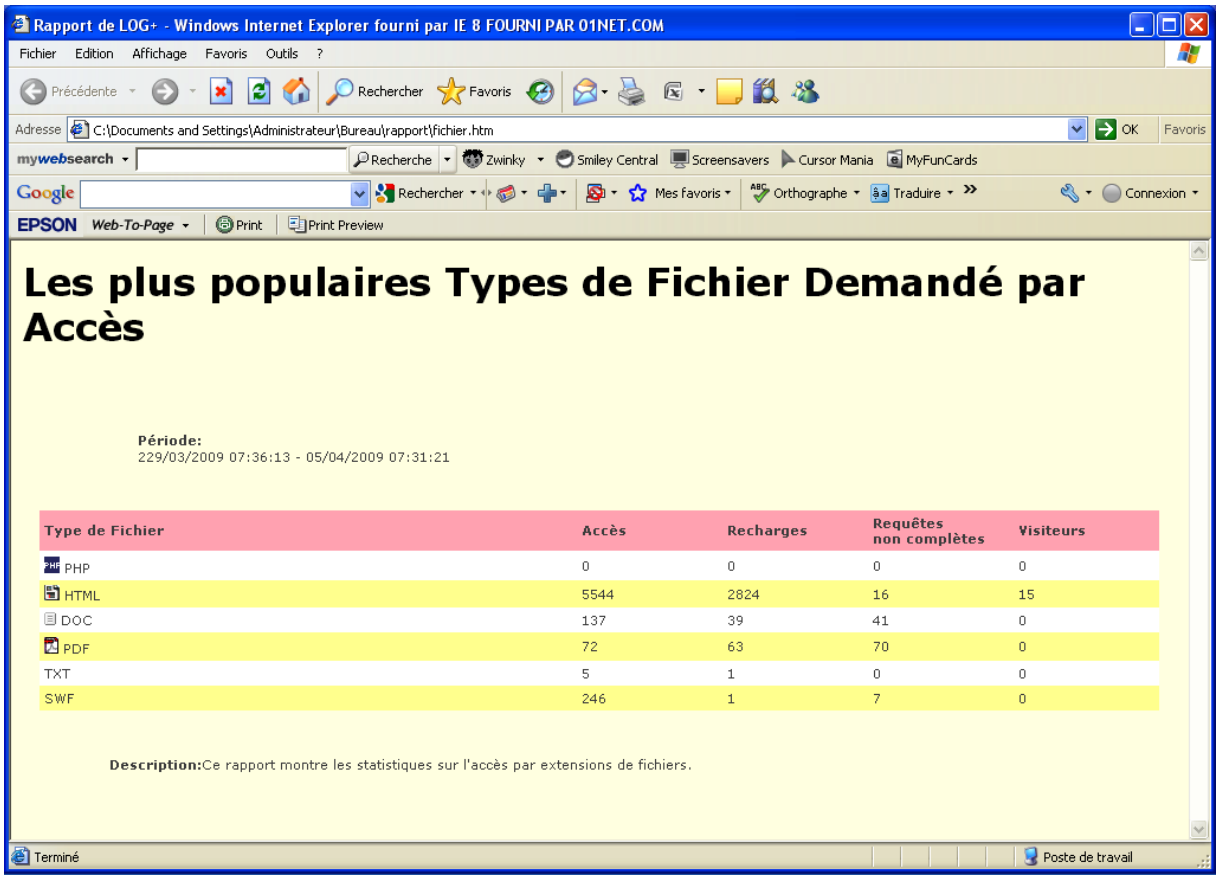


Figure 4.10: L'accès par extensions de fichiers

- **Visiteurs**
  - **Hôtes**

    Ce rapport montre l'activité des visiteurs depuis chaque hôte.
  - **Pays**

    . Ce rapport montre l'activité des visiteurs des pays différents.

    La figure 4.11 illustre le cas du choix des statistiques propres aux Pays des Visiteurs.

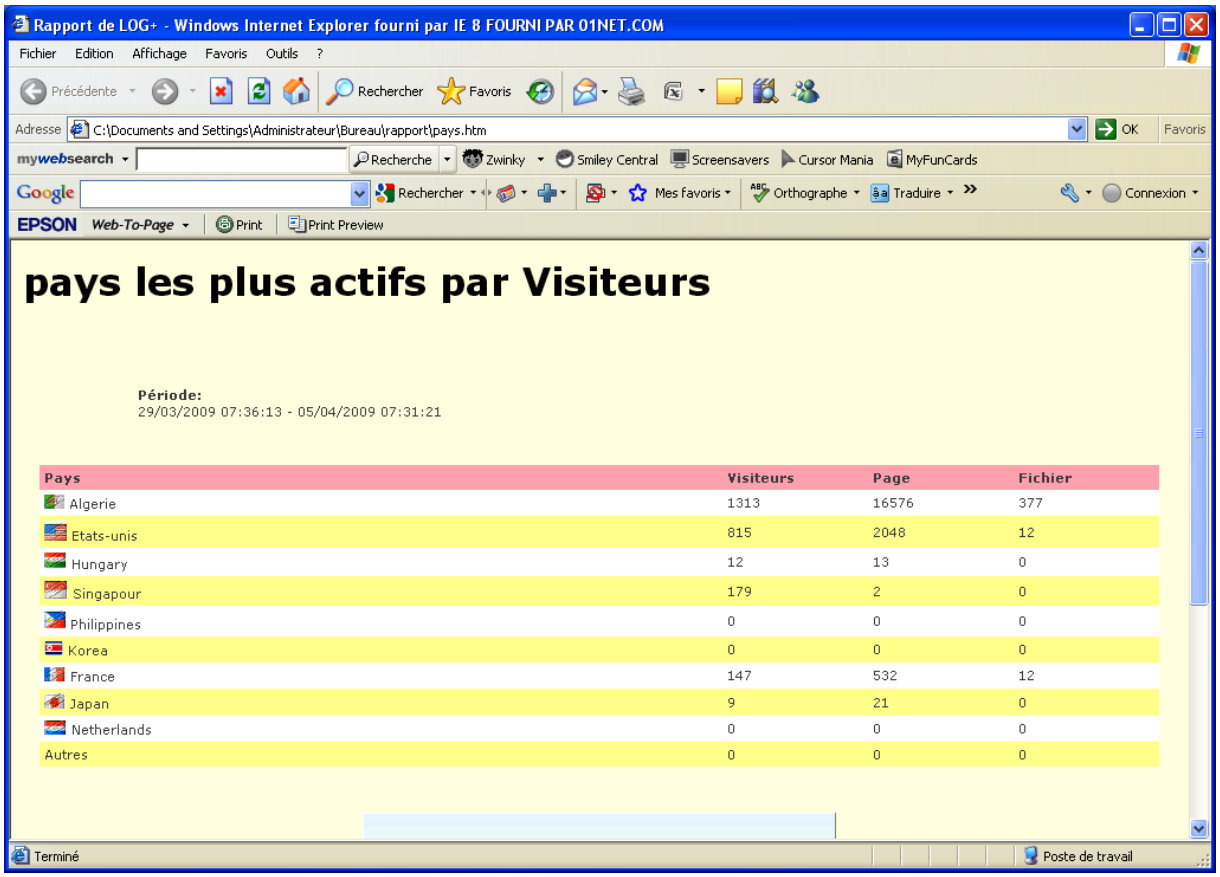


Figure 4.11: L'activité des visiteurs des pays différents.

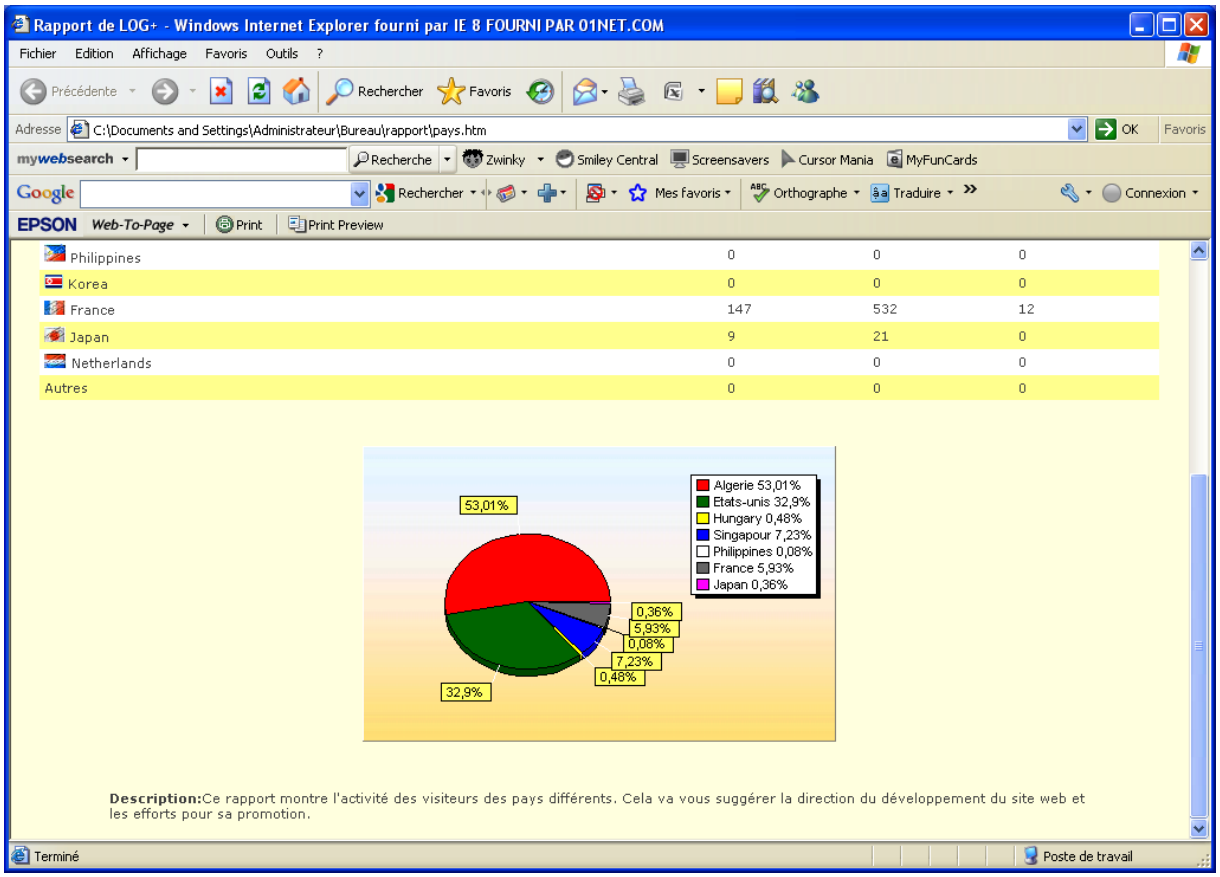


Figure 4.12: Le graphe des visiteurs des pays différents.

- **Navigateurs et Plates-Formes**

  - **Navigateurs**

    Ce rapport contient les statistiques sur les navigateurs qu'utilisent vos.

  - **Systèmes d'exploitation**

    Ce rapport contient les statistiques sur les systèmes d'exploitation qu'utilisent vos visiteurs.
    visiteurs. La figure 4.13 illustre le cas du choix des statistiques propres aux systèmes d'exploitation.

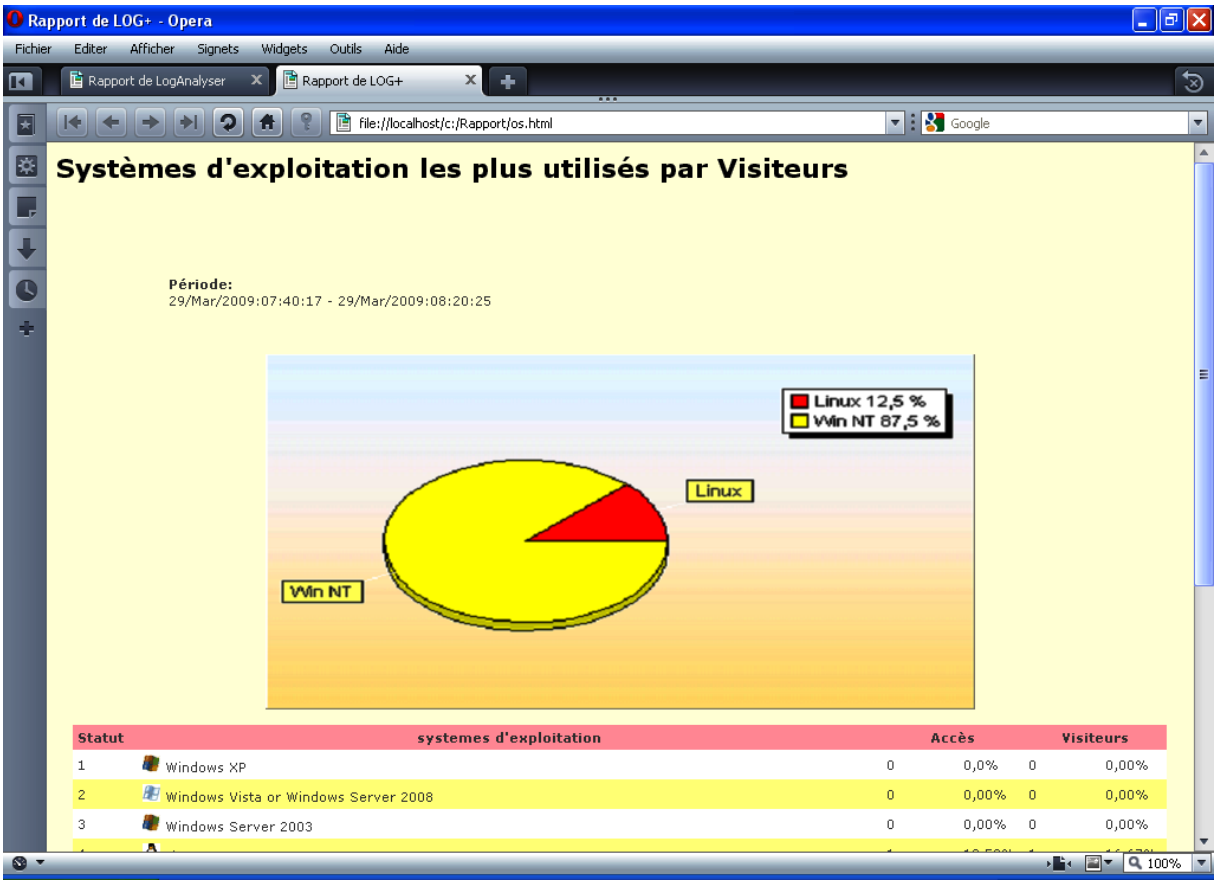


Figure 4.13: Les statistiques sur les systèmes d'exploitation

- **Erreurs**
  - **Erreur 404**

    Ce rapport contient l'information sur la production de l'erreur avec le code 404 code (Page Introuvable).La figure 4.14 illustre le cas du choix des statistiques propres aux Erreurs(404).

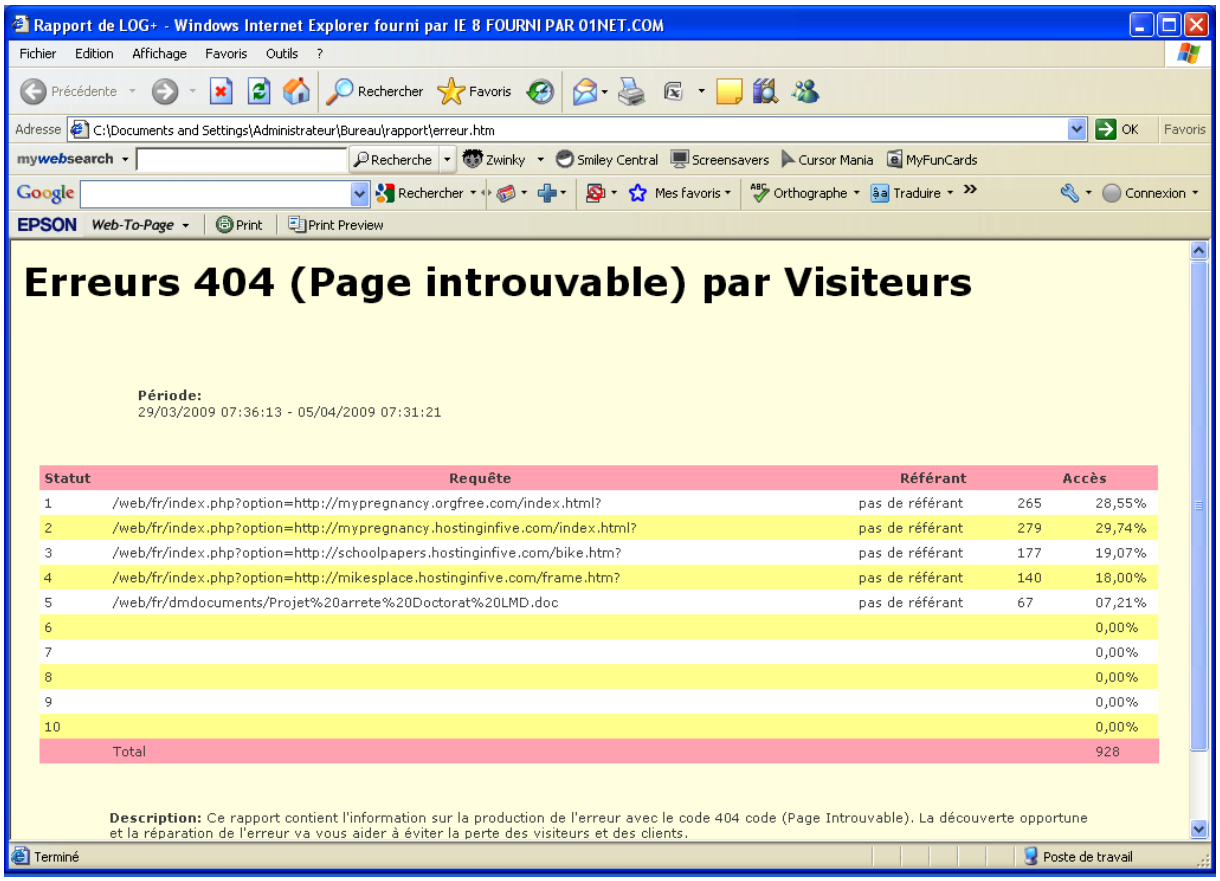


Figure 4.14: Erreur 404

4.2.4 Aide

Pour ouvrir l'aide de **LOG+** :

- Cliquer sur « **Rebrique d'aide** » dans le menu Aide.

On peut aussi cliquer sur le bouton «**Aide** » ou sur **F1**.

La figure 4.15 et la figure 4.16 illustrent l'ouverture d'aide.

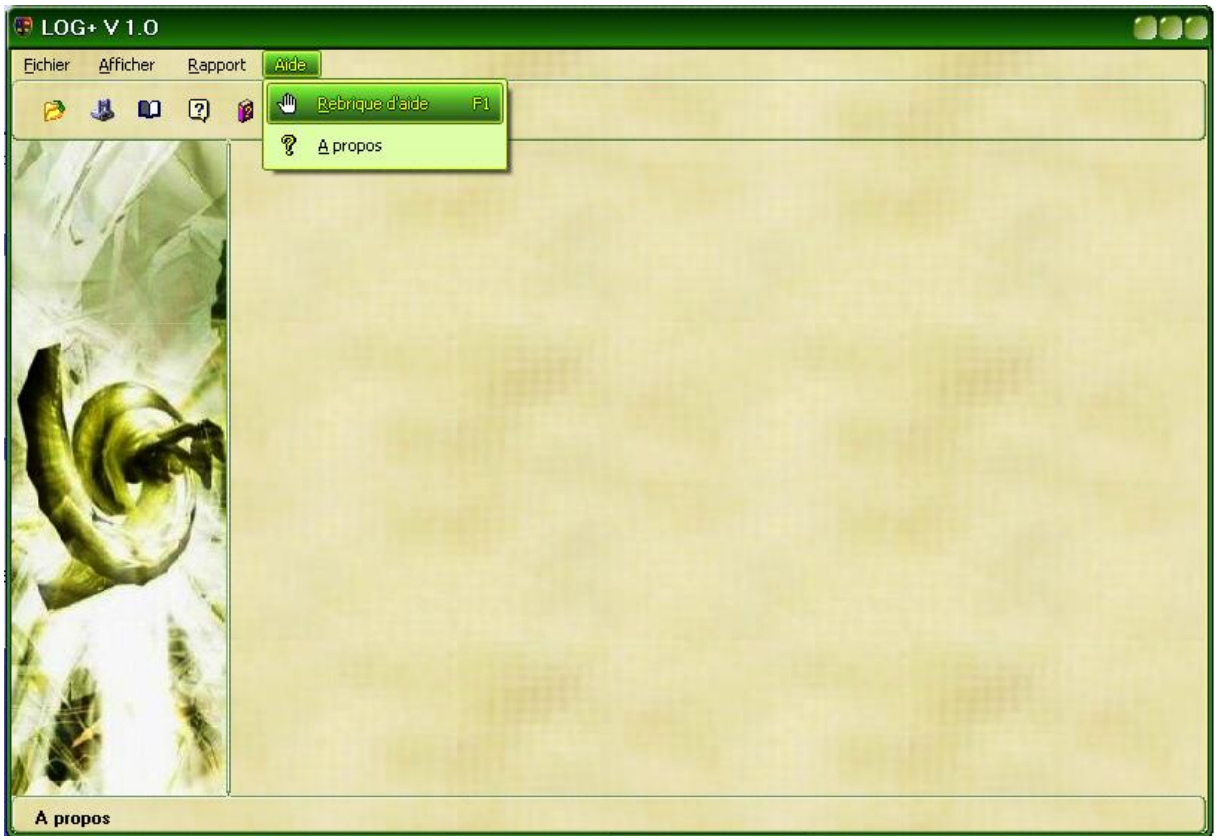


Figure 4.15 Ouverture de l'aide de LOG+

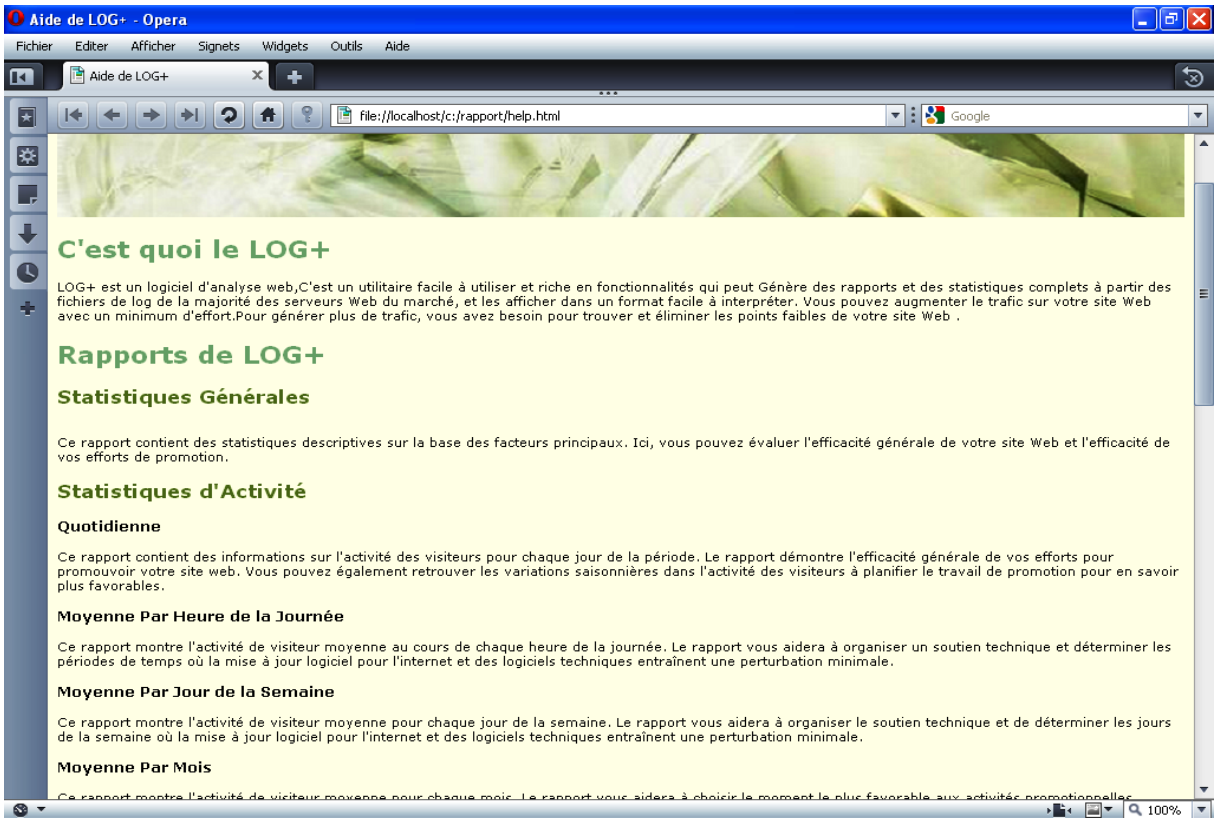


Figure 4.16 L'aide de LOG+

Conclusion générale

à l'heure où le nombre d'utilisateurs de l'Internet augmente exponentiellement et par conséquent les données à analyser deviennent de plus en plus volumineuses. les propriétaires de ce site Web cherchent à comprendre le comportement des visiteurs de leur site pour leur offrir un contenu personnalisé répondant à leurs besoins.

Dans notre travail nous nous sommes intéressés à une application dans le domaine du WUM. Nous avons réalisé une conception et une implémentation d'un outil d'aide pour le webmaster.

L'objectif étant de rechercher des connaissances permettant de mieux comprendre le comportement des visiteurs d'un site web, afin d'améliorer des services offerts sur ce site. L'outil réalisé était testé sur les données du serveur web de l'université Ammar Telidji Laghouat : www.lag-univ.dz de Laghouat.. Nous souhaitons que cet outil rendu un service au webmaster de l'université dans ça tâche journalière.

Glossaire

Ce glossaire contient les termes techniques et de spécialité les plus employés dans cette thèse.

Adresse IP Adresse de 32 bits utilisée par le protocole IP pour identifier de manière unique les hôtes (machines) sur réseau IP.

404 (Erreur 404)

C'est ce qui arrive lorsqu'aucune page n'est disponible à l'adresse que vous avez demandée. Peut-être que l'adresse est fausse (faute de frappe ?) ou que la page a déménagé.

HTML(HyperText Markup Language)

Le HTML est un système qui formalise l'écriture d'un document avec des balises de formatage indiquant la façon dont doit être présenté le document et les liens qu'il établit avec d'autres documents.

Navigateur (browser)

Le navigateur est le logiciel qui interprète les adresses des pages Web, les affiche et permet d'en exploiter les contenus. Les principaux navigateurs sont Internet Explorer de Microsoft, Mozilla (open source) et Safari (de Apple).

Pages Vues Nombre de pages affiché par l'internaute, pour un site donné, et pour une période donnée.

Page Web Unité ergonomique élémentaire d'un site Web, désignée par une URL unique. Une page contient souvent plusieurs éléments (images, *frames*, etc.), qui occasionnent chacun une requête auprès d'un serveur Web, et qui sont assemblés par le navigateur.

Requête Dans un modèle client-serveur, envoi d'une instruction d'un client vers un serveur. Dans le cas des serveurs Web, une requête HTTP demande à un serveur l'envoi du contenu d'un document (ex. : une page HTML) ou du résultat de l'exécution d'un traitement par le serveur (ex. : rechercher dans une base de données).

Serveur

Composant logiciel et/ou matériel assurant la disponibilité, la distribution, le service transactionnel de l'information. Il gère le partage, la sécurité et la cohérence de l'information.

Site Web

Page d'accueil et les pages, graphismes, documents, éléments multimédias ou autres fichiers créés dans FrontPage qui lui sont associés, stockés sur un serveur Web ou sur le disque dur d'un ordinateur.

Statistiques Les techniques statistiques sont des techniques mathématiques permettant de recueillir et d'analyser des données.

Text mining

Technique permettant d'automatiser le traitement de gros volumes de contenus texte pour en extraire les principales tendances et répertorier de manière statistique les différents sujets évoqués. Les techniques de text mining sont surtout utilisées pour des données déjà disponibles au format numérique. Sur Internet, elles peuvent être utilisées pour analyser le contenu des e-mails entrant ou les propos tenus sur des forums.

Webmaster (webmestre)

La définition varie suivant les entreprises, mais le Webmaster devrait être celui qui gère la technique du site. Pas forcément développeur, c'est lui qui peut intervenir sur le site pour régler les problèmes techniques.

World Wide Web

Ensemble intégral des documents hypertexte interconnectés résidant sur les serveurs HTTP du monde entier. Les documents du World Wide Web sont appelés pages ou pages Web. Ces pages sont écrites en langage HTML (Hypertext Markup Language). Elles sont identifiées par des URL (Uniform Resource Locators) qui spécifient l'ordinateur et le nom du chemin grâce auxquels un fichier pourra être accédé et transmis de nœud en nœud jusqu'à l'utilisateur final via le protocole HTTP (Hypertext Transfer Protocol).

XML (eXtensible Markup Language)

Langage de balisage décrivant le rôle des données dans une page de site. Plus puissant qu'HTML, XML permet de créer des liens pointant chacun vers plusieurs documents (tandis qu'HTML ne permet de pointer que vers une destination unique).

Références

[1] B.Ziani, Application des techniques de fouille de données à l'analyse des fichiers logs, mémoire de magistère, 2007.

[2] NGONGANG BADJIO, Rapport de stage Pré-Ingénieur réalisé par Guy Carrel, août – novembre 2005.

[3] http://www.memoireonline.com/07/08/1307/m_techniques-extraction-de-connaissances- appliquees-aux-donnees-du-web0.html

[4] Sulliman Omarjee, Le data mining : Aspects juridiques de l'intelligence artificielle au regard de la protection des données personnelles, 2001/2002.

[5] D. Tanasa et B. Trousse, »Le prétraitement des fichiers Logs Web dans le Web Usage Mining Multi-sites,» In Journées Francophones de la Toile, Juin-Juillet 2003.

[6] L. Tauscher et S. Greenberg, »How People Revisit Web Pages : Empirical Findings and plications for the Design of History Systems,» International Journal of Human Computer Studies, Special issue on World Wide Web Usability 47(1), pp. 97-138,1997.

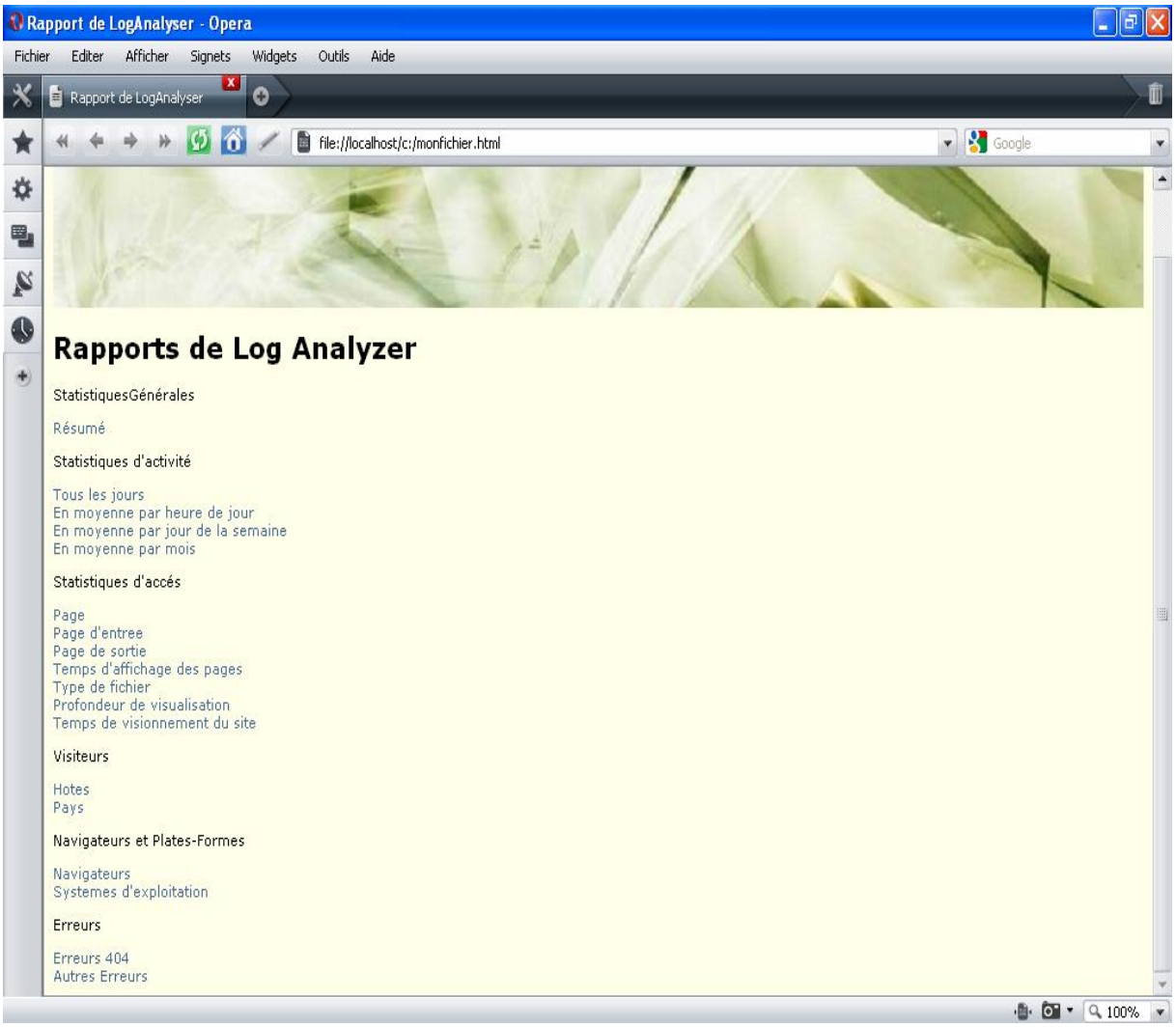
[7] D. Pierrakos, G. Paliouras, C. Papatheodorou, et C.D. Spyropoulos, »Web Usage Mining as a tool for personalization : A survey,» User Modeling and User-Adapted Interaction 13 (4), pp.311-372, 2003.

[8] K. Chevalier, V. Corruble, et C. Bothorel, »SURFMINER: Connaître les utilisateurs d'un site,» in Documents Virtuels Personnalisables, Brest, Juillet 2002.

[9] C. Michel, »Caractérisation d'usages et personnalisation d'un portail pédagogique. État de l'art et expérimentation de différentes méthodes d'analyse du Web Usage Mining,» 7ème colloque de l'AIM Affaire électronique et société de savoir : Opportunités et Défis, 29 mai-1er juin, 2002.

Résultat généré par LOG+

Dans cette annexe, nous présentons un exemple de page Web de notre application. Nous aurons détaillé la page web principale «Rapports de Log Analyser ».



Cette page est la page d'accueil de notre programme. Elle englobe les rapports importants concernant les requêtes des internautes de site.

Le premier rapport résume en générale les statistiques descriptives qui sont basées sur des facteurs essentiels. Il contient des statistiques sur :

- **Les accès** tel que : Accès au Total, Accès en moyenne par Jour, Accès en moyenne par Visiteur, Recharges au Total, Recharges en moyenne par Visiteurs, Requêtes non complètes, Requêtes échouées.

- **Les visiteurs** tel que : Visiteurs au Total, Visiteurs en moyenne par Jour, Time Moyen Passé.
- **Les Ressources Accédées** : Visualisations de Page au Total, Visualisations de Page en moyenne par Jour, Visualisations de Page en moyenne par Visiteur.

Ensuite nous décrivons les rapports détaillés les activités des utilisateurs, en effet au fur et à mesure de la navigation entre les différentes statistiques.
Il est décomposé en plusieurs rapports :

- **Tous les jours**

Ce rapport contient des informations sur l'activité des visiteurs pour chaque jour de la période. Le rapport démontre l'efficacité générale de l'effort pour promouvoir le site web. On peut également retrouver les variations saisonnières dans l'activité des visiteurs à planifier le travail de promotion pour en savoir plus favorables.

- **Moyenne Par Heure de la Journée**

Ce rapport montre l'activité de visiteur moyenne au cours de chaque heure de la journée. Le rapport aidera à organiser un soutien technique et déterminer les périodes de temps où la mise à jour logiciel pour l'internet et des logiciels techniques entraînent une perturbation minimale.

- **Moyenne Par Jour de la Semaine**

Ce rapport montre l'activité de visiteur moyenne pour chaque jour de la semaine. Le rapport aidera à organiser le soutien technique et de déterminer les jours de la semaine où les mises à jour logicielles pour l'internet et des logiciels techniques entraînent une perturbation minimale.

- **Moyenne Par Mois**

Ce rapport montre l'activité de visiteur moyenne pour chaque mois. Le rapport aidera à choisir le moment le plus favorable aux activités promotionnelles.

Ensuite nous décrivons les rapports détaillés les accès de l'utilisateur, tel que la page visitée, le type de fichier à télécharger... les rapports sont :

- **Page**

Ce rapport contient des informations sur les visites à différentes pages de votre site Web. On peut évaluer quelles sont les pages les plus intéressantes pour nos visiteurs. Nous devrions regarder de plus près les pages qui sont visitées et le moins souvent. Peut-être qu'ils sont difficiles à obtenir ou à des liens vers d'entre eux ne sont pas intéressants pour les visiteurs.

- **Page d'Entrée**

Ce rapport contient les pages à partir de laquelle les visiteurs commencent leur visualisation de notre site Web. Ces pages sont les portes de nos visiteurs et en fonction de comment ils les veulent, ils mai devenir nos clients ou à pied.

- **Page de Sortie**

Ce rapport contient les pages à partir de laquelle les visiteurs quittent notre site web. Cette mai se produire en raison de moins intéressant dans la page de contenu ou en raison de gênant la navigation. Analyser les raisons que les visiteurs quittent le site Web vous aidera à accroître l'efficacité de notre site web et d'attirer plus de clients.

- **Temps d'Affichage des Pages**

Au vu de la minimale, maximale qu'il faut un visiteur de visualiser une page, on peut évaluer la façon dont le visiteur est intéressé dans le contenu de la page. La durée de temps est définie comme un intervalle entre la demande pour voir la situation actuelle et la page suivante. L'heure de la dernière page ne peut être déterminée. Par conséquent, le nombre de visites dans le rapport est inférieur ou égal au nombre de visites.

- **Type de Fichier**

Ce rapport montre des statistiques sur l'accès par les extensions de fichier.

- **Profondeur de Visualisation**

Ce rapport indique le nombre de visiteurs qui ont consulté un certain nombre de pages Web. Le rapport qualifie le niveau d'intérêt des visiteurs dans les informations sur le site.

- **Temps de Visionnement du Site**

Ce rapport indique le temps que passent les visiteurs sur le site. L'heure affichée sur le site web est déterminée comme étant la différence entre le moment où la demande de la dernière ressource et le premier. Ce rapport caractérise le niveau d'intérêt des visiteurs dans les informations sur le site.

Après, on a détaillé les statistiques propres aux visiteurs tel que les hôtes et les pays d'où vient le visiteur.

- **Hôtes**

Ce rapport montre l'activité des visiteurs à partir de chaque hôte.

- **Pays**

Ce rapport montre l'activité des visiteurs de différents pays. Il vous propose la direction du développement de sites web et les efforts pour sa promotion.

Ensuite on a les statistique concernant les Navigateurs et Plates-Formes

- **Navigateurs**

Ce rapport contient des statistiques sur les navigateurs que vos visiteurs utilisent.

- **Systèmes d'Exploitation**

Ce rapport contient des statistiques sur les systèmes d'exploitation que vos visiteurs utilisent.

En fin on termine notre rapport de Log Analyser par les erreurs Courantes lors d'une requête telle que 404, 405,403…

- **Erreurs 404**

Ce rapport contient des informations sur les erreurs survenant avec le code 404 (page introuvable). La découverte opportune et une correction d'une erreur nous aidera à éviter de perdre des visiteurs et des clients.

- **Autre Erreurs**

Ce rapport contient des statistiques de l'erreur type ont eu lieu sur votre site web. Le temps de découverte et de correction d'une erreur nous aidera à éviter de perdre des visiteurs et des clients.