

Contributions à l'amélioration de la Qualité de la fouille de l'Usage Web

THÈSE

présentée et soutenue publiquement le 12/12/2019

pour l'obtention du

Doctorat de l'Université Paris 8

Spécialité Informatique

par

Amine Ganibardi

<i>Rapporteurs :</i>	Mme Rim Zitouni Faiz	Professeure des Universités – Université de Carthage
	M. Lakhdar Sais	Professeur des Universités – Université d'Artois
<i>Examineurs :</i>	Mme Nada Matta	Professeure des Universités – Université de Troyes
	M. Gilles Bernard	Professeur des Universités – Université Paris 8
	M. Herman Akdag	Professeur des Universités – Université Paris 8
<i>Directeur :</i>	M. Arab Ali Chérif	Professeur des Universités – Université Paris 8

Remerciements

Je tiens à exprimer, infiniment, mes vifs remerciements, ma sincère reconnaissance et ma gratitude à M. Arab Ali Cherif, professeur à l'université Paris 8 et directeur du Laboratoire d'Informatique Avancée de Saint-Denis ; d'avoir accepté de m'accueillir dans son laboratoire et de diriger cette thèse. Ses conseils précieux et ses interventions ciblées ont été le fil conducteur de l'aboutissement de cette thèse. Je le remercie, encore une autre fois, pour la confiance et l'intérêt accordés tout au long de cette thèse.

Comme je tiens à remercier les membres du jury d'avoir accepté d'évaluer et de discuter mon travail.

Mes remerciements vont aussi à ceux qui ont contribué de près ou de loin à la construction de cette thèse et à maintenir ma volonté d'arriver à son terme ; en particulier :

M. Herman Akdag, professeur à l'université Paris 8, qui s'est toujours intéressé à l'avancement de ma thèse ;

Mme. Jaziri Rakia, maître de conférences à l'université Paris 8, pour sa disponibilité à assister l'avancement de mon travail ;

M. Ivaylo Ganchev, Direction du système d'information et du numérique à l'Université Paris 8, pour sa coopération ;

Je remercie mon frère Rafik pour le temps qu'il a pu consacrer à la révision des manuscrits des publications et de la thèse ;

Enfin, je remercie tous mes collègues du LIASD et de Paris 8, particulièrement, Rabeh, Salah, Ahmet, Lisa, Nourredine, Kathia, Jérôme, et Othman ; pour leur partage de connaissances et de savoir-faire.

Je dédie cette thèse

*A ma Mère et Mon Père pour leur amour,
leurs sacrifices et leur enthousiasme qui
nourrissent chaque moment de ma vie de
fierté et guide chaque pas de mon
existence dont toute ambition tire sa
persévérance de mon aspiration à leur
témoigner mon amour et ma gratitude
A ma chère épouse pour lui exprimer ma
reconnaissance de ses sacrifices et d'avoir
eu le courage et l'altruisme de porter
avec moi mon ambition...*

*A mes chers enfants Maram, Sara, Youcef
et Mohammed pour leur curiosité de tous
les jours d'un papa qui va encore à
l'école...*

*...Pour tout ce qu'ils ont pu endurer
malgré eux, pour moi, je leur dis merci ;
et qu'ils trouvent en ces quelques mots*

*l'expression de mon amour et de ma
gratitude les plus sincères
A ma sœur Dalal et mes frères*

Résumé

Cette thèse présente nos travaux sur la qualité de trois tâches critiques de la fouille des données de l'Usage Web, i.e., le nettoyage, la structuration et la découverte de motifs d'usage. A cet égard, nos contributions analysent les limites des approches actuelles en termes de qualité, et proposent trois nouvelles méthodes d'une meilleure pertinence. En effet, notre recherche et nos contributions abordent l'analyse de l'interaction Homme-Machine sur la base des traces d'utilisation, à travers le cas de l'interaction avec le Web et les techniques de la fouille des données de l'utilisation du Web (*Fouille de l'Usage du Web*).

La première contribution présente une approche de nettoyage centrée sur la structure des données de journalisation au lieu de celles, actuellement, centrées sur le contenu de la journalisation. Cette approche est déclinée en deux variantes, i.e., une méthode heuristique et une autre basée sur les techniques de partitionnement génétique. La deuxième contribution présente une approche de structuration des données de journalisation centrée sur le flux de clics des agents au lieu de celles existantes centrées sur les attributs des agents. Notre troisième contribution propose une approche d'apprentissage semi-supervisé qui permet de découvrir des motifs d'usage pour optimiser, à la fois, plusieurs dimensions (*optimisation symbiotique*) de l'usage du Web au lieu d'une optimisation par dimension qui peut s'avérer conflictuelle et/ou contradictoire (*optimisation parasitique*).

Les résultats expérimentaux des méthodes proposées, comparés à ceux des méthodes actuelles, démontrent des améliorations significatives en termes de pertinence rapportée aux contraintes d'applicabilité et de coût. Le manuscrit présentant le contexte et les contributions de notre recherche est composé de six (06) chapitres.

Le premier chapitre introductif positionne nos travaux de recherche portant sur la qualité de la fouille de l'usage du Web par rapport au contexte et à la problématique de l'analyse de l'interaction Homme-Machine basée sur les traces d'utilisation. Ce chapitre met l'accent sur trois points, i.e., le cadre disciplinaire de l'analyse de l'interaction Homme-Machine (*Human-Computer Interaction – HCI*) et la Fouille des Données de l'Utilisation du Web « Fouille de l'Usage du Web (*Web Usage Mining – WUM*) », les avantages de la fouille des données de l'utilisation du Web en matière d'analyse de l'interaction Homme-Machine sur la base des traces d'utilisation, les limites, et les difficultés sujettes à contribution. Le deuxième chapitre présente un état de l'art général de la fouille de l'usage du Web basée sur l'analyse des données de journalisation des serveurs. Il aborde le cadre théorique et conceptuel de la discipline, sa finalité, son processus, ses techniques, et leurs applications. Aussi, il détaille les aspects techniques du système Web, du Système de Journalisation Web, des Données de Journalisation Web, et des Techniques de Fouille des Données de Journalisation Web.

Le troisième chapitre présente notre première contribution. Il aborde les méthodes actuelles de nettoyage des données de journalisation Web, qui sont centrées sur le contenu. Il

illustre leurs limites et démontre les avantages de l'approche proposée, qui est centrée sur la structure de la journalisation. L'approche en question est déclinée en deux méthodes, i.e., heuristique et une autre basée sur les techniques de partitionnement génétique. Le quatrième chapitre relatif à notre deuxième contribution porte sur les méthodes actuelles de structuration des données de journalisation, qui sont centrées sur les attributs des agents. Il montre leurs limites, et démontre les avantages de notre méthode basée sur les flux de clics des agents. Le cinquième chapitre relatif à notre troisième contribution illustre les limites des approches d'optimisation de l'usage du Web par dimension (*optimisation parasitique*). Il démontre le potentiel de notre approche prospectée pour l'optimisation de plusieurs dimensions à la fois (*optimisation symbiotique*). Notre approche d'optimisation est basée sur l'apprentissage semi-supervisé pour l'identification et le filtrage de motifs d'usage permettant l'optimisation de plusieurs dimensions à la fois. Enfin, le sixième chapitre conclut sur une vue d'ensemble de nos contributions, de leurs avantages, leurs limites, et les perspectives de recherches futures.

Abstract

Our thesis tackles the analysis of human-machine interaction based on usage traces. Our research deals with human-machine interaction through the case of Web Usage Interaction Mining. We provide contributions focused on the quality and relevance of three critical tasks, i.e., web usage data cleaning and structuring in addition to usage pattern discovery. Three new approaches related to these tasks are introduced, i.e., web usage data cleaning based on the log structure, clickstream centered structuring and usage pattern-based optimization through semi-supervised classification. The experimental results of the proposed methods, compared to those identified within the related literature, demonstrate significant improvements in terms of relevance balanced by workability and cost constraints.

First, our cleaning approach, based on the Log structure, and those identified within in the literature, based on the logging content, were tested on a panel of log files to demonstrate the relevance of our method in terms of identifying end-user clicks from their underlying user-agent hits (*noises*).

In addition, we proposed a structuring approach clickstream-centered, that deals with the limitation of the current agent-centered approaches. The comparative experimentation, we performed on several Log files, demonstrates the relevance of our method and its capacity to overcome the limits of the compared ones in terms of identification and construction of relevant single user sessions despite the constraints related to sequential logging, multiple hosting, dynamic web, and the lack of tracking information.

Finally, the preliminary outcomes of our symbiotic optimization approach, namely the optimization of several dimensions at the same time, i.e., traffic flow, Websites navigation paths and structures; tackles the issue of symbiotic optimization as a problem of classification. The target is to predict and control the optimization in order to avoid conflicts and/or contradictory optimization (*parasitic*). In this regard, we illustrate the limits of an optimization performed on distinct dimensions. Then, we introduced a semi-supervised classification approach for an optimization that handles several dimensions at a time (*symbiotic*). The experimentation on a logging sample of our university website demonstrates the capacity of our method to provide useful usage patterns for symbiotic optimization prediction, filtering and control.

Overall, the particular added value in terms of relevance balanced by workability and costs constraints, provided by each of our contributions, consists of relevance improvement of inter-referenced critical tasks, where the upstream relevance of a task references that downstream, leading to more relevant analysis, interpretation and reliable usage patterns.

Table des Matières

I.	Introduction Générale	1
I.1.	Contexte et problématique de la qualité de la fouille de l'usage Web.....	2
I.1.1.	Contexte	2
I.1.2.	Problématique	5
I.2.	Problèmes et contributions	8
I.2.1.	Définition du problème.....	8
I.2.2.	Contributions visées.....	9
I.3.	Présentation de la thèse et organisation du manuscrit.....	11
I.3.1.	Présentation de la thèse.....	11
I.3.2.	Organisation du manuscrit.....	12
II.	Etat de l'art de la fouille des données de l'Usage Web.....	17
II.1.	La fouille de l'Usage Web.....	18
II.1.1.	Discipline, finalités et applications	18
II.1.2.	Concepts, données, techniques et processus.....	20
II.2.	Système et données de journalisation Web.....	27
II.2.1.	Journalisation de l'activité Web.....	27
II.2.2.	Données de Journalisation Web	29
II.2.3.	Gestion, propriétés et caractéristiques.....	32
II.3.	Revue de littérature et analyse critique	36
II.3.1.	Classification de la littérature	36
II.3.2.	Revue de la littérature, limites et contributions visées	37
III.	Première Contribution – Nettoyage des Données de l'Usage Web	43
III.1.	Contexte, problème et objectif.....	44
III.2.	Analyse des méthodes de nettoyage – Travaux connexes.....	46
III.2.1.	Données et formalisme	46
III.2.2.	Perspectives de nettoyage.....	47
III.2.3.	Méthodes de nettoyage	47
III.3.	Contribution 1 : Nettoyage basé sur la structure de la journalisation.....	50
III.3.1.	Contribution 1.1 : Méthode heuristique	50
III.3.2.	Contribution 1.2 : Méthode de nettoyage basée sur le partitionnement génétique ...	58
III.4.	Limites et perspectives	71
IV.	Deuxième Contribution – Structuration des Données de l'Usage Web	73

IV.1.	<i>Contexte, problème et contributions visées</i>	74
IV.2.	<i>Analyse des méthodes de structuration – Travaux connexes</i>	75
IV.2.1.	<i>Concepts et problèmes de la structuration</i>	75
IV.2.2.	<i>Méthode réactives de structuration</i>	78
IV.2.3.	<i>Analyse des limites et proposition</i>	81
IV.3.	<i>Contribution 2 : Structuration centrée sur les flux de clics</i>	84
IV.3.1.	<i>Concept et approche</i>	84
IV.3.2.	<i>Méthode d'application</i>	85
IV.3.3.	<i>Méthode d'implémentation et d'évaluation</i>	87
IV.3.4.	<i>Expérimentation et évaluation</i>	93
IV.4.	<i>Limites et perspectives</i>	95
V.	Troisième Contribution–Optimisation Symbiotique de l'Usage du Web	97
V.1.	<i>Contexte, problème et contribution visée</i>	98
V.2.	<i>Illustration des méthodes et des techniques d'optimisation</i>	100
V.2.1.	<i>Contexte expérimental et organisation des données</i>	100
V.2.2.	<i>Optimisation de la structure</i>	104
V.2.3.	<i>Optimisation des chemins de navigation</i>	115
V.2.4.	<i>Optimisation du trafic</i>	119
V.3.	<i>Limites, problème et proposition</i>	122
V.3.1.	<i>Hypothèse des limites et définition du problème</i>	122
V.3.2.	<i>Illustration des limites et proposition</i>	123
V.3.3.	<i>Proposition</i>	128
V.4.	<i>Une approche d'apprentissage semi-supervisé pour l'Optimisation Symbiotique</i>	130
V.4.1.	<i>Données, indicateurs et instrument de mesure</i>	130
V.4.2.	<i>Méthode, modèle et évaluation</i>	132
V.4.3.	<i>Démonstration, discussion et perspectives</i>	134
VI.	Conclusion Générale	137
VII.	Bibliographie	141
VIII.	Liste des figures	155
IX.	Liste des tableaux	157
X.	Liste des algorithmes	159
XI.	Publications	161

Chapitre I

Introduction Générale

Analyse de l'Interaction Homme- Machine sur la base des traces d'utilisation

I.1. Contexte et problématique de la qualité de la fouille de l'usage Web

I.1.1. Contexte

Comme illustré en **Table I.1**, la relation homme-technologie est un domaine abordé sous différents angles, à savoir : comportemental/descriptif, évaluation/impact, et ingénierie/conception. L'informatique, les systèmes d'information et l'interaction homme-machine (*IHM*) sont les disciplines fondamentales qui associent ces différentes perspectives[1]–[19].

Table I-1 Problématique SI & IHM

Questionnement ?	Niveau Machine Technique	Niveau Individus Groupes (IHM)	Niveau Organisation Stratégique
Un bon Artefact ? (Valeur et modèles)	Qualité technique, coûts, et ergonomie Modèles causaux TCO	Utilisabilité, Utilité, satisfaction Acceptation, appropriation, activation Modèles causal, processus, interactive, sommative, formative	Diffusion de la technologie, adoption, Alignement stratégique SI/Entreprise Modèles statique (SAM) et dynamique (SAP)
	Symbiotique		
Mesurer son succès ? (Méthodes et Instruments)	Référentiel, standards, normes, paramètres Quantification ordinale, intervalles, rapport/ratio	Méthodes IHM orientées utilisateur, expert, analytique, UX Quantification nominale, ordinale, intervalles, rapport/ratio	Méthodes IS (gouvernance et schéma directeur SI, urbanisation SI, Alignement Stratégique, BPM, etc.) Evaluation et qualification de l'adéquation entre variables stratégiques internes et externes
	Techniques de fouille de données et d'extraction de connaissance		
Quelles données ? (Sources et collecte)	Données métriques quantitatives	Données quantitatives et qualitatives Sondage, enquête, observation	Données quantitatives et qualitatives Sondage, enquête, observation
	Traces d'utilisation, mouchards (cookies), journalisation, surveillance, flux de données de contexte spatio-temporel		

Dans les trois disciplines citées ci-dessus, le problème de l'analyse de l'IHM des artefacts tourne autour de ce qui fait leur réussite « succès de l'artefact ». Les questions les plus fréquemment posées sont les suivantes :

- Qu'est-ce qu'un bon artefact ?
- Comment mesurer son succès ?
- A travers quelles données ?

Un vaste consensus a été établi au cours de plusieurs décennies de recherches et de pratiques autour de ces trois questions au niveau Homme-Machine. Un bon artefact doit intégrer une symbiose entre qualité, utilité, utilisabilité. L'expérience utilisateur (*traces d'usage*) comme source de données offre une meilleure couverture et une compréhension approfondie. Les méthodes non intrusives dans la collecte de données objectives et l'analyse via les techniques de fouille de données et d'extraction de connaissances (*KDD*) confèrent une grande confiance en ce qui concerne la capture de comportement naturel, la découverte de modèles d'intérêt et la fourniture de résultats pertinents.

L'évaluation de l'interaction homme-machine, basée sur l'analyse des traces d'utilisation enregistrées par les systèmes de journalisation des Artefacts, à l'aide de techniques KDD reflète les tendances citées ci-dessus. La Fouille des Données de l'Usage du Web (*Web Usage Mining – WUM*) est l'une des cas d'application typique de cette tendance. Le WUM a l'ambition de palier aux limites des méthodes IHM traditionnelles appliquées à l'analyse de « l'Artefact Web », i.e., orientées utilisateur, expert, et/ou analytique.

L'analyse du Web dans ce cadre permet de tirer pleinement parti des tendances susmentionnées, à savoir la possibilité de couvrir tous les utilisateurs, tout le temps, et toutes les données d'une manière non intrusive et objective pour répondre à la question du « Comment » au titre des problématiques IHM du Web et face aux enjeux techniques et commerciaux du World Wide Web.

Le World Wide Web est la place d'échange la plus fréquentée du monde. Elle est la voie la plus accessible aux individus, groupes et/ou sociétés pour être, de manière interchangeable, fournisseur, client et/ou analyste. Cette accessibilité génère une concurrence commerciale rude, un vif intérêt à consommer et un volume énorme de données et d'informations complexes, hétérogènes, virales et volatiles à analyser. L'information de base générée par l'activité des acteurs de la Toile est appelée données d'utilisation/usage du Web/Journalisation serveurs. Les données d'utilisation du Web

sont enregistrées par les serveurs Web et fouillées afin d'extraire des connaissances à des fins différentes, afin de répondre aux besoins des artefacts, des entreprises, des utilisateurs, des concepteurs et des analystes du Web. Il s'agit de fouiller les données d'usage pour l'extraction de connaissances pour des applications diverses (*optimisation, recommandation, personnalisation, annonces*) au service des différents acteurs (*utilisateurs clients/prestataires, analystes, concepteurs*) et composants (*serveurs, sites, agents, terminaux*) du système tel qu'illustré par la **Figure I.1**.

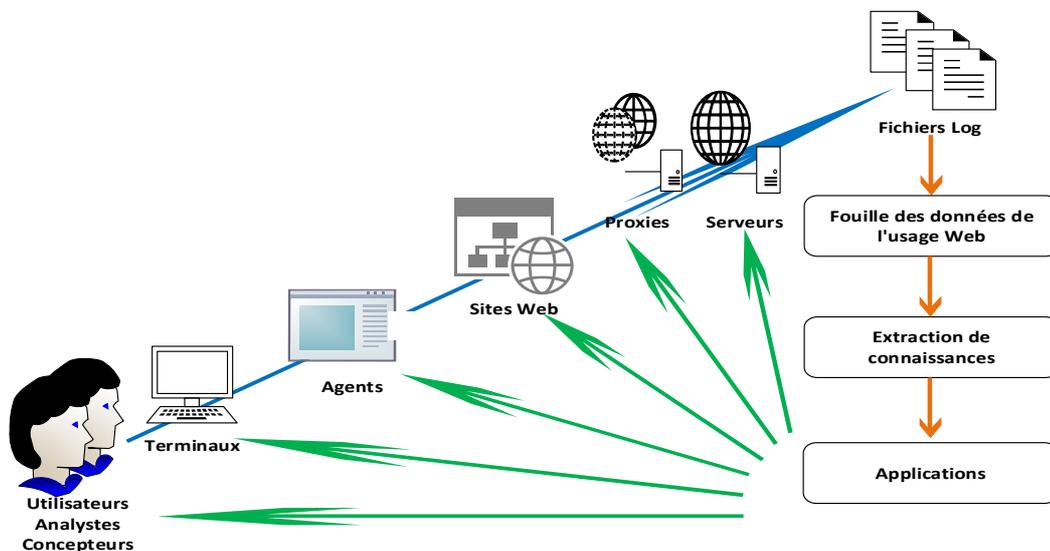


Figure I-1 Contexte de la Fouille des données de l'usage du Web

Aussi bien dans le monde universitaire que dans le monde des affaires, il est bien établi que la capacité à gérer ces données volumineuses et complexes dans un contexte aussi accessible améliore les avantages compétitifs, les positions commerciales et la durabilité des affaires. Il est décrit dans la littérature que : « Plus tôt vous apprendrez de vos données, plus vite vous pourriez laisser vos concurrents dans la poussière »[20].

La place Web est représentative des défis liés à l'utilisation du Web en termes d'apprentissage pour remplir des objectifs informatifs ou commerciaux. Les acteurs des points physiques de vente et de ceux numériques cherchent avant tout à attirer les visiteurs, à les transformer en clients, à les fidéliser et voir même gagner leur dévouement. Dans le cas d'un commerce de magasin physique, le client a tendance à être optimal et

considère les coûts de déplacement avant de passer à un autre point de vente. En revanche, en commerce électronique, rechercher la meilleure offre et basculer d'un site à un autre ne coûte pas plus qu'un clic sur les liens proposés par les services de comparaison et de recommandation.

Ainsi, apprendre des données d'utilisation pour optimiser les temps de réponse, adapter la structure du site et raccourcir les chemins de navigation sont des défis critiques pour les sites Web, notamment ceux à caractère commercial, car ils permettent d'attirer suffisamment de visiteurs pour qu'ils soient approchés et démarchés pour les convertir en clients et les fidéliser. Une structure adaptée qui met en évidence les raccourcis vers les objectifs des utilisateurs ciblés dans les meilleurs délais, conduirait probablement à garder les clients au sein du site Web. Un visiteur attiré atteignant ses objectifs par un chemin court, naviguant sur le site Web dans une structure adaptée, dans un laps de temps raisonnable ; est susceptible d'être conservé suffisamment sur le site Web pour être abordé et augmente la possibilité de le convertir en un client fidèle et dévoué.

I.1.2. Problématique

La problématique de l'optimisation de l'usage du Web peut être déclinée en trois problèmes, i.e., l'optimisation, les données et les techniques de fouille des données.

1) Problème de l'optimisation

L'optimisation de la navigation sur les sites Web porte sur trois dimensions distinctes, i.e., la navigation sur le site (*chemins et structures*) et le trafic sur le serveur et le réseau. L'optimisation de la navigation vise à fournir la meilleure expérience de navigation, où la structure des sites Web est adaptée au comportement de la navigation des utilisateurs. Cependant, l'optimisation du trafic permet de réduire la consommation de la bande passante et offrir les meilleurs temps de réponse.

L'optimisation de la navigation vise à raccourcir les chemins de navigation et adapter la structure du site en sorte que l'utilisateur atteigne ses objectifs dans les meilleurs temps en navigant des contenus adaptés à ses besoins. Cependant l'optimisation du trafic sur le serveur et le réseau vise à prédire la navigation et de mettre en cache les pages prédites sur des serveurs proxy intermédiaires. Ainsi, la bande passante du serveur de sites Web est économisée et le temps de réponse est amélioré.

À cet égard, du point de vue de l'interactivité symbiotique, un tel système d'optimisation par dimension peut s'avérer parasitique, à savoir l'optimisation d'une dimension au détriment d'une autre. A titre d'exemple, un tel système peut servir les utilisateurs en réduisant la longueur du chemin, mais ne conduit pas nécessairement à une optimisation du trafic, et peut produire l'effet inverse. A cet égard nous pouvons postuler que :

- Des chemins de différentes longueurs, passant par des pages de différentes tailles de contenus vers la même cible permettent d'étaler les flux de données dans le temps ;
- Des chemins courts et des structures de navigation rapprochées vont aiguiller et concentrer le trafic dans le temps, ce qui augmentera le flux ;
- L'optimisation du trafic est tributaire du temps de visite d'un contenu et sa taille, or le facteur Temps est hors de contrôle, ce qui rend le contrôle direct de l'optimisation impossible vu que le temps de visite des pages dépend de l'utilisateur.

Aussi, la mise en cache qui transfère une partie de l'activité du serveur Web aux serveurs intermédiaires hébergeurs du cache, peut tourner au désavantage des analystes du site car une grande partie de l'information sur la navigation n'est plus enregistrée sur leurs serveurs.

En résumé, l'optimisation par dimension peut déboucher sur des conflits et des contradictions d'optimisation et pose le problème de son contrôle vu que le facteur déterminant est le Temps, i.e., un facteur contrôlé en premier lieu par l'utilisateur.

2) Problème de données

Outre les avantages qu'offrent les données de journalisation en tant que traces d'utilisation et la promesse d'objectivité, de couverture de tous les utilisateurs, tout le temps ; cette source de données n'est pas exempte d'inconvénients. Le manque de structure, l'absence de standards obligatoires de développement Web et la complexité du système de journalisation des serveurs rendent l'analyse de ces données complexes et coûteuse.

En effet, le système de journalisation des serveurs Web enregistre les activités sur le serveur. Cette activité comprend celles des utilisateurs finaux (*Clics*) et des agents (*Hits*). Ainsi, les requêtes des utilisateurs finaux, des robots et des administrateurs sont

enregistrées, dans les fichiers de journalisation (*Fichiers Log*), entrelacées dans un ordre chronologique indépendamment de leurs sources ou types. Etant donnée que l'analyse de l'usage du Web est intéressée par l'activité des utilisateurs finaux ; distinguer les requêtes de l'activité des utilisateurs finaux de celles des agents n'est pas une tâche triviale.

Ainsi, les fichiers de log manquent de structure au sens de de la statistique (les requêtes des utilisateurs finaux et agents sont enregistrées entrelacées dans un ordre chronologique indépendamment de leurs sources ou types). Les données enregistrées représentent, donc, des informations brutes non structurées.

Aussi, à l'exception des recommandations de la communauté W3C (*The World Wide Web Consortium*), il n'y a pas de standards internationaux obligatoires en matière de conception et développement Web. Les requêtes par clics (*utilisateurs finaux*) supposées être reconnaissables via leur objet portant sur des ressources de type (*extension*) HTML ou équivalent ; peuvent porter sur des ressources d'autre type/extension, e.g., documents, graphique, vidéo, son, etc. Il s'agit là de cas envisageables (*recommandations de la communauté W3C non obligatoires*) au-delà des cas particuliers des sites de galeries d'image ou équivalent, e.g., Wikipedia, blogs, etc. Enfin, les technologies de conception et de développement de sites Web Dynamiques/Adaptatifs génèrent une grande variété de requêtes d'utilisateurs finaux/agents qui sont confuses.

3) Problème de techniques de fouille de données

L'application des techniques de fouille de données aux données de journalisation Web/Log vise différents objectifs, i.e., l'optimisation du système, la personnalisation et la recommandation, le placement publicitaire [21]–[23]. En fonction de l'objectif et du contexte de l'analyse, les données Log sont complétées par d'autres données et sont fouillées à l'aide de différentes techniques de fouille, e.g., statistique, classification et partitionnement, règles d'association et séquentielles, modélisation des processus [24]–[26].

L'application des techniques de fouille de données de journalisation pour l'optimisation de l'usage Web est confrontée au problème lié au données de journalisation et à la fiabilité de l'optimisation par dimension, i.e., difficulté à identifier les requêtes agents à nettoyer (*bruit*), à grouper les requêtes par utilisateur individuel (*structuration*),

optimisation pluridimensionnelle de l'usage Web (*optimisation symbiotique*).

Dans ce contexte, la promesse, de d'objectivité, de couverture de tous les utilisateurs, tout le temps, de l'application des techniques de fouille de données à l'analyse de l'IHM sur la base des données de journalisation (*traces d'utilisation*) ; est confrontée à des défis de qualité qui risquent de compromettre cette promesse. Il s'agit là de :

- La qualité du nettoyage des données de journalisation et leur structuration qui conditionnent le reste du processus de fouille et la qualité de découverte des motifs de l'usage du Web ;
- La qualité des motifs d'usage découvert à servir pour une optimisation symbiotique de l'usage du Web.

Ainsi nos contributions aborderont les problèmes suivants :

- Qualité du nettoyage des données de journalisation ;
- Qualité de la structuration des données de journalisation Web ;
- Qualité des motifs d'usage découverts au titre de l'optimisation Web.

I.2. Problèmes et contributions

I.2.1. Définition du problème

1) Fouille des données de journalisation

La fouille de l'usage du Web est l'application des techniques de fouille de données aux données de journalisation des serveurs Web pour l'extraction de connaissances servant les besoins des applications Web, e.g., optimisation, recommandation, personnalisation, placement publicitaire sur le Web. Les données d'utilisation Web, appelées données de journalisation Web ou données Log, sont les données, enregistrées par les serveurs et les caches des navigateurs, décrivant les requêtes des utilisateurs, e.g., l'adresse IP, heure de la requête, page Web demandée [27].

Les techniques d'analyse les plus référencées relèvent de la statistique descriptive, les règles d'association et séquentielles, la classification, le partitionnement et la modélisation des processus. Ces techniques sont utilisées séparément ou combinées pour découvrir des connaissances utiles sur les utilisateurs ainsi que sur le système.

Le processus de fouille des données Log comprend trois étapes principales, i.e., le prétraitement, la découverte de motifs et l'analyse de motifs découverts. L'étape de

prétraitement vise à nettoyer, structurer et transformer les données en fonction des techniques de fouilles envisagées. La découverte de motifs consiste en l'application de techniques de fouille de données pour identifier des motifs utiles en fonction des objectifs visés. L'analyse des motifs vise à filtrer le modèle découvert pour garder ceux utiles pour le problème adressé [21], [22].

2) Caractéristiques des données et problème

La journalisation Web enregistre les requêtes des utilisateurs finaux et d'autres agents dans des fichiers Log dans un ordre chronologique indépendamment de leurs sources, type, ou objet. Les ressources Web objets de ces requêtes sont les URI (*html*) des pages Web demandées par les utilisateurs ainsi que leurs composantes d'affichage (*média*) chargées par les agents [27]. A l'exception des recommandations de la communauté W3C (*The World Wide Web Consortium*), il n'y a aucun standard obligatoire de format de ressources Web en matière de pages web et leurs composantes. Ainsi, des ressources média peuvent être demandées directement par clics des utilisateurs finaux en dehors du format html. Avec le Web adaptatif et dynamique doté de système de recommandation et de personnalisation, les ressources sont indexées sans extension de format et change en permanence [28].

Dans ce contexte, la fouille des données de l'usage Web est intéressée par le comportement des utilisateurs finaux. Donc, les requêtes formulées par les agents sont considérées comme du bruit à nettoyer avant la fouille des données. A ce titre, le fait que les ressources Web peuvent être demandées interchangeablement par les utilisateurs finaux et les agents, il n'est pas évident de distinguer les requêtes des utilisateurs finaux de celles des agents. Aussi, dans le cas d'utilisateurs réticents à l'égard de l'authentification et l'acceptation des cookies, la journalisation séquentielle constitue une difficulté majeure à la structuration des requêtes par utilisateurs et sessions.

La pertinence du nettoyage et de la structuration des données de journalisation conditionne la fiabilité des motifs d'usage découverts au titre des objectifs de la fouille des données de l'usage Web, tels que l'optimisation.

I.2.2. Contributions visées

1) Nettoyage des données

Les méthodes actuelles, de nettoyage des données Log, sont le nettoyage

conventionnel et avancé [21], [28]. Les deux méthodes reposent sur une heuristique qui filtre les ressources sur la base de connaissance apriori sur les ressources destinées aux utilisateurs finaux et agents.

Ces deux méthodes nécessitent une connaissance apriori sur le contenu des sites web des données Log à nettoyer. Ainsi, elles ne sont pas généralisables et sont consommatrices en ressources pour le maintien à jours d'une base de connaissance apriori exhaustive. Aussi, la qualité des résultats de ces méthodes demeure fortement perturbée par les contraintes du Web Dynamic et Adaptatif.

A cet égard, deux méthodes centrées sur la structure de la journalisation ont été proposées pour surmonter les difficultés des deux méthodes centrées sur le contenu. Les résultats obtenus démontrent l'avantage des méthodes proposées en matière de pertinence, de contraintes et de couts d'application.

2) Structuration des données

La structuration des données Log, dans le cas d'absences d'authentification ou acceptation de cookies, qualifiée d'approche réactive ; vise à identifier les requêtes par utilisateurs singuliers et leurs sessions successives de visites. Cette structuration est nécessaire pour l'identification des transactions.

Trois méthodes réactives de structuration sont souvent référencées, i.e., orientées sur le temps, la topologie du site, le graph du site. Les sessions générées par ces méthodes ne sont jamais identiques aux sessions réelles et nécessitent toujours des améliorations [29].

La méthode proposée pour l'amélioration de la qualité des sessions à générer repose sur l'enrichissement de leurs heuristiques par des contraintes de pertinences. La méthode proposée a été implémentée via une fonction-objectif qui maximise le nombre de sessions pertinentes. La qualité des sessions obtenues est significativement supérieur aux méthodes réactives génériques.

3) Contrôle de l'optimisation de l'usage Web

L'optimisation de l'utilisation du Web basée sur les données Log aborde séparément trois dimensions, i.e., l'optimisation du trafic, de la structure, et des chemins traversés. Les facteurs déterminants dans l'optimisation de ces dimensions sont le temps

d'accès aux ressources, leur taille, leur nombre ou la longueur du chemin parcouru [13], [30], [31]. Face à l'absence de tendance de corrélation liant ces trois facteurs à la fois ; le contrôle du produit symbiotique (*équilibré*) ou parasitique (*conflictuel et/ou contradictoire*) de l'ensemble des optimisations distinctes n'est pas évident dès lors que la facteur « Temps » est hors contrôle.

La contribution prospecte une méthode basée sur l'apprentissage semi-supervisé devant permettre de prédire et contrôler/filtrer les optimisations entreprises par un système d'adaptation automatique d'un site Web dynamique, et ceci en fonction de leur produit symbiotique/parasitique prédit.

I.3. Présentation de la thèse et organisation du manuscrit

I.3.1. Présentation de la thèse

Cette thèse approche la problématique de l'évaluation de l'interaction Homme-Machine à travers le cas de l'interaction avec le Web. L'analyse de l'interaction avec le Web relève du champ interdisciplinaire de la fouille des données de l'utilisation du Web, en l'occurrence, les données de journalisation Web, i.e., serveurs et agents, enrichies par des données des profils des utilisateurs et de contextes spatiotemporels.

Nos travaux portent sur la qualité et la pertinence de trois tâches critiques de la discipline en question, i.e., le nettoyage des données d'utilisation, leur structuration, et la découverte de motifs d'usage. A ce titre, nous proposons trois nouvelles approches liées à ces tâches dont les résultats expérimentaux, comparés à ceux des méthodes actuelles, démontrent des améliorations significatives en termes de pertinence rapportée aux contraintes d'applicabilité et de coûts.

La première contribution relative à l'approche de nettoyage est centrée sur la structure des données de journalisation au lieu de celles actuelles centrées sur le contenu de la journalisation. Cette approche est déclinée en deux variantes, i.e., une méthode heuristique et une autre basée, par analogie de concepts, sur les techniques de partitionnement génétique.

La deuxième contribution, qui porte sur la structuration des données, est une approche centrée sur le flux de clic des agents au lieu de celles existantes centrées sur les attributs des agents.

Enfin, notre troisième contribution présente nos résultats préliminaires sur notre prospection d'une approche de classification semi-supervisée qui permet de découvrir des motifs d'usage devant permettre d'optimiser, à la fois, plusieurs dimensions de l'usage du Web, en filtrant les optimisations par dimension qui peuvent s'avérer conflictuelles et/ou contradictoires (*parasitiques*).

Le manuscrit présentant le contexte et les contributions de notre thèse est composé de sept (07) chapitres.

I.3.2. Organisation du manuscrit

Le premier chapitre introductif aborde le contexte et la problématique de l'analyse de l'interaction Homme-Machine sur la base des traces d'utilisation, et ceci à travers le cas de l'interaction avec le Web relevant de la discipline de la fouille des données de l'utilisation du Web.

Le deuxième chapitre présente un état de l'art générique de la fouille de l'usage du Web, décrit le système de journalisation Web et pointe les problématiques dont celles objets de nos trois contributions. A cet égard, nous signalons que notre recherche porte sur trois contributions différentes mais qui portent sur trois tâches liées au processus de fouille de données de l'usage Web.

La fouille de l'usage du Web est l'application des techniques de fouille de données aux données de journalisation des serveurs Web pour l'extraction de connaissance pour servir les besoins des applications Web, e.g., optimisation, recommandation, personnalisation, placement publicitaire sur le Web. Les données d'utilisation Web, appelées données de journalisation Web ou données Log, sont les données, enregistrées par les serveurs et les caches des navigateurs, décrivant les requêtes des utilisateurs, e.g., l'adresse IP, heure de la requête, page Web consultée.

Les techniques d'analyse les plus référencées relèvent de la statistique descriptive, les règles d'association et séquentielles, la classification, le partitionnement et la modélisation des processus. Ces techniques sont utilisées séparément ou combinées pour découvrir des connaissances utiles sur les utilisateurs et le système. Le processus de fouille des données Log comprend trois étapes principales, i.e., le prétraitement, la découverte de motifs et l'analyse de motifs découverts. L'étape de prétraitement vise à nettoyer, structurer et transformer les données en fonction des techniques de fouille

envisagées. La découverte de motifs consiste en l'application des techniques de fouille de données pour identifier des motifs utiles en fonction des objectifs et des applications visés de l'analyse, e.g., optimisation, recommandation, personnalisation, placement publicitaire. L'analyse des motifs vise à filtrer le modèle découvert pour garder ceux utiles pour le problème adressé.

Aussi, ce chapitre détaille le système et les données de journalisation Web et ses caractéristiques qui représentent des facteurs de difficulté en matière de fouille et l'analyse des données de l'utilisation du Web.

La journalisation Web enregistre les requêtes des utilisateurs finaux et celles d'autres agents dans des fichiers Log dans un ordre chronologique indépendamment de leurs sources, type, ou objet, à savoir une journalisation séquentielle. Les ressources Web objet de ces requêtes sont identifiables via les URI (*Uniform Resource Identifier*) des pages Web demandées par les utilisateurs ainsi que leurs composantes (*page design and content*) d'affichage (*content and accessorial média*) chargées à l'écran par les agents.

En effet, en matière de design et contenu des pages Web, à l'exception des recommandations de la W3C, il n'y a aucun standard obligatoire de format de ressources Web en matière de pages web et leurs composantes, e.g., des extensions html ou équivalent pour les URI des pages Web et d'autres formats pour leurs composantes. Ainsi, des ressources média peuvent être demandées directement par clics des utilisateurs finaux en dehors du format html. Aussi, le Web adaptative et dynamique doté de système de recommandation et de personnalisation, repose sur des ressources indexées sans extension de format et change en permanence.

Ces caractéristiques représentent des facteurs de difficulté en matière de la fouille des données de journalisation en termes de qualité de nettoyage, de structuration et découverte de motifs d'usage. En effet, la fouille des données de l'usage Web est intéressée par le comportement des utilisateurs finaux. Ainsi, les requêtes formulées par les agents sont considérées comme du bruit à nettoyer avant la fouille des données.

Dans ce contexte, du fait que les ressources Web peuvent être demandées interchangeablement par les utilisateurs finaux et les agents, il n'est pas évident de distinguer les requêtes des utilisateurs finaux de celles des agents pour nettoyer celles des agents. Aussi, dans le cas d'utilisateurs réticents à l'égard de l'authentification et

l'acceptations des cookies, la journalisation séquentielle constitue une difficulté majeure à la structuration des requêtes par utilisateurs et sessions. Les difficultés en matière de nettoyage et structuration et leur qualité se répercute sur la pertinence et la fiabilité des motifs d'usage du Web.

Au titre de notre état de l'art, et pour des raisons de lisibilité et de clarté, nous avons présenté, en premier lieu, un état d'art générique traitant du processus de la fouille des données de l'usage Web et ses tâches. Par contre, la revue de littérature détaillée des travaux connexes à nos contributions sont présentées dans les chapitres y afférents. Ainsi, en amont de chaque revue de littérature détaillée dans les chapitres des contributions ; certains concepts génériques de l'état de l'art sont repris pour contextualiser la contribution, définir le problème et l'objectif visé.

Le troisième chapitre aborde les méthodes actuelles de nettoyage des données de journalisation, leurs limites, et la proposition d'une approche basée sur la structure de la journalisation au lieu de son contenu. Les méthodes actuelles, de nettoyage des données Log, sont le nettoyage conventionnel et celui avancé. Les deux méthodes reposent sur une heuristique qui filtre les ressources sur la base de connaissance a priori sur les ressources destinées aux utilisateurs finaux et les agents. Ces deux méthodes nécessitent une connaissance a priori sur le contenu des sites web des données Log à nettoyer. Ainsi, elles ne sont pas généralisables et sont consommatrices en ressources pour le maintien à jours d'une base de connaissance a priori exhaustive. Aussi, la qualité des résultats de ces méthodes demeure fortement perturbée par les contraintes du Web Dynamic (*création et annulation automatique de ressources*) et Adaptatif (*ressources sans extension indiquant leur type*).

A cet égard, nous avons proposé et tester deux méthodes centrées sur la structure de la journalisation pour surmonter les difficultés des deux méthodes centrées sur le contenu. Les résultats obtenus démontrent l'avantage des méthodes proposées en matière de pertinence, de contraintes et de couts d'application.

Le quatrième chapitre traite des méthodes actuelles de structuration des données de journalisation, leurs limites, et la proposition d'une méthode basée sur le flux de clics des agents au lieu des attributs des agents.

La structuration des données Log, dans le cas d'absences d'authentification ou

acceptation de cookies, qualifiée d'approche réactive, vise à identifier les requêtes par utilisateurs singuliers et leurs sessions successives de visites. Cette structuration est nécessaire pour l'identification des transactions. Trois méthodes réactives de structuration sont souvent référencées, i.e., orientée sur le temps, la topologie du site, le graph du site. Les sessions générées par ces méthodes ne sont jamais identiques aux sessions réelles et nécessitent toujours des améliorations.

La méthode proposée pour l'amélioration de la qualité des sessions produites repose sur l'enrichissement de leurs heuristiques par des contraintes de pertinences. Cette méthode a été implémentée via une fonction-objectif qui maximise le nombre de sessions pertinentes. La qualité des sessions obtenues est significativement supérieure aux méthodes réactives génériques.

Le cinquième chapitre illustre les limites des approches d'optimisation de l'usage du Web par dimension et propose une approche d'optimisation intégrée sur la base de la découverte de motifs d'usage couvrant plusieurs dimensions à la fois.

L'optimisation de l'utilisation du Web basée sur l'analyse des données de journalisation traite séparément trois dimensions, i.e., l'optimisation du trafic, de la structure, et des chemins traversés. Il s'agit d'optimisations reflétant l'intérêt des acteurs concernés par l'usage Web, i.e., optimisation menée par l'analyste/propriétaire au profit de l'utilisateur, l'artefact (*le site Web*) espérant un retour à son profit (*analyste/propriétaire*).

L'optimisation de ces dimensions repose sur trois facteurs, i.e., le temps d'accès/consultation aux/des ressources, leur volume, et leur nombre ou la longueur du chemin parcouru. Ces indicateurs peuvent donner, respectivement, un aperçu descriptif explicatif de l'optimisation de chaque dimension et la valeur du produit de l'optimisation globale des trois dimensions. Cependant, vu l'absence de tendance de corrélation liant ces facteurs à la fois, et comme le facteur Temps dépend de l'utilisateur ; le contrôle direct du produit global de l'ensemble des optimisations distinctes par dimension n'est pas évident.

La contribution ciblée vise à fournir une méthode d'apprentissage semi-supervisée pour contrôler, filtrer et préserver l'équilibre de la valeur symbiotique/parasitique de l'interaction de l'usage Web servant les intérêts des différents acteurs concernés. A cet

égard, nous tenons de préciser que ce chapitre porte les résultats préliminaires de notre recherche en la matière et présente nos perspectives à ce sujet.

Enfin, **le sixième chapitre** conclut sur une vue d'ensemble sur nos contributions, de leurs avantages, leurs limites, et les perspectives des recherches futures.

Chapitre II

Etat de l'Art

La Fouille des Données de l'Usage Web

II.1. La fouille de l'Usage Web

II.1.1. Discipline, finalités et applications

1) Discipline

- *La Fouille du Web*

Deux définitions ont été évoquées lors de la WEBKDD'99, i.e., centrée sur le processus [32] et une autre sur les données [33] qui est la plus citée. La définition centrée sur le processus indique que la fouille du Web consiste à découvrir et à analyser des informations utiles provenant du World Wide Web. À côté de cette définition centrée sur le processus, l'autre définition centrée sur les données stipule que la fouille du Web est l'application de techniques d'exploration de données pour extraire des connaissances à partir de données Web. La **Figure II.1** présente la taxonomie de la fouille du Web la plus référencée. Elle comprend trois sous-catégories, i.e., fouille des données de contenu Web, de structure Web et de l'utilisation du Web [20].

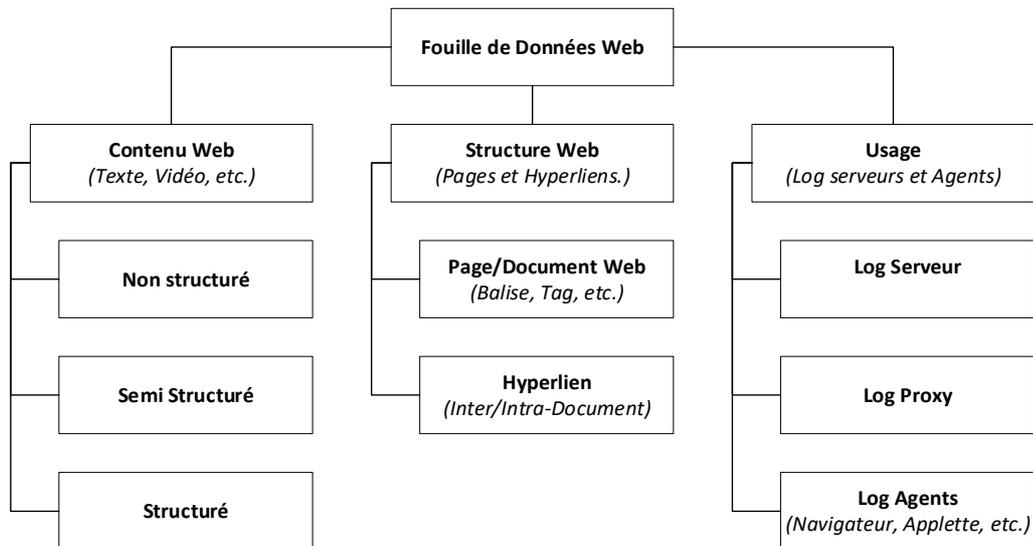


Figure II-1 Taxonomie sommaire de la fouille des données Web

La fouille du contenu Web consiste à extraire des informations utiles à partir des documents Web (*multimédia et texte*). La fouille de la structure Web est axée sur l'exploration des graphs Web qui représentent les pages Web en nœuds et les hyperliens en arêtes. La structure est fouillée pour analyser les topologies pour divers motifs, e.g.,

optimisation de structure, visibilité sur la toile, classification des pages Web, etc. Généralement, l'analyse de la structure Web porte sur les hyperliens (*URI, URN, URL*) qui structure le site (*analyse inter pages*) et/ou les balises (*HTML, XML*) de page qui l'organisent en arborescence (*analyse intra pages*). Enfin, la fouille de l'utilisation du Web, qui fait l'objet de notre recherche et contributions, se concentre sur les données d'utilisation, appelées données log ou journalisation serveur, enregistrées par les serveurs Web pour capturer le comportement des utilisateurs et les interactions avec les sites Web [21], [22], [27], [32]–[35].

- ***Fouille de l'usage du Web***

Le terme « Fouille des Données Log/Journalisation Serveur » est souvent utilisé de manière interchangeable avec celui de « Fouille des Données de l'Usage Web ». La fouille de l'utilisation du Web est l'application de techniques de fouille de données visant à découvrir des modèles d'utilisation intéressants à partir des données d'utilisation du Web, afin de comprendre et de mieux répondre aux besoins des applications Web[21], [33]. De plus, le fait que la fouille des données de l'utilisation Web soit effectuée sur les données Log, il ressort clairement de l'analyse de la littérature, que la fouille de l'utilisation du Web peut être effectuée sur des données de structure, de même que les données Log peuvent servir à la fouille de structure [20]–[22], [27], [33]. À cet égard, il convient de noter que l'enrichissement des données de journalisation, par d'autres données, e.g., les données de contexte (*profils, temps, espace*), illustre le fait que Web Usage Mining peut être exécuté sur d'autres données en plus des données de journalisation. À ce titre, notre recherche se limite à la qualité d'analyse des données Log sans enrichissement au vu qu'il représente à la fois une source de pertinence et d'impertinence.

2) Finalités et applications

- ***Finalités***

Les données de journalisation serveur peuvent offrir un aperçu précieux de l'utilisation du site Web. Elles reflètent l'utilisation réelle dans des conditions de travail naturelles non intrusives, comparée aux données collectées dans d'autres cadres d'observation et de collecte expérimental. Les données Log présentent l'avantage de couvrir un grand nombre d'utilisateurs sur une longue période à même tous les utilisateurs tous le temps[27], [30], [35], [36].

La fouille des données Log a des objectifs et des applications variés. Ces objectifs couvrent trois dimensions, i.e., l'artefact, l'utilisation, et le service. Les objectifs centrés sur les artefacts visent l'amélioration de la qualité intrinsèque technique site Web. Les objectifs centrés sur l'utilisation concernent le facteur humain, la facilité d'utilisation et l'ergonomie. Les objectifs centrés sur le service portent sur l'utilité de l'artefact. A noter qu'il s'agit-là des dimensions canoniques de l'analyse IHM, i.e., la qualité techniques, l'utilisabilité, et l'utilité. Ces objectifs servent les intérêts des parties prenantes d'un artefact, i.e., l'analyste/concepteur (*qualité*), l'analyste/métiers (*utilité*), l'utilisateur (*utilisabilité*), et l'artefact lui-même (*qualité*).

- ***Applications***

Les objectifs ci-dessus peuvent être illustrés à travers plusieurs applications, notamment [22], [27] :

- Optimisation de l'usage par l'adaptation de la structure du site à la structure de navigation des utilisateurs, de la navigation par le raccourcissement des chemins de navigation vers le contenu recherché, du trafic par la mise en cache basée sur la prédiction de la navigation.
- Des applications à des fins commerciales, e.g., la recommandation, la personnalisation et le placement publicitaire.
- Analyse des comportements des utilisateurs pour la découverte de modèles d'usage pour diverses fins, e.g., modélisation et conception de nouveaux services, détection des intrusions et contrôler les activités virales des robots, etc.

II.1.2. Concepts, données, techniques et processus

1) Concepts

Au vu de la nature interdisciplinaire de la recherche en fouille des données Web, il n'y a pas de cadre théorique et conceptuel unifié en la matière. Néanmoins certaines contributions ayant couvert un large spectre en termes d'objet, méthodes, et finalités de la discipline permettent de lever certaines confusions et/ou utilisations interchangeables de concepts liées à l'activité Web et à la fouille des données y afférentes [21], [26], [29]–[34], [37].

L'**activité** Web fait référence aux **utilisateurs-finaux**, **agents-navigateurs** et

autres **agents-robots** qui requêtent des **ressources** de sites **Web** contenues dans des **pages Web** [27], [34]. Les pages Web sont identifiables sur le Web par l'intermédiaire de leurs **URI** (*Uniform Resource Identifier*), i.e., leur **URN** (*Uniform Resource Name*) et leur **URL** (*Uniform Resource Locator*). Ainsi, les pages Web constituent un ensemble de ressources qui comprennent principalement des :

- URI identifiant la page Web sur la toile.
- Contenu multimédia destiné aux utilisateurs-finaux.
- Médias accessoires liés à la conception de la page.
- Services destinés aux utilisateurs finaux.
- Services destinés aux agents.
- Styles, frames et scripts qui contrôlent le contenu et l'affichage de la page sur les terminaux.

Un utilisateur-final est une personne physique qui navigue sur le Web via un agent-utilisateur. Elle accède aux ressources du Web en envoyant des demandes/requêtes, via l'agent-navigateur, au serveur Web. Une demande d'un utilisateur-final est appelée un clic. Les **flux de clics** sont des demandes successives émanant d'un seul utilisateur-finale unique ou singulier. Le termes **utilisateur unique** et **singulier** sont utilisés pour designer respectivement :

- Un flux de clics émanant d'un personne unique reconnue par authentification ou éventuellement cookies dans le cas supposé d'un terminal personnel.
- Un flux de clics regroupés par sources distinctes sans pouvoir assigner les sources distinctes à des utilisateurs unique. Il s'agit là du cas de l'absence d'information d'authentification ou cookies.

Un **aperçu** de page est le rendu de l'ensemble des ressources/fichiers composant le contenu de la page Web. L'**utilisateur-final** demande, par l'intermédiaire d'agents-navigateurs, les pages par **clic** sur les hyperliens (*URI*) ou en les tapants. Les fichiers/ressources nécessaires pour afficher l'aperçu des pages sont demandées par les agents-navigateurs. Les **requêtes** des **agents-navigateurs** et ceux des robots d'indexation ou autres robots ou scriptes sont des **hits**. Une **session utilisateur** est la séquence de pages consultées par un utilisateur sur l'ensemble du Web (plusieurs sites Web) à chaque accès au Web. La séquence de pages consultées par site Web particulier, à chaque accès au site,

est appelée une **session serveur/visite**. Par transposition les sessions d'utilisateurs uniques/singuliers sont appelées, respectivement, **sessions uniques/singulières**. Tout sous-ensemble/séquence sémantiquement significatif/utiles d'une session utilisateur ou serveur est appelé **épisode/transaction**.

2) Techniques de fouille

Statistique Descriptive : Analyse de données de journalisation brutes ou structurées. Elle permet d'avoir un aperçu global sur l'utilisation du site, e.g., les ressources les plus consultées, temps de consultation, longueurs des chemins de navigation, les erreurs fréquentes. Il s'agit de rapports statistiques sur l'utilisation du site et la performance du système. C'est un support de monitoring et d'aide à la décision qui sert à des fins d'amélioration technique des systèmes et contenus des sites Web.

Classification : Analyse de données de journalisation structurées. Elle permet d'identifier les caractéristiques de classes aux attributs prédéfini à des fins de profilage de groupes et d'individus, e.g., les clients potentiels, les visiteurs par curiosité, les fraudeurs. Les techniques les plus citées en la matière sont les Arbres de Décision, les Réseaux de Neurones, les classifieurs Bayésiens Naïfs.

Partitionnement : Analyse de données de journalisation souvent structurées et enrichies, e.g., profils utilisateurs, données spatiotemporelles de contexte. Elle permet d'identifier des communautés d'utilisateurs, de contenus, et leurs relations, e.g., les attributs communs des utilisateurs qui consultes des contenus aux caractéristiques similaires. Il s'agit d'analyse pour des fins de personnalisation, recommandation. Les techniques les plus citées en la matière sont : le Partitionnement Hiérarchique, en K-Moyennes, Flou, basé sur la Densité, basé sur un Modèle, Hybride.

Règles d'Association : Analyse de données de journalisation structurées. Elle permet d'identifier les associations fortes entre ressources souvent consultées lors de la même visite/transaction même si elles ne sont pas liées. Il s'agit d'analyse pour des fins marketing, e.g., analyse panier ; et technique, e.g., adaptation de la structure du site. A cet égard, une variété d'algorithmes sont proposés dans la littérature, e.g., Apriori, FP-Growth, Eclat, SaM, SETM, AIS, Recursive.

Règles séquentielles : Analyse des données de journalisation structurées. Elle permet d'identifier les séquences de navigation fréquentes. Il s'agit du même principe des

règles d'association incluant la contrainte de l'ordre. C'est une technique utile pour l'analyse des chemins traversés pour l'optimisation de la navigation. Une grande variété d'algorithmes sont proposés dans la littérature, i.e., méthodes basées sur le principe de recherche en largeur d'abord tel que GSP, SPAD ; méthodes basées sur le principe de recherche en profondeur d'abord tel que PSP, PREFIXSPAN, SPAM ; méthode de séquences fréquentes fermées tel que CLOSSPAN, BID ; méthodes de séquences fréquentes incrémentales tel que ISE, ISM, IUS, FASTUP, KISP ; méthodes d'extraction de séquences fréquentes sous contraintes de temps, item et/ou longueur.

Les règles séquentielles comme les règles d'association, communément, appelée règles de motifs fréquents (règles d'items fréquents), constituent aussi une méthode pour extraire des règles de prédiction de la navigation dans l'absence d'un modèle prédictif. Les différentes variantes d'algorithmes reflètent l'évolution vers des méthodes optimisées en termes de coûts, de pertinence, de portée de la connaissance fournie et son actualisation.

Modélisation : Analyse des données de journalisation structurées. Elle permet d'identifier des modèles de navigation pour des fins de prédiction, détection des intrusions, profilage des comportements de navigation, et ceci pour diverses applications, e.g., commerciales, techniques, sécuritaire. Les techniques les plus citées en la matière sont : les Chaînes de Markov, les réseaux bayésiens, les Séries Temporelles.

Les techniques ci-dessus peuvent être utilisées séparément ou combinées dans le cadre du processus de fouille, et ceci en fonction de la finalité et de l'application.

3) Processus

Ce qui a fait émerger la fouille des données du Web dont la fouille de l'utilisation du Web en une discipline c'est, d'une part, ses concepts spécifiques cités ci-dessus, et d'autre part, la particularité de son processus, ses tâches et les concepts y afférents, comparés au processus générique de fouille de données [20]–[22], [27], [32]–[35], [38]–[44].

La **Figure II.2** présentant le processus de fouille des données de l'usage du Web et illustre les étapes et tâches génériques d'un processus de fouille de données en plus de celles spécifiques à la fouille des données de l'usage Web.

Taches communes : Le processus de fouille de données de l'usage Web, d'un point de vue générique, repose sur les étapes canoniques d'un processus de fouille de données, à savoir :

- La préparation des données, i.e., collecte, manipulation et extraction d'attributs utiles, nettoyage, structuration, transformation, enrichissement des données (*profils et contexte*), etc. ;

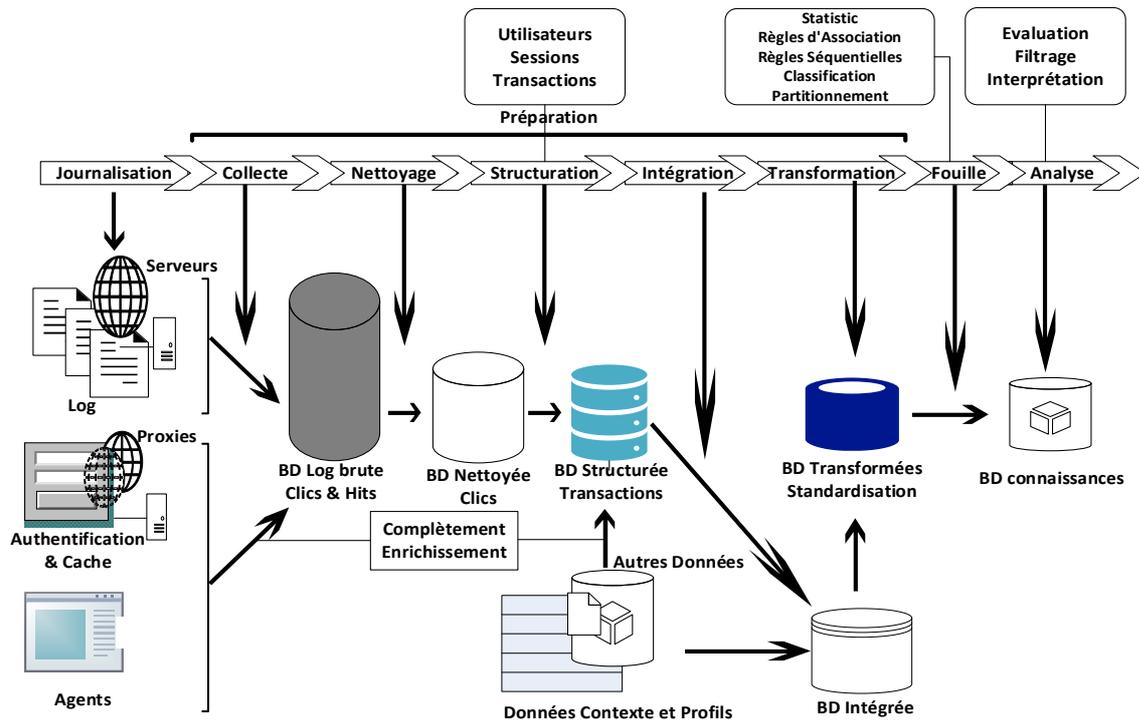


Figure II-2 Processus de fouille des données de l'usage Web

- La préparation des données, i.e., collecte, manipulation et extraction d'attributs utiles, nettoyage, structuration, transformation, enrichissement des données (*profils et contexte*), etc. ;
- La découverte de motifs d'usage Web qui consiste en l'application des techniques de fouille suscitées aux données Log préparées pour la découverte de connaissance, i.e., motifs, règles et/ou modèles d'usage Web ;
- L'analyse, des connaissances découvertes, qui a trait à l'abstraction, le stockage de cette connaissance et son filtrage pour retenir que les motifs, règles et/ou modèles

d'intérêt. A ce titre il souvent question des techniques d'abstraction et de stockage (*table, cube, etc.*), d'interrogation (*SQL, MDX, etc.*), et seuilles d'intérêt (*subjective, objective*).

Tâches spécifiques : Les tâches spécifiques au processus de fouille des données de l'usage Web qui reposent sur des concepts qui lui sont propre relève de l'étape de prétraitement, i.e., nettoyage en l'occurrence l'identification des clics parmi les hits afin de nettoyer ces derniers ; la structuration en utilisateurs uniques/singuliers, sessions uniques/singulières ; complètement des chemins parcourus, et identification des transactions.

Identification des clics/hits (Prétraitement – Nettoyage des données) : La tâche de nettoyage peut être décomposée en trois couches.

- La première est destinée à identifier les clics (*utilisateur final*) hors des hits (*agents*) sous-jacents, vu que la fouille des données de l'usage Web s'intéresse en premier lieu aux comportements des utilisateurs finaux.
- La seconde consiste à sélectionner, parmi les clics identifiés, ceux qui sont significatifs pour les objectifs et le contexte de l'analyse, par exemple les analyse des requêtes ayant abouti ou non, exclusion de celles des proxy et robots à IP uniques connus.
- La troisième est effectuée en aval de la structuration des données et consiste plutôt en un nettoyage intelligent basé sur des modèles de navigation destinés à la détection de robots et de valeurs aberrantes.

Ainsi, la couche de nettoyage centrale qui s'attaque au bruit générique est l'identification des clics provenant de hits.

Identification des utilisateurs uniques/singuliers (Prétraitement – Structuration) : L'identification des utilisateurs est destinée à regrouper les requêtes par utilisateur distinct sur la base de certains attributs utiles des données de journalisation, i.e., IP, Login, Agent-Navigateur. Lorsque les informations de d'authentification, cookie, ou éventuellement appliquestes Java sont disponibles, des utilisateurs uniques peuvent être identifiés. Sinon, seuls des utilisateurs singuliers peuvent être identifiés sur la base des requêtes provenant de paires distinctes d'adresses IP/Agent-Navigateur.

Construction/reconstruction des sessions (Prétraitement – Structuration) : Une session de serveur consiste en une séquence de pages consultées par un seul utilisateur au cours d'une seule visite sur le site Web. Lorsque les informations d'authentification ou de cookie sont disponibles (*système de filature/tracking systems*), la structuration des données de journalisation en utilisateurs uniques puis en sessions est triviale. Les requêtes sont triées par identifiant (*Login*) et/ou par cookies (*code numérique*) pour identifier et grouper les requêtes par utilisateur. Selon certains seuils de temps d'inactivité ou d'intervalle, les séquences de requêtes des utilisateurs unique sont divisées en sessions qui représentent des visites uniques.

L'identification de session basée sur de tels systèmes de filature est appelée méthode proactive et fournit des sessions presque réelles. Le terme proactif désigne la coopération de l'utilisateur (*authentification*) ou l'acceptation (*cookies*) à être identifié et/ou suivi. En cas de réticence des utilisateurs envers les systèmes de filature et en raison de la journalisation séquentielle et du Web dynamique, il n'est pas simple d'identifier les utilisateurs et les sessions individuels. Dans ce cas, il est recouru aux méthodes réactives, i.e., centrée sur le temps et/ou la topologie du site Web. Dans un contexte de méthodes réactives les concepts de sessions et d'utilisateurs singuliers se valent et sont permutables à même utilisés dans la littérature sans distinction [24], [26], [27], [29], [33], [36], [37], [43], [45], [46].

Identification des transactions (Prétraitement – Structuration) : Une transaction ou un épisode est un sous-ensemble significatif d'une session qui représente un intérêt particulier ou un format approprié pour la découverte de motifs d'usage. L'identification des transactions à partir des séquences de navigation au sein des sessions -en fonction de la méthode de reconstruction de session choisie- est basée sur l'une des méthodes ci-après :

- Transaction définie par le temps de navigation (*Reference Length -based*), où les pages naviguées sont classées en deux catégories, i.e., page auxiliaire ou de navigation (*temps de visite relativement court*) et page de contenu ou cible (*temps de visite relativement plus long*). Ains, en fonction de l'objectif de l'analyse, une transaction est soit toute séquence de pages de navigation/auxiliaires achevées par une page cible/contenu ; ou la séquence de pages de contenu/cible au sein d'une session.

- Transaction définie par la référence maximale en avant (*Maximal Forward Reference*), où la séquence de pages naviguées jusqu'à celle avant que l'utilisateur fasse navigation en arrière. Une page de navigation en avant est une page qui n'apparaît pas dans l'ensemble des pages déjà naviguées. Une page de navigation en arrière est une page qui apparaît dans l'ensemble des pages déjà naviguées, sinon la fin d'une transaction est référée par la dernière pages liées à la page antécédente dans la séquence de navigation.
- Transaction définie par une fenêtre temporelle (*Time Window*), où la séquence de pages naviguées dans une intervalle de temps ne dépassant pas un seuil défini. Cette méthode est basée sur l'hypothèse qu'on peut associer une norme de moyenne de temps pour les transactions significatives en fonction du contenu du site. Cette méthode débouche sur des transactions uniques par session et peut générer des transactions qui valent des sessions.

La session et les transactions sont les informations élémentaires pour la fouille des données de l'usage Web et dont la pertinence d'identification conditionne la fiabilité des motifs d'usage découverts.

Complètement des chemins de navigation (Prétraitement – Complètement des données) : Dû au système de mise en cache au niveau des proxys et des agents-navigateurs, les pages qui y sont stockées ne sont pas journalisées au niveau du serveur tant que leurs périodes d'expiration ne sont pas consommées. Ainsi, les pages mises en cache représentent une information manquante qui affecte la pertinence des sessions construites et transactions identifiées. Dans ce cas plusieurs alternatives se présentent :

- Analyser sans récupération/inférence des références/pages manquantes car la récupération/inférence peut être une source d'impertinence et d'erreur à cause du décalage d'horodatage des serveurs, proxys et terminaux.
- Le cas échéant de possibilité technique, récupération du cache des proxys et agents-navigateurs et considération de la possibilité d'erreur de décalage d'horodatage.
- L'inférence des références manquantes sur la base de la topologie réelle ou inférée du site Web dans le contexte de construction proactive de sessions.

II.2. Système et données de journalisation Web

II.2.1. Journalisation de l'activité Web

Tel que illustre par la **Figure II.3**, l'activite d'utilisation du Web designe l'activite des utilisateurs finaux, des agents-navigateurs et d'autres agents robots qui demandent des ressources de sites Web fournies via des pages Web [27], [37].

Les pages Web peuvent etre demandees sur le Web par le biais de leurs URI (*Uniform Resource Identifier*), i.e., URN (*Uniform Resource Name*) ou URL (*Uniform Resource Locator*). Ainsi, les pages Web est un ensemble de ressources qui comprend principalement :

- URI identifiant la page Web sur le WWW.
- Contenu multimedia (*texte, son, video*) destine aux utilisateurs finaux.
- Supports accessoires lies a la presentation de la page.
- Services destines aux utilisateurs finaux.
- Services destines aux agents.
- Styles, cadres et scripts qui controlent l'affichage de la page sur le terminal.

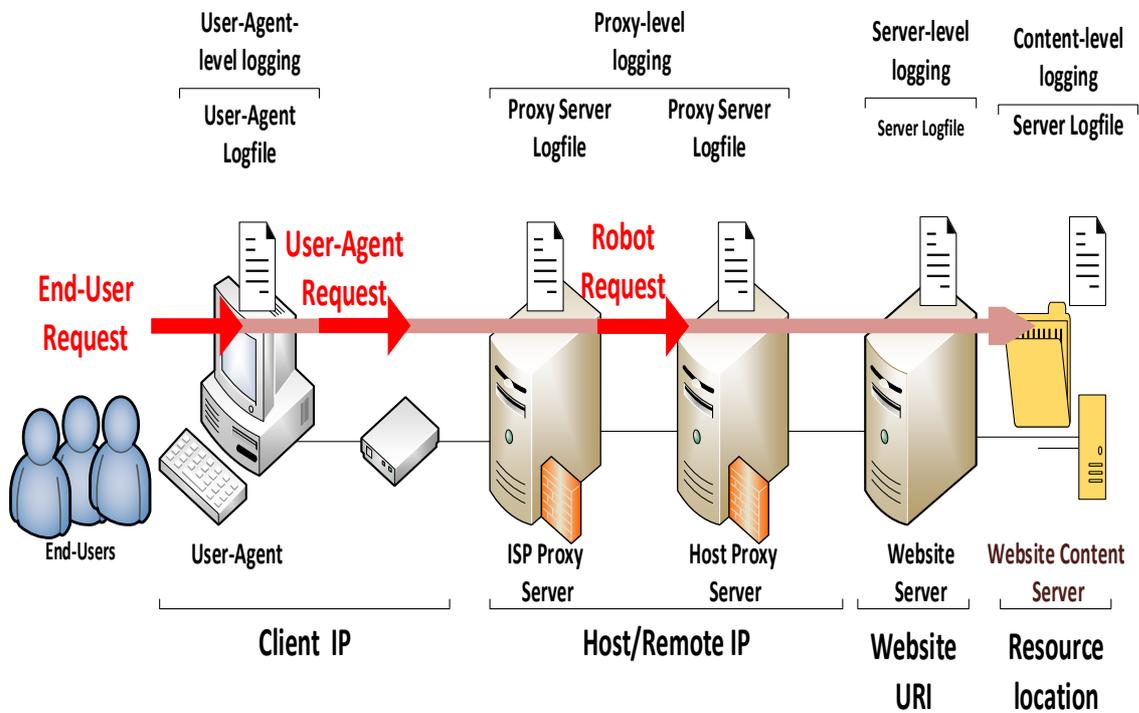


Figure II-3 Systeme d'activite Web

Ainsi, une page Web se compose de son URI cliquable destiné aux utilisateurs finaux et de ses composants chargées par les agents-navigateurs pour afficher la page sur les terminaux. Quand un utilisateur final clique/valide en tapant un URI, l'agent-navigateur charge les composants de l'aperçu de la page, et d'autres agent-robots accèdent (*hits*) aux différentes ressources du site Web pour diverses fins, e.g., indexation, commerciales.

Les clics des utilisateurs finaux et les hits agents navigateurs et robots représentent l'activité d'utilisation du Web. Les informations sur l'activité d'utilisation sont appelées de manière interchangeable Données de l'Utilisation du Web, Données d'Usage Web ou Données Log (*Web Log Data*), et sont enregistrées/journalisées à différents niveaux, i.e., client, proxy, serveur. Le niveau de journalisation principal est le serveur [21], [27], [37].

II.2.2. Données de Journalisation Web

Le format le plus utilisé des données Log dans les serveurs Web est le format de journalisation le plus ancien du National Centre for Supercomputing Applications (*NCSA*), i.e., Common Log Format (*CLF*) et Combined Log Format (*CedLF*). Ce format est adopté dans différents types de fichiers de journalisation et communément appelé format de journal commun W3C (*Common Log Format -CLF*) et format de journal étendu (*Extended Log Format ELF*)[29], [36].

CLF et ELF représente, respectivement, la configuration de journalisation par défaut et étendu des fichiers Log (*Access Log File – ALF*) gérés par les serveurs Web. La première est composée de 7 entrées/attributs, alors que la seconde, complétée par deux entrées supplémentaires, est composée de 9 entrées. Un exemple représentatif d'une instance d'un fichier ALF tiré d'un tutoriel Apache[47], [47] est donné ci-dessous.

```
[127.0.0.1] [-] [frank] [10/Oct/2000:13:55:36 -0700] [GET/apache_pb.gif HTTP/1.0] [200] [2326] [http://www.example.com/start.html] [Mozilla/4.08 (Win*; ... ;Nav*)]
```

Les entrées entre crochets sont présentées dans la **Table II.1**. Pour mieux décrire le problème de nos contributions, les attributs significatifs et le formalisme sous-jacent utilisés dans le présent manuscrit sont donnés ci-dessous en **Figure II.4**. L'accent est mis sur les attributs utiles pour analyser les méthodes identifiées objet de nos contributions et démontrer l'efficacité de nos méthodes proposées.

Table II-1 Entrées et Formats d'un Fichier ALF

[N°] [Entrée] [Format ALF]	Description
[1] [IP address] [CLF]	Adresse du client (l'hôte distant) qui a envoyé la requête au serveur
[2] [Client identity] [CLF]	Identité RFC 1413, du client, déterminée par ident sur la machine du client
[3] [Login] [CLF]	Identifiant de connexion de l'utilisateur final demandant une ressource (authentification HTTP)
[4] [Date and time] [CLF]	Heure à laquelle le serveur a fini de traiter la demande
[5] [Request] [CLF]	La ressource demandée, la méthode de requête et la version du protocole http
[6] [Status code] [CLF]	Code de réponse que le serveur renvoie au client (succès, erreur, etc.)
[7] [Size] [CLF]	La volume de l'objet renvoyé au client
[8] [Referrer] [ELF]	La ressource référente indiquée par le navigateur client
[9] [User-agent] [ELF]	Information sur le navigateur client donnée par lui-même (navigateur et système d'exploitation)

Les attributs utiles peuvent être représentés en matrice comme ci-après.

Le formalisme adopté aux termes de nos contributions est le suivant :

Soit le tuple (I, A, O) où :

- (I) est un ensemble de (m) requêtes (REQ) qui portent sur (A) représentant un ensemble de (n) attributs.
- Les occurrences (I_m, A_n) sont notées $(O_{m,n})$ aux termes d'une représentation matricielle.
- Aux termes d'une représentation transactionnelle les occurrences (O) sont indexées $I_n(A_m)$.
- Par souci de clarté, nous adoptons le formalisme transactionnel.
- E.g., $O = req_res_n$ dans la matrice ci-dessus est notée $req_n(req.res)$.

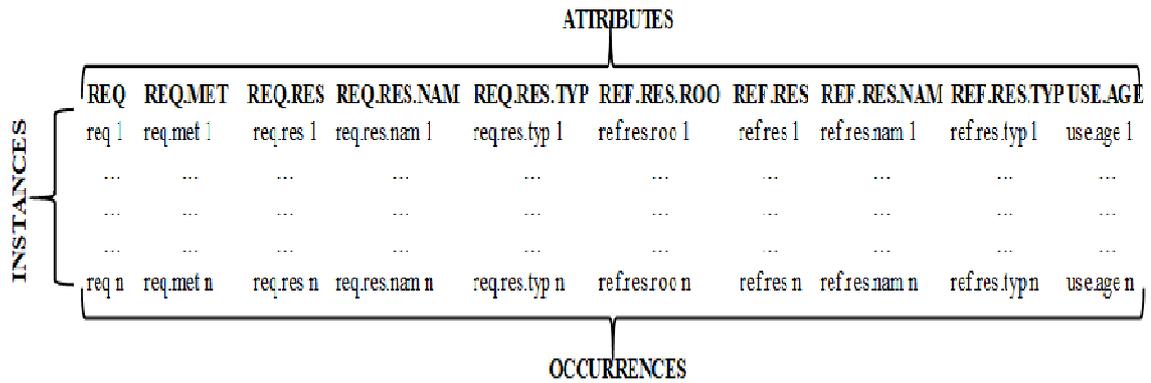


Figure II-4 Matrice des Attributs Utiles

Chaque entrée du fichier Log peut être traitée pour l'extraction de champs et attributs utiles, comme illustré en **Table II.2**, et ceci en fonction des objectifs de l'analyse, e.g., ATT.8 & 9 pour le nettoyage des données, ATT.1,3,16 pour l'identification des utilisateurs, ATT. 1,3,8,9 & 16 pour l'analyse des transactions.

Table II-2 Attributs Utiles d'un Fichier Log

ENTREE	CHAMP	ATTRIBUT	ABREVIATION	ATT N
IP Address	IP Address	IP Address	IP	ATT. 1
User Identity	User Identity	User Identity	USE_IDENT	ATT. 2
User log name	User log name	User log name	USE_LOG_NAM	ATT. 3
Date	Date and Time	Time	TIM	ATT. 4
Request	Request Method	Request Method	REQ_MET	ATT. 5
	Request String	Requested Resource Root	REQ_RES_ROO	ATT. 6
		Requested Resource Path	REQ_RES_PAT	ATT. 7
		Requested Resource File/URI	REQ_RES_NAM	ATT. 8
		Requested Resource Extension/Type	REQ_RES_TYP	ATT. 9
		Requested Resource Parameter	REQ_RES_PAR	ATT. 10
Request Protocol	Request Protocol	REQ_PRO	ATT. 11	
Status Code	Status Code	Status Code	STA_COD	ATT. 12
Size	Size	Size	SIZ	ATT. 13
Referer	Referer	Referring Resource Root	REF_RES_ROO	ATT. 14
		Referring Resource Path	REF_RES_PAT	ATT. 15
		Referring Resource URI	REF_RES_NAM	ATT. 16
		Referring Resource Extension/Type	REF_RES_TYP	ATT. 17
		Referring Resource Parameter/Header	REF_RES_PAR	ATT. 18
User Agent	User Agent	User Agent	USE_AGE	ATT. 19
	User Agent Parameter	User Agent Parameter	USE_AGE_PAR	ATT. 20
Cookie Code	Cookie Code	Cookie Code	COO_COD	ATT. 21

II.2.3. Gestion, propriétés et caractéristiques

1) Gestion de la journalisation

Dans ce qui suit nous abordons la gestion des données de journalisation telles que décrite dans le tutoriel Apache[47], [47].

Le journal d'accès au serveur, à savoir les ALF, enregistre toutes les demandes traitées par le serveur. L'emplacement et le contenu du journal d'accès sont contrôlés par les directives (CustomLog) du fichier Apache httpd. La directive LogFormat est utilisée pour définir le contenu des journaux, i.e., CLF de 7 entrées, ECLF de 9 entrées, et d'autres configurations possibles.

- **Configuration d'un CLF:**

La configuration d'un CLF est donnée comme suit :

```
LogFormat "%h %l %u %t \"%r\" %>s %b" common
CustomLog logs/access_log common
```

Les directives ci-dessus associent un nom à une chaîne de format de journal particulière. La chaîne de format consiste en indications (%) directives qui demande au serveur de consigner une information particulière, des directives en littérales après les indications directives définissant le type d'information à journaliser. Le caractère guillemet ("), le cas échéant ou il s'agit d'une indication de forme du contenu, doit être précédé par une barre oblique inversée pour qu'il ne soit pas interprété comme la fin de la chaîne de format. La chaîne de format peut également contenir des caractères de contrôle spéciaux, e.g., ($\backslash n$) indiquant les fins de lignes, ($\backslash t$) pour la tabulation.

La configuration ci-dessus écrira les entrées de journal dans le format CLF comme ci-dessous.

```
127.0.0.1 - frank [10/Oct/2000:13:55:36 -0700] "GET
/apache_pb.gif HTTP/1.0" 200 2326
```

Ce format standard peut être produit par de nombreux serveurs Web et lu par de nombreux programmes d'analyse de journaux.

- **Configuration d'un ECLF:**

La configuration d'un ECLF est donnée comme suit :

```
LogFormat "%h %l %u %t \"%r\" %>s %b \"%{Referer}i\"  
\"%{User-agent}i\" combined  
CustomLog log/acces_log combined
```

Il s'agit du format CLF avec deux entrées supplémentaires. Pour chacune de ces entrées les champs de directives supplémentaires utilisent les indications directives et littérales :

- \"%{Referer}i\", pour instruire la journalisation de la ressource référente, à savoir, celle à partir de laquelle l'utilisateur a demandé la ressource objet de la journalisation, e.g., page précédente, e.g., le site ou la page Web ayant pointé l'utilisateur vers le site ou page Web du serveur de journalisation.
- % {header} i, peut être utilisée pour instruire la journalisation dans la ressource référente l'en-tête de requête http, e.g., le mot de recherche ayant été utilisé par la requête.
- \"%{User-agent}i\", pour instruire la journalisation des informations que l'agent-navigateur du client rapporte sur soi et le client.

Le journal d'accès sous ce format se présentera comme suit :

```
127.0.0.1 - frank [10/Oct/2000:13:55:36 -0700] "GET  
/apache_pb.gif HTTP/1.0" 200 2326  
"http://www.example.com/start.html" "Mozilla/4.08 [en]  
(Win98; I ;Nav) "
```

- ***Configuration de journalisation multiple :***

Pour des raisons de gestion de l'espace de stockage des fichiers de journalisation et le contenu d'intérêt, plusieurs journaux peuvent être créés en spécifiant plusieurs directives CustomLog dans le fichier de configuration. Par exemple, les directives ci-dessous créeront trois fichiers de journaux d'accès. Le premier contient les informations de base contenues dans un CLF, tandis que les autres reçoivent, respectivement, les informations sur le référant et l'agent-navigateur.

C'est de la même manière que les requêtes ayant généré des erreurs peuvent être centralisées dans un journal distinct en ajoutant la directive ci-dessous :

```
ErrorLog [root]/[path]/nomfichier
```

Le journal des erreurs du serveur, dont le nom et l'emplacement sont définis par la directive ErrorLog, est un fichier journal important. C'est à cet endroit que le serveur enregistrera les erreurs rencontrées lors du traitement des requêtes. Il s'agit d'une information très utile pour le diagnostic des erreurs survenues.

```
[Wed Oct 11 14:32:52 2000] [error status code ] [client
127.0.0.1] client denied by server configuration:
/export/home/live/ap/htdocs/test
```

La nature de l'erreur est classée, comme les autres réponses du serveur aux requêtes dans les CLF et ECLF, par code de réponse. Les classes des codes de réponse sont :

- Classe 1xx : portant sur les requêtes et trafic d'échange d'information.
- Classe 2xx : portant sur les requêtes ayant été traitées avec succès.
- Classe 3xx : pour les requêtes ayant été redirigées.
- Classe 4xx : pour les requêtes ayant échoué dont l'erreur incombe au client.
- Classe 5xx : pour les requêtes ayant échoué dont l'erreur incombe au serveur.

Enfin, d'autres directives sont proposées pour la gestion de la rotation des fichiers Log, la virtualisation en plusieurs serveurs, drainage direct de la journalisation vers une autre location ou processus d'analyse, etc.

2) Caractéristiques du système de journalisation

- ***L'entrelacement***

Le système de journalisation consiste en un enregistrement chronologique des informations sur les requêtes des utilisateurs tant ceux finaux que ceux des agents, et ceci indépendamment de la source (*émetteur*) de la requête ou la type de la ressource demandée (objet de la requête). Ainsi, les requêtes des ressources Web (*pages Web*) utilisateurs finaux et ceux des agents-utilisateur portant sur les composantes d'affichage des pages demandées, en plus des requêtes d'autres agents-robot (*indexation*) ; sont enregistrées entrelacées dans les ALF dans un ordre séquentiel.

- ***La mise en cache***

Pour l'optimisation du trafic sur le réseau et les serveurs Web, les agents-

navigateurs des clients et les serveurs intermédiaires (*proxy*) stockent localement des copies des ressources (*pages Web*) consultées. Ainsi, quand les utilisateurs demandent ces mêmes ressources ; elles sont accessibles localement au lieu de recourir au serveur Web pour économiser la bande passante et les capacités des serveurs Web.

- ***Le décalage d'horodatage***

Les systèmes de mise en cache réduisent tant le trafic sur le serveur Web que le contenu de la journalisation car les pages consultées localement ne sont pas journalisées sur le serveur. Une journalisation incomplète affecte la pertinence de l'analyse ce qui implique, le cas échéant, la récupération des caches pour compléter la journalisation quand c'est possible. Reste que cette récupération pourrait constituer une source d'impertinence, le cas échéant de l'absence de l'information sur le décalage d'horodatage entre la source de récupération (*agent-utilisateur, proxy*) et le serveur Web. Il s'agit là d'un décalage dont l'importance dépend du débit du trafic sur le réseau.

- ***L'hébergement/convoyage multiple des requêtes***

Les requêtes d'un seul utilisateur-final peuvent être émises à partir de plusieurs agents-utilisateurs. Aussi, les requêtes de plusieurs utilisateurs-finaux peuvent être émises à partir d'un seul agent-utilisateur. Ces requêtes sont convoyées vers le serveur Web transitent par plusieurs hôtes (*serveurs intermédiaires/proxy*) Web. Ainsi, l'adresse IP de la journalisation des requêtes peut porter sur plusieurs cas de figure, i.e., Adresse IP/Utilisateur-final singuliers, Adresse IP Singulière/Utilisateurs-finaux Multiples, Adresse IP Multiple/Utilisateur-final Singulier.

- ***Les standards de développement Web***

A ce titre il importe de signaler, qu'à l'exception des normes/recommandations de W3C, il n'existe pas de standards universels coercitives en matière de conception et développement Web. L'objectif de la W3C, fondée en 1994, est favoriser -via la normalisation- l'exploitation optimale des potentialités du Web sur plusieurs plans dont, entre autres, le format des ressources Web et l'analyse des données Web « Web of Data and Services » [48]. La normalisation des formats des ressources Web joue un rôle important en matière de pertinence des tâches critiques de la fouille des données de l'usage du Web et les résultats des analyses de la fouille.

- ***Les technologies de conception Web***

Contrairement au Web statique, le format, le contenu et la structure du Web dynamique et adaptatif sont générés par un script en fonction des paramètres passés dans les requêtes des utilisateurs, e.g., l’affichage, les liens entre les pages, le contenu des pages en fonction des spécifications du client (*système d’exploitation, navigateur, etc.*), des profils utilisateurs-finaux, historique des chemins traversés, etc.

Ces caractéristiques représentent des contraintes majeures qui affectent la pertinence des tâches critiques du processus de fouille des données de l’usage Web, i.e., le nettoyage et la structuration des données. Cet impact contraignant affecte aussi la pertinence des analyses et des applications [21], [22], [25]–[27], [29]–[31], [33], [35], [36], [43], [49], e.g., l’optimisation de l’usage Web, basées sur la fouille des données d’usage Web et le retour escompté d’une « analyse de toutes les données, de tous les utilisateurs, tout le temps d’utilisation » et son avantage par rapport aux méthodes traditionnelles d’analyse IHM du Web, i.e., orientées utilisateur, expert, analytique.

II.3. Revue de littérature et analyse critique

II.3.1. Classification de la littérature

1) Recherches originales fondatrices

Les contributions fondatrices et/ou originales sont les premières recherches initiatrices de la discipline et/ou celle fournissant les premières méthodes et techniques conventionnelles d’analyse des données de l’usage Web ; ou des méthodes alternatives originales qui traitent des limites des méthodes conventionnelles, e.g., [21], [27]–[29], [31], [31], [32], [34], [35], [35], [37], [49].

2) Réplication et validation

Il s’agit des travaux empiriques expérimentant les méthodes proposées et des pratiques d’application les reproduisant à des fins différentes. Les recherches intéressées par la validation visent souvent l’évaluation de l’efficacité des méthodes et techniques proposées. Les répliques ne traitent pas nécessairement de la qualité des méthodes. Il s’agit plutôt de l’application des méthodes proposées pour la résolution de problèmes spécifiques via la fouille des données de l’usage Web, e.g., [25], [30], [31], [50]

3) Implémentation et automatisation

De nombreux formalismes et algorithmes de mise en œuvre des méthodes et techniques liées au processus de fouille des données de l'usage du Web sont proposés. Plusieurs outils d'implémentation et d'automatisation sont rapportés dans la littérature. S'agissant des projets de la communauté universitaire, les outils dédiés et les projets identifiés en la matière ne sont plus disponibles [51]–[58]. L'analyse et la fouille des données de l'usage du Web est prise en charge, actuellement, par les différents services et applications de préparation et traitement de données et les écosystèmes Big Data comme « Logstash » de « Elasticsearch ».

4) Etat de l'art et revue de littérature

Des recherches descriptives, analytiques et revues de littératures couplées à des pratiques commerciales ont été au cœur de l'émergence de la fouille des données de l'usage Web en discipline universitaire et professionnelle. Outre les premières contributions fondatrices, il existe une batterie de papiers ayant entretenu régulièrement un aperçu actualisé de la discipline, e.g., [20], [22], [23], [26], [38], [54], [54], [55], [58]–[65].

II.3.2. Revue de la littérature, limites et contributions visées

1) Nettoyage des données de journalisation

- *Travaux connexes*

Les premiers travaux [21], [32], [34], [35], [37] sur le nettoyage des données de journalisation, visant la distinction des requêtes des utilisateurs finaux de celles des agents, proposent des heuristiques de filtrage centrées sur le contenu de la journalisation. Il s'agit d'une méthode de nettoyage conventionnelle qui est basée sur les recommandations de la W3C de mettre les ressources destinées aux utilisateurs-finaux sous le format HTML ou équivalent pour les distinguer de ceux destinées aux agents.

Avec le développement du Web adaptatif basé sur des frames sans extensions de type de fichier, l'heuristique de filtrage par type de ressource devient obsolète. Une méthode de nettoyage avancé est proposée [28] pour surmonter cette contrainte. Cette méthode repose sur le même principe de filtrage heuristique. Par contre, elle se sert d'une base de connaissance pour reconnaître les ressources destinées aux utilisateurs-finaux. Cette base de connaissance contient l'identification des ressources consultables par les utilisateurs-finaux à partir des URI qui leurs sont destinés.

- **Limites**

Les deux méthodes reposent sur des heuristiques de filtrage centrées sur le contenu de la journalisation. Ces méthodes de nettoyage sont limitées en termes de pertinence, et de contraintes d'applicabilité et de coût, dans le contexte d'un contenu Web Dynamique/Adaptatif. L'absence de standards obligatoires en termes de format des ressources (*contenu de la journalisation*) destinées aux utilisateurs finaux conjuguée au changement en continu des ressources du Web dynamique/adaptatif affecte la pertinence du nettoyage dans l'absence d'une mise à jour en temps réel de la base de connaissance identifiant les ressources destinées aux utilisateurs finaux. Par conséquent, la mise à jour en temps réel de cette base de connaissance est un processus qui génère un surcoût et susceptible de saturer le serveur dans le cas du Web dynamique à même impossible [24].

- **Première Contribution visée**

Etant donné que les limites des méthodes de nettoyage et les contraintes du contexte du Web Dynamique/Adaptatif sont liées au contenu de la journalisation ; notre contribution vise l'introduction d'une méthode de nettoyage basée sur la structure de la journalisation. En effet, la structure de journalisation dépend seulement de son format. Le format de la journalisation est insensible aux répercussions du Web Dynamique/Adaptatif en termes de contenu. Ainsi, l'objectif visé est de proposer une méthode de nettoyage basée sur les règles de la structure de journalisation devant garantir des sortants pertinents, sans avoir besoin de base de connaissance a priori, et sans générer des surcoûts.

2) Structuration des données de journalisation

- **Travaux connexes**

Comme le nettoyage des données de l'usage Web, les méthodes de structuration en utilisateurs et sessions remontent aux premières travaux [21], [32], [34], [35], [37] de recherche en fouille des données Web. Les méthodes de structuration s'inscrivent dans l'une des deux approches suivantes :

- L'approche proactive quand l'information d'authentification ou de filature (*cookies/tracking system*) est disponible.
- L'approche réactive dans l'absence de cette information, le cas échéant de réticence de l'utilisateur-final.

En fonction de l'approche adoptée, trois méthodes peuvent être utilisées pour structurer/grouper les requêtes par utilisateur, session et transaction, i.e., générique, orientée temps de navigation, topologie réelle ou inférée du site Web [21], [26], [29], [32], [34], [37]. Ces trois méthodes sont basées sur l'identifiant de l'agent, i.e., identifiant utilisateur (*cas d'authentification*), Adresse IP, nom de l'Agent (navigateur et système d'exploitation). En effet, le processus de base consiste à regrouper les requêtes des utilisateurs par agent, puis les structurer en sessions transactions en fonction de la méthode adoptée.

Ces deux méthodes, qui remontent à la période de l'émergence de la fouille des données Web comme discipline 1997-2002, demeurent à ce jours la base des différents outils et services d'analyse des données de journalisation Web. A cet égard, il est important de signaler que le domaine de l'analyse des données de l'usage du Web est à forte valeur ajoutée pour la sphère commerciale[20], [33]. Par conséquent, seuls les outils d'analyse issues de cette sphère qui ont prévalu sur ceux proposés par le monde académique, pourtant plus explicite en matière des techniques utilisées. A ce jour, tous les outils qui ont été proposées par le monde académique n'ont pu s'assurer une pérennité.

- ***Limites***

En effet, le cas échéant d'absence d'authentification et dans un contexte de structuration réactive, les approches centrées sur l'agent ne peuvent pas distinguer plusieurs utilisateurs utilisant le même client/machine (même navigateur et système d'exploitation) ou passant par un proxy à IP unique, ou un utilisateur par un proxy à IP multiple. Aussi, dans le cas d'indisponibilité de l'information sur la topologie réelle du site (*Web Dynamique*), les méthodes de structuration en sessions tendent à construire des sessions plus courtes que celles réelle qui auraient pu être identifiées, le cas échéant de disponibilité de l'information d'authentification (*approche proactive*). En plus, les méthodes réactives ne peuvent pas identifier des sessions d'utilisateurs ayant navigué via plusieurs IP et clients lors de la même session. Enfin, les systèmes de mise en cache constituent aussi une source d'impertinence de la structuration. D'une part, la récupération du cache n'est pas sans erreur en l'absence de l'information sur le décalage d'horodatage (*Serveur/Client*), et d'autre part, la non journalisation des ressources consultées à partir du cache conduit à la construction de sessions courtes qui ne reflètent

pas les sessions réelles [26], [29], [33], [43].

Ainsi, la pertinence des méthodes actuelles de structuration, particulièrement celles réactives, demeurent fortement affectée des contraintes liées à l'évolution des technologies Web, i.e., Web dynamique/adaptatif, mise en cache, Proxy à IP unique/Multiples.

- ***Deuxième Contribution visée***

Les limites des méthodes de structuration, présentées ci-dessus, reviennent au fait que la structuration en sessions est basée sur l'identification des agents. Or, l'évolution du contexte des technologies Web rendent complexe à même impossible cette identification le cas échéant de réticence des utilisateurs à l'authentification et aux systèmes de filature. A cet égard, notre contribution vise l'introduction d'une méthode centrée sur les flux de clics au lieu des agents émetteur des clics. Le processus de notre méthode de structuration -piloté par des contrainte de pertinence- reconstruira des sessions dont la longueur sera la plus proche possible des sessions réelles, et ceci tout en ayant la possibilité d'identifier des sessions à hébergement multiples.

3) Evaluation et optimisation de l'usage Web

- ***Travaux connexes***

L'évaluation de l'usage Web sur la base des données de journalisation constitue une application de la fouille des données d'usage Web à l'analyse de l'interaction Homme-Machine basée sur les traces d'utilisation [13]. Elle représente une approche récente en matière d'évaluation IHM. Cette approche orientée Expérience Utilisateur (*UX*) est destinées à surmonter les limites des approches déjà établies, i.e., orienté utilisateur, expert, et analytique qui sont basées sur des instruments tels que les référentiels, l'enquête, l'observation, etc. [5], [6], [12], [14]–[16], [36], [66]–[68].

Ces méthodes sont limitées en termes d'objectivité, neutralité et exhaustivité sur le plan des dimensions et de la population analysée. Les méthodes récentes basées sur l'expérience utilisateur, les traces d'usage et la symbiotique visent plus d'exhaustivité et d'objectivité en termes d'évaluation IHM. L'analyse des traces d'utilisation permet d'avoir un aperçu objectif sur l'expérience réelle tandis que l'observation de la dimension symbiotique prend en charge l'ensemble des partis prenantes (*utilisateur, concepteur,*

analyste, artefact, etc.) et les valeurs à mesurer y afférentes (*qualité technique, utilité, utilisabilité, etc.*). Néanmoins, les méthodes développées à cet effet demeurent intrusives et moins exhaustives en termes de la population incluse tout en générant un important volume de données à traiter.

C'est dans ce contexte que l'application des techniques de fouille de l'usage Web aux données de journalisation pour l'évaluation de l'usage du Web (*évaluation IHM du Web*) se présente comme une alternative dont l'ambition est de : « Analyser toutes les données de tous les utilisateurs, tout le temps d'une manière objective et neutre ». A ce titre, l'objectif principal de l'évaluation est l'optimisation de l'usage en amont et/ou en aval de la conception initiale/nouveaux services. La revue de littérature à cet égard [13], [21], [25]–[27], [30], [31], [33], [35], [36], [69]–[72] a permis d'identifier trois axes d'optimisation, i.e., le trafic sur le serveur et le réseau, les chemins de navigation et la structure du site Web.

L'optimisation repose sur la fouille des données de journalisation via techniques appropriées, e.g., Chaîne de Markov, Règles d'Associations, Règles Séquentielles, etc., pour l'extraction de connaissance visant l'amélioration des temps de réponse (*optimisation du trafic*) par la mise en cache, permettre aux utilisateurs d'atteindre les contenus d'intérêts via des chemins raccourcis (*optimisation des chemins traversés*), et l'adaptation de la structure du site Web à leurs habitudes de navigation (*optimisation de la structure du site Web*), et ceci en liant les pages souvent visitées durant la même, le cas échéant d'absence de lien. Dans le contexte de Web dynamique, ces optimisations sont entreprises automatiquement par des systèmes intelligents, e.g., les systèmes de recommandations, de personnalisation, de placements publicitaires, etc.

- ***Limites***

Une telle démarche d'optimisation par dimension n'est pas symbiotique car elle peut déboucher sur des conflits d'optimisations par rapport aux intérêts des différents acteurs concernés, à savoir :

- L'artefact, en l'occurrence le site Web, le serveur, et le réseau dont l'intérêt réside dans la fluidité du trafic et l'optimisation de la consommation des ressources, la consommation de la bande passante, etc. ;
- Les utilisateurs-finaux dont l'intérêt réside dans l'optimisation des temps de réponse,

des structures adaptées à leur habitude de navigation ;

- L'analyste qui peut être le concepteur, le propriétaire du serveur ou du commerce dont l'intérêt porte sur la disponibilité de l'information de journalisation pour l'analyse devant servir la conception de services futures, marketing, publicité, etc.

Par conséquent, et à titre d'exemple :

- La mise en cache pour l'intérêt de l'artefact réduit la portée et la pertinence de l'analyse car les pages mises en cache ne sont pas journalisées sur le serveur ;
- Le raccourci des chemins de navigation dans l'intérêt des utilisateurs peut générer un fort flux sur le réseau par l'effet de l'aiguillage de la navigation et sa concentration dans le temps. Ainsi, une telle optimisation peut se retourner contre l'intérêt de l'utilisateur.
- L'adaptation de la structure peut influencer sur le trafic de par la taille des ressources et la longueur des nouveaux chemins de navigation générés.

- ***Troisième Contribution visée***

L'optimisation conventionnelle de l'utilisation du Web basée sur les données de journalisation porte distinctement sur trois dimensions, i.e., le trafic sur le réseau et le serveur, la structure du site Web, et les chemins de navigation. Il s'agit de dimensions pas nécessairement corrélées et qui dépendent de trois facteurs, i.e., le temps de consultation des ressources/pages Web, leur volume, et la longueur des chemins parcourus/nombre de ressources sollicitées. Les analyses statistiques peuvent donner un aperçu explicatif sur l'optimisation globale. Cependant, comme le facteur temps est hors de contrôle, elles ne sont pas en mesure de fournir une solution pour l'équilibre et le contrôle du produit symbiotique de l'optimisation. Ainsi, la contribution ciblée vise à fournir une méthode d'apprentissage semi-supervisé pour équilibrer/contrôler la valeur symbiotique de l'interaction et éviter des optimisations de dimensions au détriment d'autres.

Chapitre III

Première Contribution

Nettoyage des Données De l'Usage du Web

III.1. Contexte, problème et objectif

En premier lieu, au titre de notre contribution nous analysons les limites des méthodes de nettoyage des données de l'usage Web en termes de pertinence, applicabilité et de coûts. En deuxième lieu, nous proposons deux méthodes de nettoyages d'une meilleure pertinence, sans besoins de connaissances apriori et sans facteurs de coûts supplémentaires.

La fouille des données de l'utilisation du Web est l'application des techniques de fouille de données aux données de l'utilisation du Web, communément appelés données de journalisation Web (*Weblog Data - WLD*), à des fins différentes, e.g., l'amélioration et l'adaptation du système, la personnalisation, la recommandation et le placement des annonces publicitaires [21], [22]. En fonction de l'objectif et du contexte de l'analyse, les WLD sont complétées par d'autres données [24] et sont fouillées à l'aide de différentes techniques de fouille de données, e.g., statistique descriptive, classification, partitionnement, règles d'association/séquentielles et régressions. Au titre des étapes classiques d'un processus de fouille de données et extraction des connaissances, l'étape de pré-traitement du processus de fouille des WLD se caractérise par trois tâches critiques, i.e., le nettoyage des données, la structuration des données en utilisateurs et visites uniques/singuliers et l'identification des épisodes/transactions. [21], [38], [44], [73].

La tâche de nettoyage des données conditionne la pertinence de l'ensemble du processus de fouille de données et d'extraction de connaissance [26], [27], [29]. Cette tâche consiste à traiter la Base de Données de WLD brutes (*Raw Weblog Data - R. WLDB*) pour fournir une Base de Données de WLD sans bruit/Nettoyée (*Cleaned Weblog Data - C. WLDB*).

Une R. WLDB fait référence aux données enregistrées par les serveurs Web dans des fichiers de journalisation (*Access Logfiles - ALF*). Ils reflètent l'activité d'utilisation sur les serveurs de sites Web, c'est-à-dire les clics de l'utilisateur final et les hits sous-jacents d'agent d'utilisateur [21], [27]. Notez que les clics de l'utilisateur final et les hits de l'agent utilisateur portent, respectivement, sur l'identifiant uniforme de ressource (*Uniform Resource Identifier - URI*) des pages Web et à leurs composants de la vue d'écran affichées.

Étant donné que la fouille de l'usage du Web s'intéresse au comportement des utilisateurs finaux, les hits des agents utilisateurs sont considérés comme bruit à identifier et à nettoyer avant la fouille. Filtrer les hits parmi les clics n'est pas une tâche simple pour deux raisons : les serveurs enregistrent les requêtes entrelacées dans un ordre séquentiel, indépendamment de leur source ou de leur type ; les ressources des sites Web peuvent être requêtées de manière interchangeable par les utilisateurs finaux et les agents utilisateurs [21], [27], [28].

Sur la base de la littérature relative à la fouille de l'usage du Web depuis WEBKDD'99 jusqu'à 2017 [20]–[24], [38], [44], [73], [74], la tâche de nettoyage peut être divisée en trois couches. La première est destinée à identifier les clics et les hits. La seconde consiste à sélectionner, parmi les clics identifiés, ceux qui sont significatifs aux fins de l'analyse et du contexte, e.g., les requêtes abouties/échouées, les serveurs proxy à IP uniques et/ou robots connus. [24], [26], [44]. La troisième intervient en aval de la structuration des données. Elle est plutôt un nettoyage centré comportemental et destiné à la détection de robots et des valeurs aberrantes. [24], [27], [30].

Ainsi, la couche de nettoyage centrale qui s'attaque au bruit générique est l'identification des clics des utilisateurs finaux parmi les hits des agents. À cet égard, les méthodes les plus citées et mises en œuvre sont le nettoyage conventionnel et celui avancé. Ces deux méthodes reposent sur des heuristiques de filtrage centrées sur le contenu. Elles sont basées sur l'attribut de ressource demandé de la R. WLDB. Ces méthodes de nettoyage sont limitées en termes de pertinence, et de contraintes d'applicabilité et de coût, dans le contexte d'un contenu Web Dynamique/Adaptatif (*Dynamic and Responsive Web Design*).

Afin de faire face aux contraintes du Web Dynamiques/Adaptatif, une approche de nettoyage basée sur la structure de la journalisation, centrée sur les règles de la structure de journalisation, au lieu du contenu, est introduite dans cette thèse. Comme la structure de journalisation est insensible au contenu de la journalisation, l'expérimentation de la méthode de nettoyage basée sur les règles montre des avantages significatifs par rapport au nettoyage conventionnel et avancé, qui sont basés sur le contenu de la journalisation.

III.2. Analyse des méthodes de nettoyage – Travaux connexes

III.2.1. Données et formalisme

Pour mieux décrire les méthodes de nettoyage ; le contenu des données de journalisation, les formats, les attributs et le formalisme sous-jacent sont présentés ci-dessous.

Les formats des fichiers de journalisation (*Access Log File – ALF*) les plus utilisés sont ceux du format commun (*Common Log Format – CLF*) et celui étendu (*Extended Common Log Format – ECLF*). Ces fichiers contiennent des entrées d'informations sur l'activité d'utilisation, Ils contiennent, respectivement, 7 et 9 entrées [27]. Le contenu des entrées ALF décrit dans la **Table III.1** représente le contenu de la R. WLDB.

Table III-1 Entrées d'un fichier de journalisation ALF/ECLF

Entrée	Description
IP address	Adresse du client (l'hôte distant) qui a envoyé la requête au serveur
Client identity	Identité RFC 1413, du client, déterminée par ident sur la machine du client
Login	Identifiant de connexion de l'utilisateur final demandant une ressource (authentification HTTP)
Date and time	Heure à laquelle le serveur a fini de traiter la demande
Request	La ressource demandée, la méthode de requête et la version du protocole http
Status code	Code de réponse que le serveur renvoie au client (succès, erreur, etc.)
Size	La volume de l'objet renvoyé au client
Referrer	La ressource référente indiquée par le navigateur client
User-agent	Information sur le navigateur client donnée par lui-même (navigateur et système d'exploitation)

Soit une R. WLDB de (n) requêtes (*REQ*), contenant des instances de 7 entrées d'un ALF, comme suit :

$REQ_n = \{127.0.0.1, -, frank, 10/Oct/2000 :13:55:36 -0700, GET/apache.gif HTTP/1.0, 200, 2326, http://apache.org/example.html, Mozilla/4.08 (Win*; ... ;Brow*) \}$.

Notez, ci-dessous, les attributs et abréviations utiles pour notre analyse.

- [*GET*], la méthode de la requête (*the request method – REQ.MET*) ;
- [*apache.gif*], la ressource demandée (*the requested resource – REQ.RES*) ;
- [*apache*], le nom de la ressource demandée (*the requested resource name – REQ.RES.NAM*) ;

- *[gif]*, le type de la ressource demandée (*the requested resource type – REQ.RES.TYP*);
- *[apache.org]*, la racine de la ressource référente (*the referring resource root – REF.RES.ROO*);
- *[example.html]*, la ressource référente (*the referring resource – REF.RES*);
- *[example]*, le nom de la ressource référente (*the referring resource name – REF.RES.NAM*);
- *[html]*, le type de la ressource référente (*the referring resource type – REF.RES.TYP*);
- *[Mozilla/4.08 (Win*; ... ;Brow*)]*, l'agent-utilisateur (*the user-agent – USE.AGE*).

Chaque occurrence d'un attribut est indexée $req_n(att_m)$. E.g., $req.res=apache.org$ dans REQ_n ci-dessus est indexée $req_n(req.res)$.

III.2.2. Perspectives de nettoyage

Sur la base de la revue de littérature depuis WEBKDD'99 jusqu'à 2017 [20]–[23], [46], [59], [74], [75], nous distinguons trois couches de nettoyage, à savoir un nettoyage générique, contextualisé et comportemental. Le nettoyage comportemental est effectué en aval de la structuration des données. Ce type de nettoyage vise plutôt la détection des valeurs aberrantes et les robots. C'est un nettoyage effectué en aval de la structuration des données en utilisateurs uniques/ singuliers. Il est communément appelé nettoyage intelligent [27]–[29], [29]–[31] et dépasse le cadre de notre recherche.

Ainsi, la couche de nettoyage centrale qui s'attaque au bruit générique est l'identification des clics parmi les hits. À cet égard, les méthodes les plus citées et appliquées, dans la littérature, sont le nettoyage conventionnel et celui avancé.

III.2.3. Méthodes de nettoyage

1) Nettoyage Conventionnel

Le nettoyage conventionnel repose sur l'hypothèse que les ressources de type HTML ou équivalent qui sont destinées aux requêtes/clics de l'utilisateur final, comme le recommande le World Wide Web Consortium (W3C) [21], [23]. Toutes les demandes liées à des types de ressources non html sont supposées être des requêtes d'agent d'utilisateur, soit des hits/bruit. Ainsi, l'attribut de nettoyage c'est le « *REQ.RES.TYP* ».

En pratique, les « *REQ.RES.TYP* » non désirés sont définis dans une base de données de connaissances filtrante (*filtering knowledge database – FLT.KDB*) servant

l'heuristique de nettoyage. Les ressources y afférentes sont supprimées (*FILTER-OUT*) de la R. WLDB pour créer une base de données d'usage du Web nettoyée (*cleaned Weblog Database – C. WLDB*).

Le processus de nettoyage Conventionnel est donné dans l'**algorithme III.1**.

Algorithme III-1 Nettoyage Conventionnel	
01	INPUT DATA
02	R. WLDB
03	FLT.KDB
04	PROCESS
05	SCAN R. WLDB
06	IF req _n (req. res. typ) ∈ FLT. KDB FILTER-OUT req _n
07	OUTPUT DATA
08	C.WLDB

2) Nettoyage Avancé

Le nettoyage avancé est basé sur des connaissances préalables sur les URI de sites Web ou sur une extraction en temps réel de celles intégrant des ressources cliquables [23], [28]. Le but est de construire une base de données de connaissances de validation (*validation knowledge database – VLD.KDB*) contenant toutes les ressources incorporées dans les URI du site Web, qui sont destinées/pouvant être demandées aux/par clics de l'utilisateur final. Toutes les demandes contenant des ressources appartenant à la VLD.KDB sont supposées être des clics et sont validées/filtrées vers (*FILTER-IN*) la C. WLDB. Ainsi, l'attribut de nettoyage est le « *REQ.RES* ».

Le processus du Nettoyage Avancé est donné dans l'**algorithme III.2**.

Algorithme III-2 Nettoyage Avancé	
01	INPUT DATA
02	R. WLDB
03	VLD.KDB
04	PROCESS
05	SCAN R. WLDB
06	IF req _n (req. res) ∈ VLD. KDB FILTER-IN req _n
07	OUTPUT DATA
08	C.WLDB

3) Limites et proposition

Pour obtenir des résultats de haute qualité, le Nettoyage Conventionnel et celui Avancé nécessitent une connaissance préalable exhaustive, d'autres connaissances que les données d'usage du Web/données de journalisation (*données extra-journalisation*) et d'autres coûts, en plus du coût du processus de nettoyage (*extra-coûts*).

Le Nettoyage Conventionnel nécessite des connaissances préalables exhaustives et des données extra-journalisation relatives aux types de ressources destinées aux requêtes des utilisateurs finaux ou ceux destinées à des agents utilisateurs, appelées bruit. En l'absence de normes obligatoires à cet égard, la construction et la mise à jour de la FLT.KDB représentent un facteur de coût supplémentaire en plus du coût du processus de nettoyage. De plus, le nettoyage conventionnel devient obsolète dans le cas de sites Web basés sur des frames sans extensions de type de fichier, qui sert d'attribut de nettoyage.

Le Nettoyage Avancé est basé sur une base de connaissance de validation « VLD.KDB » listant les ressources « REQ.RES » intégrées à la structure du site Web, donc destinées à l'utilisateur final. Ainsi, cette méthode de nettoyage a besoin de données extra-journalisation liées à la structure du site. Dans le cas d'une conception Web dynamique basée sur une personnalisation/adaptation automatique du contenu/structure ; le contenu et la structure du site Web changent constamment. Dans ce cas de figure, le Nettoyage Avancé induira une surconsommation de la bande passante, peut provoquer la saturation du serveurs, et même impossible dans le cas de personnalisation automatique de contenu [24].

Ces cas de figure sont dû au fait que la construction et la mise à jour de la base VLD.KDB nécessitent une extraction continue des URI destinées aux utilisateurs finaux du site Web. Un tel processus est très complexe à réaliser, représente un facteur d'extra-coût et peut manquer de filtrer, vers la base VLD.KDB, les contenus créés/annulés ; si le processus d'extraction vers la base VLD.KDB n'est pas synchronisé avec le processus de nettoyage en temps réel.

Enfin, le fait que les deux méthodes nécessitent une connaissance préalable exhaustive du contenu du site Web, de sa structure et de ses ressources destinées aux utilisateurs finaux pour la mise en place de l'heuristique de filtrage constitue une faiblesse

critique. L'obtention d'une connaissance préalable assez exhaustive dans la perspective analytique des propriétaires de serveur n'est pas évidente, car les propriétaires de serveur ne possèdent pas nécessairement cette connaissance, contrairement aux propriétaires de sites Web [13].

III.3. Contribution 1 : Nettoyage basé sur la structure de la journalisation

III.3.1. Contribution 1.1 : Méthode heuristique

1) Concept et approche

L'objectif est de proposer une méthode de nettoyage garantissant les mêmes avantages du Nettoyage Avancé (*sans bruit*), et ceci sans avoir besoin de données extra-journalisation, de connaissances préalables ou de facteurs de coût supplémentaire (*extra-coût*) au-delà du coût du processus de nettoyage. Une telle méthode vise à être applicable dans le contexte d'un Web dynamique/adaptatif, et exploitable dans les deux perspectives d'analyse, i.e., propriétaires de serveurs et/ou de sites Web.

Notre méthode filtre la R. WLDB en fonction des règles de la structure de journalisation. Étant donné que la structure de journalisation dépend uniquement du format de journalisation, cette approche est insensible aux répercussions du Web dynamique/adaptatif en termes de structure et de contenu de sites Web. Ainsi, l'objectif du nettoyage, basé sur les règles de la structure de journalisation, est de permettre de surmonter les contraintes de nettoyage liées au nettoyage avancé et conventionnel.

La méthode consiste en trois étapes principales, à savoir :

- L'identification de motifs structurels de journalisation liés aux clics distinctement des hits ;
- L'abstraction des motifs en règles de nettoyage,
- L'implémentation des règles via un algorithme de nettoyage qui sera expérimenté et évalué.

Le nettoyage basé sur la structure repose sur les règles liées aux motifs réguliers de la structure de journalisation des clics distinctement des hits. Une liste pertinente de requêtes/clics de navigation-utilisateur final (*liste pertinente de navigation*) a été prédéfinie. La liste pertinente de navigation représente les différentes manières via lesquelles un utilisateur final accède aux URI/Ressources d'un site Web, e.g., un accès en

cliquant sur les liens d'un moteur de recherche ou d'un site Web tiers, en tapant l'URI dans la barre de recherche du navigateur, en cliquant un lien déjà enregistré comme favori ou la navigation en arrière. A cet égard, une liste non exhaustive est donnée dans la **Table III.2.**

Sur la base de la liste pertinente de navigation ; une simulation de navigation d'utilisateurs finaux a été effectuée via une application dédiée (*Webserver Stress Tool 8.0.0.1010*) sur une copie d'un site Web (*Simple English Wikipedia*), sous serveur local (*serveur Apache*), configuré pour générer un fichier ALF de format ECLF.

Table III-2 Liste pertinente de requêtes de navigation-utilisateur final

N°	Liste pertinente	Catégorie
1	Accéder à une page d'accueil en tapant des mots clés dans un moteur de recherche	Accès via page d'accueil/URN
2	Accéder à une page d'accueil en cliquant sur un lien/l'URL indexé dans le moteur de recherche	
3	Accéder à une page d'accueil en tapant/collant un lien/URN dans le navigateur	
4	Accéder à une page d'accueil par un lien/URN marqué favori dans le navigateur	
5	Accéder à une page d'accueil par un lien externe/URN indexé dans un site Web tiers	
6	Accéder à une page de contenu par un lien/l'URL marqué favori dans le navigateur	Accès via page de contenu/URL
7	Accéder à une page de contenu par lien indexé/URL dans un moteur de recherche	
8	Accéder à la page de contenu en tapant/collant un lien/l'URL dans le navigateur	
9	Navigation de liens internes/URL via un seul navigateur	Navigation mono/multi navigateur
10	Navigation de liens internes/URL via plusieurs navigateurs	
11	Fin de navigation sur un site Web via lien externe/URI	Fin de navigation
12	Fin de navigation par temps mort (<i>timeout</i>)	
13	Navigation du cache navigateur (<i>navigation par retour en arrière ou de liens marqués favoris dans le navigateur</i>)	Navigation du cache

L'ALF généré est collecté et les requêtes liées aux URI parcourus (*liste pertinente/clic d'utilisateur final*) sont labélisées distinctement des hits sous-jacents (*hits d'agent utilisateur*). Les motifs réguliers identifiés en termes des attributs REQ.RES et REF.RES de la structure de journalisation des clics sont abstraits pour déduire les règles sous-jacentes aux clics en dehors des hits. Les règles inférées sont combinées dans un processus de filtrage qui filtre les éléments pertinents vers la C. WLDB.

Pour effectuer un test de nettoyage sur des ALF de sites Web tiers nous avons conçu un algorithme d'implémentation sous R via l'API Apache Spark. Les résultats du nettoyage basé sur des règles sont comparés au nettoyages Conventionnel et Avancé.

2) Méthode d'application

Les caractéristiques des attributs et fonctions des motifs identifiés de la structure de journalisation sont présentées dans la **Table III.3**. Trois caractéristiques génériques (*a*, *b*, *c*) sont identifiées dans la structure d'ALF :

- Accès au site Web par URN (page d'accueil–Home Page ; Nom de domaine–DN ; Nom de la ressource – URN) identifiable par un attribut de ressource demandé (*REQ.RES*) vide, entre les attributs de requête (*REQ*) et de protocole http (*REQ.PRO*) **Table III.3 (a)**.
- (*b*) Navigation sur le site Web via les URL (*page de contenu*) identifiables par la relation entre l'attribut URL de la page de contenu demandée (*REQ.RES*) et l'attribut de ressource référente (*REF.RES*) des composantes de la page demandée (*REQ.RES*) **Table III.3 (b)**.
- (*c*) Bruit interférant avec les motifs réguliers ci-dessus qui est dû aux requêtes de Web adaptatif, à savoir les requêtes de style (*SQ*) qui sont entrelacées dans les requêtes des ressources référentes (*REF.RES*) enregistrées dans l'entrée REF de l'ALF **Table III.3 (c)**.

Table III-3 Motif régulier de structure de journalisation

REQ	REQ.RES [nome & type]	REF.RES [nom & type]	Clic/Hit
(a) HOMEPAGE ACCESS f (REQ.RES n)			
R n	empty (URN)	Internal V External Referer/Page	Click
R $n+1$	Media 1	Home Page/URN	Hit
...	...	Home Page/URN	...
...	Media m	Home Page/URN	Hit
(b) CONTENT PAGE ACCESS f (REQ.RES n & REQ.RES$n+1$)			
...	Content Page (URL)	Internal V External Referer/Page	Click
...	Media 1	Content Page/ URL	Hit
...	...	Content Page/ URL	...
...	Media m	Content Page/ URL	Hit
(c) INTERFERING NOISE f (REF.RES.TYP n) [style query type]			
...	Homepage/Content Page (URI)	Internal V External Referer/Page	Click
...	Media name	Media style query	Hit
...	...	Content Page/ URL	Hit
R $n+x$	Media m	...	Hit

Les motifs réguliers identifiés peuvent être formalisés en trois règles, à savoir, accès/clics URN, accès/clics URL, et bruit/hits interférant réguliers.

$$\text{ACCESS URN} \Rightarrow \text{req}_n((\text{req. met} \ \& \ \text{req. pro}) \neq \emptyset \ \& \ \text{req. res} = \emptyset) \quad \text{Eq. III.1}$$

$$\text{ACCESS URL} \Rightarrow \text{req}_n(\text{req. res}) = \text{req}_{n+1}(\text{ref. res}) \quad \text{Eq. III.2}$$

$$\text{BRUIT INTERFERENT} \Rightarrow \text{req}_n(\text{ref. res. typ}) \in \text{SQ} \quad \text{Eq. III.3}$$

L'inférence et le rajout automatiques du DN du site Web dans l'attribut vide (req.res) amènent la règle d'accès URN à la règle d'accès à l'URL. L'inférence automatique du DN donnée en **Eq. III.4** est expliquée dans ce qui suit.

$$\text{DN} = \arg. \max_{\text{freq}} (\text{REQ}(\text{ref. roo})) \quad \text{Eq. III.4}$$

L'entrée REF d'un ALF porte sur des ressources référentes (*REF.RES*) internes et externes (URI). Comme illustré dans la **Table III.3**, un clic est référé par un référent externe et autant de référents internes que le nombre de composants de la page pointés. Étant donné que les ressources référentes représentant des URI internes se composent de la même racine (*DN du site Web*) ; et que celles représentant des URI externes portent des racines différentes ; le DN du site Web est la racine la plus fréquente dans l'entrée REF de l'ALF. Ainsi, le DN du site Web est la racine la plus fréquente des ressources référentes.

L'inférence automatique du bruit interférant représenté par les requêtes de style/média (SQ) donnée en **Eq. III.5** est expliquée dans ce qui suit.

Considérant que les requêtes de style sont le composant unique de page représentant des REQ.RES.TYP de type SQ susceptible d'apparaître dans l'entrée REF du ALF ; il s'agit du type de ressource le moins fréquent de l'intersection des deux ensembles REQ.RES.TYP et REF.RES.TYP.

$$SQ = \arg. \min_{\text{freq}} (\text{REQ.RES.TYP} \cap \text{REF.RES.TYP}) \quad \text{Eq. III.5}$$

Les **Eq. III.4 & 5** représentent des propriétés statistiques de la structure de journalisation d'un ALF–ECLF. Elles ont été testées et validées sur 6 ALF qui serviront, plus loin, dans la section expérimentation du présent chapitre. Nous supposons donc qu'ils sont généralisables pour tout ALF–ECLF.

Ainsi, après l'inférence automatique du DN et du SQ, nous n'avons que deux règles, i.e., accès/clics et bruit. Ces deux règles sont représentées par les **Eq. III-6 et 7**.

$$\text{ACCESS} \Rightarrow \text{req}_n(\text{req.res}) = \text{req}_{n+1}(\text{ref.res}) \quad \text{Eq. III.6}$$

$$\text{BRUIT} \Rightarrow \text{req}_n(\text{ref.res.typ}) \in \text{SQ} \quad \text{Eq. III.7}$$

3) Implémentation

L'**algorithme III.3** que nous avons conçu présente l'implémentation du nettoyage basé sur les règles, i.e., accès et bruit. L'algorithme commence par analyser la R. WLDB pour, d'abord, inférer le DN du site Web de l'ALF et l'éditer quand l'instance de l'attribut REQ.RES est vide alors que les occurrences des attributs de méthode de requête (*REQ.MET*) et le protocole de requête (*REQ.PRO*) sont renseignés par la journalisation. Ensuite, il en déduit REQ.RES.TYP de type SQ susceptible d'apparaître dans l'entrée REF du ALF. Une fois que le type SQ est déduit, les requêtes dont les ressources référentes (*REF.RES*) contiennent le type de ressource SQ sont supprimées.

Algorithme III-3 Nettoyage basé sur les règles de la structure

```

01  INPUT DATA
02  R. WLDB
03  PROCESS
04  [reqn(req.met & req.pro = ∅)] ← arg. maxfreq (REQ (ref.roo))]
05  SQ=arg. minfreq (REQ.RES.TYP ∩ REF.RES.TYP)
06  WHILE reqn(ref.res.typ = SQ) FILTER-OUT reqn WEND
07      IF reqn(req.res) = reqn+1(ref.res) FILTER-IN reqn
08  OUTPUT DATA
09  C.WLDB

```

Ainsi, le filtre sur la base de la règle d'accès peut être appliqué pour retenir les requêtes remplissant la condition de la règle d'accès ; qui sont à considérer comme des clics à renvoyer à la C. WLDB.

4) Données et paramètres expérimentaux

L'algorithme de nettoyage basé sur des règles ainsi que les algorithmes de nettoyage Conventionnel et Avancé ont été testés sur un échantillon représentatif de 6 ALF de sites Web tiers. La description des données expérimentales est donnée dans la **Table III.4**. Ce panel expérimental a pour objectif d'être représentatif des différentes pratiques en termes de conception et contenu Web, et de ratio clics /hits.

Les ALF concernés concernent : une administration publique (*AL1.GOV*), des sites Web commerciaux (*AL2.COM*, *AL3.COM*, *AL4.COM*, *AL5.COM*) et le site Web de notre laboratoire (*AL6. LAB*). La taille des ALF est donnée en nombre de requêtes. L'expérimentation consiste à nettoyer ces ALF avec les **algorithmes 1,2 et 3** des méthodes de nettoyage en question, implémentés sous R via l'API Apache Spark.

Table III-4 Données expérimentales

ALF	DN du Site Web	Source de l'ALF	Volume	Clics	Hits	Ratio
AL1.GOV	www.khanyounis.mun.ps	https://www.mosa.gov.ps/khanyounis.mun.ps/log/access.log	26 168	2 564	23 604	1/9
AL2.COM	almhuetten-raith.at	http://www.almhuetten-raith.at/apache-log/access.log	31 433	13 108	18 325	1/2
AL3.COM	www.facades.fr	http://igm.univ-mlv.fr/~cherrier/download/L1/access.log	5 945	987	4 958	1/5
AL4.COM	www.boring-log.com	https://www.scisoftware.com/boring-log.com/logs/apache-access.log	17 676	3 164	14 512	1/4
AL5.COM	www.megapeloteros.com	http://salablanda.com.ar/megapeloteros.com.access.log	680	48	632	1/13
AL6.LAB	www.ai.univ-paris8.fr/	Site Web LIASD	145 227	25 391	119 836	1/6

Les sorties de l'application des méthodes de nettoyage sont évaluées selon les termes d'une matrice [76] de confusion où :

- La taille (*TAI*) fait référence à la taille, de l'ALF traité, exprimée en nombre de requêtes ;
- Positif (*P*) et Négatif (*N*), respectivement, indiquent le nombre de clics utilisateurs

finaux et hits utilisateurs agents contenus dans l'ALF ;

- Vrai Positif (TP), Faux Positif (FP), Vrai Négatif (TN) et Faux Négatif (FN) indiquent, respectivement, les requêtes valides et non valides, filtrées comme clics ou hits par l'algorithme de nettoyage ;
- Le taux de faux positif ($FPR = \frac{FP}{FP+TN}$) mesure le taux de bruit ;
- Le taux de faux négatif ($FNR = \frac{FN}{FN+TP}$) mesure le taux de silence ;
- La précision ($ACC = \frac{TP+TN}{P+N}$) mesure la pertinence du nettoyage.

La référence de validation (*label/nombre de P et N*) est composée de ressources valides pouvant être demandées par un utilisateur final sur les sites Web des FAL concernés. Cette référence de validation est définie par un logiciel de contrôle de site Web (*Xenu's Link Sleuth 1.3.8*) qui rapporte, entre autres, les URI incorporant des ressources cliquables destinées aux utilisateurs finaux.

Notez que le nombre des requêtes labélisées, par le logiciel, comme clics ou hits, peut changer en fonction du temps écoulé entre la date du fichier ALF et l'extraction des URI, car le contenu et la structure Web peuvent changer. Les tests de méthodes doivent donc être effectués sur la base de référence de validation obtenue à partir de la même extraction d'URI.

5) Evaluation et discussion des résultats

Les résultats présentés dans la **Table III.5** démontre que le R.CLE est la méthode la plus avantageuse en termes de pertinence, rapportée aux contraintes d'applicabilité et de coûts. Elle fournit des sorties de haute qualité avec un seul facteur de coût (*le processus de nettoyage*). Elle n'a pas besoin de données extra-journalisation, ce qui la rend insensible aux répercussions du Web dynamiques/Adaptatif sur le contenu de la journalisation. De plus, puisqu'elle ne nécessite aucune connaissance préalable, elle est utilisable tant pour les analyses des propriétaires de site Web que pour et ceux des serveurs.

L'analyse du taux de bruit (FPR) de l'algorithme du Nettoyage Conventionnel (*CON.CLE*) montre que 53% se compose d'images sans extension de type de fichier, en plus d'autre extensions de type de fichier inattendues de Web adaptatif. 47% du taux de silence concerne différents formats de contenu multimédia cliquable (*ressources non*

html) demandés par les utilisateurs finaux. Les résultats en termes de FPR du CON.CLE confirment son inadéquation pour le Web adaptatif. Les résultats en termes de FNR démontrent le défi de l'exhaustivité en termes de FLT.KDB et de connaissance apriori. Globalement, le CON.CLE est très bruyant (ACC 88%).

Table III-5 Résultats des nettoyages

ALF	TAI	P	N	TP	FP	TN	FN	FPR	FNR	ACC
Nettoyage Avancé (ADV.CLE)										
AL1.GOV	26 168	2 564	23 604	2 564	0	23 604	0	0%	0%	100%
AL2.COM	31 433	13 108	18 325	13 108	0	18 325	0	0%	0%	100%
AL3.COM	5 945	987	4 958	987	0	4 958	0	0%	0%	100%
AL4.COM	17 676	3 164	14 512	3 164	0	14 512	0	0%	0%	100%
AL5.COM	680	48	632	48	0	632	0	0%	0%	100%
AL6.LAB	145 227	25 391	119 836	25 391	0	119 836	0	0%	0%	100%
SUM/AVG	227 129	45 262	181 867	45 262	0	181 867	0	0%	0%	100%
Nettoyage basé sur les règles (R.CLE)										
AL1.GOV	26 168	2 564	23 604	2 564	0	23 604	0	0%	0%	100%
AL2.COM	31 433	13 108	18 325	13 108	0	18 325	0	0%	0%	100%
AL3.COM	5 945	987	4 958	987	0	4 958	0	0%	0%	100%
AL4.COM	17 676	3 164	14 512	3 164	0	14 512	0	0%	0%	100%
AL5.COM	680	48	632	48	0	632	0	0%	0%	100%
AL6.LAB	145 227	25 391	119 836	25 391	0	119 836	0	0%	0%	100%
SUM/AVG	227 129	45 262	181 867	45 262	0	181 867	0	0%	0%	100%
Nettoyage Conventionnel (CON.CLE)										
AL1.GOV	26 168	2 564	23 604	2 252	312	22 947	657	1%	23%	96%
AL2.COM	31 433	13 108	18 325	12 725	383	6 474	11 851	6%	48%	61%
AL3.COM	5 945	987	4 958	795	192	4 852	106	4%	12%	95%
AL4.COM	17 676	3 164	14 512	3 040	124	11 715	2 797	1%	48%	83%
AL5.COM	680	48	632	43	5	627	5	1%	10%	99%
AL6.LAB	145 227	25 391	119 836	20 270	687	119 149	5121	1%	20%	96%
SUM/AVG	227 129	45 262	181 867	39 125	1 703	165 764	20 537	2%	27%	88%

Les résultats du nettoyage sont illustrés dans la Table III.5. Notre méthode a réalisé son objectif, en l'occurrence un résultat de nettoyage sans bruit (ACC 100%). Notre méthode obtient le même résultat, sans bruit, que celui de la méthode ADV.CLE car les deux partagent un attribut de nettoyage commun (REF.RES). L'ADV.CLE basé sur la topologie réelle du site Web cherche l'information sur cet attribut par l'extraction des URIs du site concerné par la journalisation pour l'utiliser comme filtre de nettoyage. Par contre, notre méthode n'a pas besoin de la topologie du site, car elle exploite seulement l'information disponible sur la structure de journalisation des occurrences de cet attribut existant déjà dans les données de journalisation. Notre méthode obtient le même résultat que l'ADV.CLE d'une manière optimisée, à savoir sans besoin d'information hors données de journalisation ou de coûts d'extraction de cette information d'une source externe.

Les tests du Nettoyage Avancé (ADV.CLE) et celui basé sur les règles (R.CLE) donnent des taux sans bruit (FPR) ni silence (FNR). Cela est dû au fait qu'ils reposent sur

les mêmes attributs de filtrage (*REQ.RES*) mais extraits de deux sources différentes. Le R.CLE l'inclut (*REQ.RES*) dans la règle de filtrage (*Règle d'accès – Eq.III.6*) qui est basée les URI contenu dans l'entrée référente (*REF*) de l'ALF. Cependant, L'ADV.CLE le (*REQ.RES*) récupère à partir des URI du site Web qu'il extrait dans leur totalité. Puisque les URI de sites Web extraits contiennent toutes les ressources pouvant être demandées par clic, l'ADV.CLE fournit des sorties sans bruit. Par contre, le R.CLE atteint le même résultat de manière optimisée, car il prend en compte uniquement les URI des ressources demandées par les clics qu'il identifie via la règle d'accès (*Règle d'accès – Eq.III.6*).

Le R.CLE est la méthode la plus avantageuse en termes de pertinence, rapportée aux contraintes d'applicabilité et de coûts. Elle fournit des sorties de haute qualité avec un seul facteur de coût (*le processus de nettoyage*). Elle n'a pas besoin de données extra-journalisation, ce qui la rend insensible aux répercussions du Web dynamiques/Adaptatif sur le contenu de la journalisation. De plus, puisqu'elle ne nécessite aucune connaissance préalable, elle est utilisable tant pour les analyses des propriétaires de site Web que pour et ceux des serveurs.

L'ADV.CLE, qui donne également des résultats sans bruit, est très couteux pour les serveurs en termes de bande passante. Aussi, il s'avère irréalisable en cas de changement automatique et continu du contenu et de la structure du site (*Web Dynamic*), notamment en cas de contenu personnalisé [24], [25], [27], [28].

Le CON.CLE est très bruyante et s'avère obsolète dans le cas du Web Adaptatif, basé sur des frames sans extension de type de fichiers. L'ADV.CLE et le CON.CLE ont besoin de connaissances exhaustives et préalables sur le site Web de l'ALF à analyser, ce qui les limite aux analyses des propriétaire du site Web. Globalement, l'ADV.CLE et le R.CLE peuvent être combinés, où le premier peut être optimisé par le dernier, car ils partagent un attribut de filtrage commun provenant de deux sources différentes.

III.3.2. Contribution 1.2 : Méthode de nettoyage basée sur le partitionnement génétique

1) Concept et approche

Le contexte du Web Dynamic/Adaptatif affecte les méthodes de nettoyage, Conventionnel et Avancé, en termes de pertinence, de contraintes d'applicabilité, et de

coût, de sorte qu'elles deviennent obsolètes dans la perspective analytique des propriétaires de serveurs. Les facteurs limitants de ces méthodes sont dus à leurs approches centrées sur le contenu. Ces méthodes sont basées sur des heuristiques de filtrage reposant sur le contenu de l'ALF. Ainsi, ces méthodes, centrées sur le contenu, sont inadaptées au Web Dynamique/Adaptatif qui génère un contenu/structure changeant tant au niveau frontal qu'au niveau de la journalisation.

Différemment, notre contribution traite le problème de nettoyage du point de vue de la fouille des clics d'utilisateur final (*CLICKS*) parmi les hits agent-utilisateurs (*HITS*) sur la base des caractéristiques de la structure de journalisation. La présente contribution calque la structure de journalisation des *CLICKS* et *HITS* sur les caractéristiques de reproduction génétique afin de les différencier sur la base des concepts de partitionnement génétique.

Étant donné que la structure de journalisation dépend uniquement du format de l'ALF, la méthode proposée est insensible aux répercussions du Web Dynamique/Adaptatif sur le contenu de la journalisation. Cette méthode est destinée à surmonter les facteurs du Web Dynamiques/Adaptatif et leur impact sur la pertinence du nettoyage Conventionnel et la fiabilité du nettoyage Avancé et celui basé sur les Règles dans le contexte de Données Massives, à savoir le cas de la perspective analytique Propriétaire de Serveurs.

2) Concepts de Partitionnement Génétique

La **Figure III.1** présente l'objet et les caractéristiques de la génétique tels qu'ils sont présentés dans la discipline de la Génétique des Populations et les applications de fouille de donnée et extraction de connaissances (*KDD*) sous-jacentes [77]–[79], [79]–[86]. Une base de données génétiques (*Genetic Database – GDB*) représente une population de chromosomes décrits par des allèles (*occurrences/instances*) présents dans des loci (*attributs/variables*). La population initiale est constituée de chromosomes dont les loci comprennent des allèles appartenant à un ensemble d'allèles, appelés pool d'allèles.

Les populations sont décrites en termes de fréquence d'allèles, de séquences d'allèles et/ou de distance des loci. L'évolution de la population parentale à celle de la progéniture dépend d'opérateurs de reproduction et de facteurs de contexte, i.e.,

recombinaison (*croisement*), mutation, migration, sélection et valeur adaptative (*fitness/apptitude*).

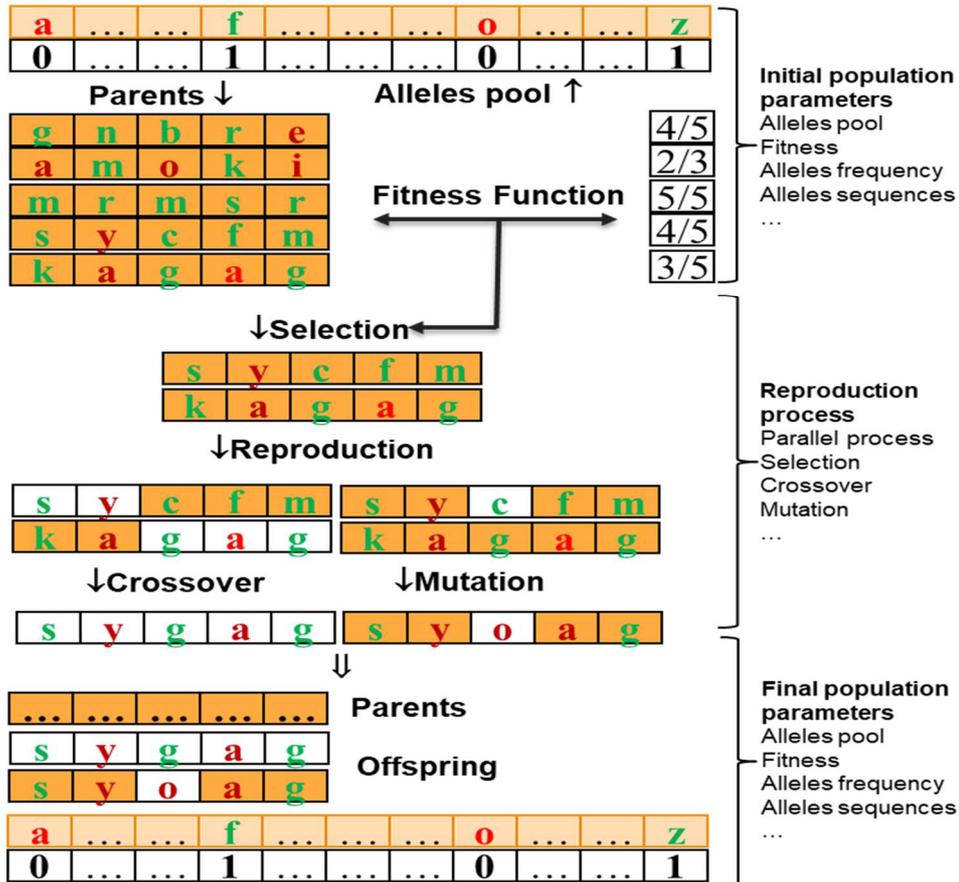


Figure III-1 Objet et caractéristique génétiques

Comme le montre la **Figure III.1**, il est à noter que dans le cas de la sélection naturelle, les chromosomes avec les allèles les plus aptes (*à forte valeur adaptative*), e.g., les consonnes majuscules, sont susceptibles de survivre et de se reproduire par croisement et mutation. Cependant, dans des cas de sélection aléatoire, tous les chromosomes ont la même probabilité de se reproduire. La différenciation des chromosomes, en termes de leurs allèles descripteurs, est basée sur les caractéristiques de croisement et de mutation, dans le cas d'une sélection aléatoire.

La recombinaison génère des chromosomes en croisant les allèles de locus des chromosomes parents. La mutation se produit lorsque le chromosome de progéniture généré prend un ou plusieurs nouveaux allèles, parmi ceux du pool allélique, qui

n'appartiennent pas à ses chromosomes-parents. L'évolution de la population sous de tels opérateurs et la sélection aléatoire n'affectent pas la fréquence allélique initiale au niveau de la population générée.

Cependant, une dérive peut affecter la fréquence allélique d'une génération à une autre pour deux raisons :

- Dans un contexte contraignant et sous l'effet de la sélection naturelle, seuls les chromosomes les plus aptes survivent pour se reproduire et conserver leurs paramètres alléliques sous-jacents (*modalité, fréquence absolue, croissance de la fréquence relative*) au sein de la population globale ;
- La disparition, la migration ou l'arrivée de nouveaux chromosomes affecte la population en termes de paramètres alléliques (*modalité, fréquence absolue, fréquence relative*).

Aux termes de la statistique, les concepts loci, allèle et pool allélique représentent des variables catégorielles (*loci*) aux modalités (*allèles*) interchangeable appartenant à la même classe (*pool allèles*). [77]–[79].

La différenciation entre les chromosomes est basée sur la distance génétique à deux niveaux, i.e., la fréquence allélique et la séquence et/ou le niveau physique.

- La distance génétique allélique basée sur la fréquence (*Freq.GDist*) est l'application d'une distance canonique à la matrice de fréquence des allèles en termes de loci utile de partitionnement. Cette perspective ne prend pas en compte les facteurs de dérive. Elle cartographie les chromosomes liés par croisement ou mutation en partitions distinctes.
- Des distances génétiques dédiées qui mesurent une distance génétique basée sur les séquences de loci/allèles (*Seq.GDist*) ou une distance physique de loci (*Pyh.GDist*) sont utilisées pour prendre en considération les facteurs de dérive et l'analyse des expressions géniques.

Globalement, le choix de la méthode de partitionnement appropriée et les loci utiles sous-jacents dépendent principalement des opérateurs et facteurs de reproduction considérés, ainsi que des objectifs de l'analyse [77]–[79].

Pour notre approche de nettoyage basée sur les concepts de partitionnement génétique, la perspective qui nous intéresse est la différenciation entre chromosomes sur la base de la Freq.GDist, qui distingue des groupes de chromosomes liés par un croisement ou une mutation.

3) Nettoyage basé sur les concepts de partitionnement génétique

La **Table III.6** décrit les caractéristiques de la structure de journalisation (*motifs de la structure de journalisation – Motifs*) sur la base du tutoriel du serveur Apache et d'un test de simulation de navigation effectué par un logiciel dédié (*Webserver Stress Tool 8.0.0.1010*) sur une copie d'un site Web (*Simple English Wikipedia*) que nous avons dupliqué sous serveur local Apache. Les caractéristiques génétiques d'intérêt de la structure de journalisation sont les suivantes : CLICS, HITS sous-jacents et Ratio CLICKS / HITS.

Table III-6 Motifs Génétiques de la structure de journalisation

Utilisateur	Requête	REQ.RES	REF.RES
End-user x	Click 1	PAGE B URI	PAGE A URI
User-agent x	<i>Hit 1.1</i>	<i>PAGE B Component 1</i>	<i>PAGE B URI</i>
	<i>Hit 1.2</i>	<i>PAGE B Component 2</i>	<i>PAGE B URI</i>
	<i>PAGE B URI</i>
	<i>Hit 1. n</i>	<i>PAGE B Component n</i>	<i>PAGE B URI</i>
End-user x	Click 2	PAGE C URI	PAGE B URI
User-agent x	<i>Hit 2.1</i>	<i>PAGE C Component 1</i>	<i>PAGE C URI</i>
	<i>Hit 2.2</i>	<i>PAGE C Component 2</i>	<i>PAGE C URI</i>
	<i>PAGE C URI</i>
	<i>Hit 2. n</i>	<i>PAGE C Component n</i>	<i>PAGE C URI</i>
End-user x	Click 3	PAGE D URI	PAGE C URI
User-agent x	<i>Hit 3.1</i>	<i>PAGE D Component 1</i>	<i>PAGE D URI</i>
	<i>Hit 3.2</i>	<i>PAGE D Component 2</i>	<i>PAGE D URI</i>
	<i>PAGE D URI</i>
	<i>Hit 3. n</i>	<i>PAGE D Component n</i>	<i>PAGE D URI</i>
Known Robot-agents	Hit 1	Robot dedicated resource	NA
	NA
	Hit n	Robot dedicated resource	NA
Unknown Robot-agents	Hit 1	PAGE URI or/and Component	...

	Hit n	PAGE URI or/and Component	...

Les motifs de structure de journalisation illustrés dans la *Table III.6* montrent le rapport de croisement (couleur verte) entre les clics et celui de mutation (couleur orange) entre les hits. C'est cette analogie de motifs de structuration génétique qui servira notre concept de nettoyage basé sur les techniques de partitionnement génétique.

Aux termes de la statistique, les attributs REQ.RES et REF.RES et les occurrences sous-jacentes représentent des variables catégorielles (*attributs/descripteurs*) avec des modalités interchangeables (*occurrences/instances*) dont les valeurs appartiennent à la même classe (*URI et Composants*). Les Motifs structurels de journalisation ci-dessous illustrent la journalisation des requêtes en termes des attributs « REQ.RES » et « REF.RES ». Les clics des utilisateurs finaux et les hits des agents-utilisateurs associés sont, respectivement, liés à l'URI de la page Web et aux composants d'affichage associés (*contenu multimédia, média accessoire, etc.*).

Les Motifs consistent en des paires d'entrées où les URI des pages et leurs composants (*REQ.RES*) sont référés (*REF.RES*) par les ressources à partir desquelles ces pages ont été demandées (*REQ.RES*), respectivement. Ainsi, l'URI de la page en cours est référé par l'URI de la *page demandée précédemment* ; puisqu'*elle* a été pointée à partir de *celui-ci*. Les composants de la page en cours sont référés par l'URI de la *page en cours* ; car elles ont été pointées à partir de *celle-ci*.

Les clics sur les URI des pages, considérées en dehors des HITS sous-jacents, concernent des paires d'entrées d'attributs REQ.RES et REF.RES avec des modalités (*occurrences/instances*) interchangeables. Les attributs REQ.RES, REF.RES, leurs modalités et les requêtes impliquées représentent, respectivement, des loci et des allèles croisés de chromosomes. Ce motif de structure représente un croisement/recombinaison génétique.

Ainsi, ce croisement peut être exprimé comme une caractéristique génétique de la structure de journalisation comme suit :

$$(\mathbf{req}_n(\mathbf{req.res}) \cap \mathbf{req}_{n+x}(\mathbf{ref.res}) \neq \emptyset)$$

&

Eq. III.8

$$(\mathbf{req}_n(\mathbf{req.res}) \cap \mathbf{req}_{n+x}(\mathbf{req.res}) = \emptyset)$$

La structure de journalisation des hits pointant les composants des pages, considérée en dehors des clics sous-jacents, comprend des paires d'attributs REQ.RES et

REF.RES avec les mêmes modalités pour l'attributs REF.RES et des modalités changeantes pour l'attributs REQ.RES. L'attribut REQ.RES et ses modalités (*occurrences/instances*) changeantes représentent, respectivement, des loci et des allèles mutants de chromosomes. Ce motif de structure représente une mutation génétique.

Ainsi, cette mutation peut être exprimée comme une caractéristique génétique de la structure de journalisation comme suit :

$$\begin{aligned}
 & (\mathbf{req}_n(\mathbf{req.res}) \cap \mathbf{req}_{n+x}(\mathbf{req.res}) = \emptyset) \\
 & \qquad \qquad \qquad \& \qquad \qquad \qquad \text{Eq. III.9} \\
 & (\mathbf{req}_n(\mathbf{ref.res}) \cap \mathbf{req}_{n+x}(\mathbf{ref.res}) \neq \emptyset)
 \end{aligned}$$

Pour ce qui est de l'activité/hits des robots, il est à noter ce qui suit :

- Les robots conventionnels tels ceux d'indexation (*respectant les standards en la matière*) naviguent directement les liens des sites Web à partir de listes d'URI. Il s'agit là d'une navigation directe sans référent (*REF.RES*). Cependant, les robots malveillants (*ou ne respectant pas les standards de l'activité d'indexation*) ont un comportement imprévu qui peut prendre différentes formes, telles que la simulation de la navigation d'un utilisateur final [24], [29]–[31], [45] ;
- Le premier cas est un cas banal où les requêtes dont l'occurrence de l'attribut REF.RES est vide (*valeurs manquantes – NA*) ne seront pas prises en compte ;
- Le deuxième cas est un problème de détection de robot basé sur nettoyage intelligent à effectuer en aval de la structuration des données. Ce problème ne rentre pas dans le cadre de notre recherche.

Globalement, la structure de la R. WLDB présente les mêmes caractéristiques statistiques et génétiques de croisement et de mutation en termes des attributs de nettoyage, i.e., REQ.RES et REF.RES. Ces attributs représentent des facteurs avec des modalités (*occurrences*) interchangeable et/ou communes appartenant à la même variable catégorielle. A ce titre, les caractéristiques génétiques d'intérêt qui serviront notre approche sont développées ci-dessous.

- Le rapport CLICK/HITS est de Singulier/Multiple, soit « $1:n, n > 1$ », et varie en fonction du contenu de la page Web et de la nature du site. Le ratio de « $1/10$ », souvent cité dans la littérature académique [24]–[27], est adopté à titre représentatif à cet égard dans le cadre de notre contribution.
- Les caractéristiques génétiques de la structure de journalisation en termes de CLICKS et de HITS se limitent, respectivement, au croisement et à la mutation. Les attributs qui présentent des caractéristiques génétiques sont : REQ.RES et REF.RES. Ils sont donc les attributs qui serviront le partitionnement destiné à distinguer les clics et les hits.
- La mesure appropriée de différenciation entre CLICKS et HITS en termes d'attributs de partitionnement est la Freq.Gdist, car elle est la distance génétique appropriée lorsque les caractéristiques génétiques sont limitées au croisement et à la mutation. Freq.Gdist est censé différencier, par deux partitions distinctes et exclusives, les CLICKS en tant que croisement et les HITS en tant que mutation.

Étant donné que le nombre de partitions ciblées est limité à deux partitions exclusives, et en raison de l'aspect Big Data de la R. WLDB ; l'algorithme CLARA (*Clustering Large Application*) basé sur la variante de partitionnement autour des Médoïdes (*Partitioning Around Medoids – M*) et la distance de Manhattan (*Man.Dist*) semble être la méthode de partitionnement la plus appropriée pour trois raisons :

- Le partitionnement (*PC*) est approprié pour le cas de nombre limité et connu de partitions ;
- Le partitionnement autour des Médoïdes (*PAM*) et basée sur la distance de Manhattan (*Man.Dist*) constituent une méthode insensible aux valeurs aberrantes ;
- L'algorithme CLARA est optimisé pour traiter des données massives.

Étant donné que le rapport CLICK/HITS est Singulier/Multiple, soit « $1:n, n > 1$ », les partitions de petites et grandes tailles sont, respectivement, supposées être celles des CLICKS et HITS. Dans le cas de sites Web dont toutes les pages ne contiennent qu'une seule composante où le ratio CLICK / HITS est Singulier/Singulier, soit « $1:n, n = 1$ », les partitions présentant des caractéristiques de croisement ou de mutation sont, respectivement, supposées être celles des CLICKS ou HITS. Ceci peut être vérifié sur la base de l'expression des caractéristiques sous-jacentes indiquée ci-dessus

(Eq. III.8 & 9).

4) Méthode d'application

- **Données d'entrée (R. WLDB)**

Etant donnée une R. WLDB de CLICKS et HITS enregistrée dans une ALF comme suit :

```
[127.0.0.1] [-] [frank] [10/Oct/2000:13:55:36 -0700] [GET/apache_pb.gif HTTP/1.0]
[200] [2326] [http://www.example.com/start.html] [Mozilla/4.08 (Win*; ...;Bro*)]
```

- **Préparation des données (Clustering Attributes)**

Les attributs d'intérêt pour le partitionnement sont :

- REQ.RES soit [... apache.gif ...] ;
- REF.RES soit [... start.html ...].

- **Transformation des données (Matrice de fréquence relative)**

Etant donnée (X) de (n) éléments x_n . La fréquence relative (f) d'un élément x_n de (X) est la fréquence absolue ($F_x = x_n \Sigma n$) rapportée à la taille exprimée par le nombre d'éléments de l'échantillon en question ($X_n \Sigma n$). La matrice de fréquence relative se calcule au niveau cellule de la matrice des occurrences des attributs REQ.RES et REF.RES puisqu'ils appartiennent à la même classe.

$$f_x = \frac{x_n \Sigma n}{X_n \Sigma n} \quad \text{Eq. III.10}$$

- **Distance de partitionnement (Différenciation/Dissimilitude)**

La distance Man. Dist entre deux points (A, B) se considère dans un espace quadratique. Elle est la somme de la distance linéaire absolue entre leurs coordonnées cartésiennes (X, Y), respectifs.

$$\text{Man. Dist}(A, B) = |X_B - X_A| + |Y_B - Y_A| \quad \text{Eq. III.11}$$

- **Méthode de partitionnement (Agrégation/Similitude)**

L'algorithme CLARA est configuré pour générer deux partitions exclusives qui agrègent les requêtes sur la base du point médoïde (*MP*) des valeurs calculées de leurs distances. Dans un espace (*k*) de points, le *MP* est le point (*m*) parmi les points existant avec la distance (*d*) la plus proche ($arg. \min_m$) du centre de tous les points (*m*).

$$m_{MP} = arg. \min_m \sum_{i=1}^n d(i, m) \quad Eq. III.12$$

- **Analyse et interprétation des sorties**

La plus petite partition en nombre de requêtes est supposée être la partition des CLICKS. Dans le cas de taille égale des deux partitions, celle contenant des requêtes avec des motifs de structure de journalisation exprimant un croisement ; est supposée être la partition des CLICKS. Cela peut être vérifié par l'équation sous-jacente (**Eq. III 8**).

5) Implémentation

L'algorithme **III.4** proposé pour le nettoyage, basé sur le partitionnement génétique, traite la R. WLDB sur la base des attributs de partitionnement (*CLUST.ATT*), i.e., REQ.RES et REF.RES.

La fonction de partitionnement (*CLUST.FUNC*) est chargée de calculer la transformation de données nécessaire, à savoir la matrice de fréquence relative (*RFM*) en tant que transformation objet du partitionnement (*TRAN=RFM*).

Ensuite, l'algorithme se chargera de générer deux partitions (*K. CLUST=2*) sur la base de la distance de Manhattan (*DIST=Man.Dist*).

Enfin, il compare la taille des deux partitions, exprimée en nombre de requêtes ($\sum n_{req.k1} < > \sum n_{req.k2}$), et ceci pour exclure de la R. WLDB, les requêtes appartenant à la partition de la plus grande taille exprimée en nombre de requêtes (*arg. max*) supposée être la partition des HITS. Ainsi, la R. WLDB débarrassée des HITS représente la C. WLDB.

Algorithme III-4 Nettoyage basé sur le Partitionnement Génétique

```

01 INPUTS
02   R. WLDB
03   CLUST.ATT ← {REQ.RES, REF.RES}
04   CLUST.FUNC ← (TRAN=RFM, CLUST.METH=CLARA, K. CLUST=2,
    DIST=Man.Dist)
05 PROCESS
06   APPLY CLUST.FUNC to R. WLDB[CLUST.ATT,]
07   Filter-out REQ ∈ arg.max( $\sum n_{req.k1} < > \sum n_{req.k2}$ )
08 OUTPUT DATA
09   C.WLDB

```

6) Données et paramètres expérimentaux

L’algorithme de nettoyage basé sur le partitionnement génétique (*GEN.CLE*) ainsi que les algorithmes de nettoyage conventionnel (*CON.CLE*) et celui avancé (*ADV.CLE*) ont été testés sur un échantillon représentatif de 6 ALF. La description des données expérimentales est donnée dans la **Table III.7**. Il s’agit d’un panel représentatif des différentes pratiques en termes de conception Web, de contenu Web et de ratio CLICK/HITS.

Les ALF en question concernent une administration publique (*AL1.GOV*), des sites Web commerciaux (*AL2.COM*, *AL3.COM*, *AL4.COM*, *AL5.COM*) et le site Web de notre laboratoire (*AL6.LAB*). La taille des ALF est donnée en nombre de requêtes. Ils ont été traités avec les méthodes de nettoyage concernées par le biais de leurs **algorithmes** respectifs **III.1 (CON_CLE)**, **2 (ADV_CLE)** & **4 (Notre méthode GEN_CLE)**, que nous avons conçu et implémentés sous R dans Apache Spark API.

Table III-7 Descriptif des données expérimentales

ALF	DN du Site Web	Source de l’ALF	Taille	Clics	Hits	Ratio
ALF1.GOV	www.khanyounis.mun.ps	https://www.mosa.gov.ps/khanyounis.mun.ps/log/access.log	26 168	2 564	23 604	1/9
ALF2.COM	almhuetten-raith.at	http://www.almhuetten-raith.at/apache-log/access.log	31 433	13 108	18 325	1/2
ALF3.COM	www.facades.fr	http://igm.univ-mlv.fr/~cherrier/download/L1/access.log	5 945	987	4 958	1/5
ALF4.COM	www.boring-log.com	https://www.scisoftware.com/boring-log.com/logs/apache-access.log	17 676	3 164	14 512	1/4
ALF5.COM	www.megapeloteros.com	http://salablanda.com.ar/megapeloteros.com.access.log	680	48	632	1/13
ALF6.LAB		Site Web LIASD	61 288	8 266	52 962	1/7

Les sorties de l'application des méthodes de nettoyage sont évaluées selon les termes d'une matrice de confusion [76] où:

- La taille (*TAI*) fait référence à la taille, de l'ALF traité, exprimée en nombre de requêtes ;
- Positif (*P*) et Négatif (*N*), respectivement, indiquent le nombre de clics utilisateurs finaux et hits utilisateurs agents contenus dans l'ALF ;
- Vrai Positif (*TP*), Faux Positif (*FP*), Vrai Négatif (*TN*) et Faux Négatif (*FN*) indiquent, respectivement, les requêtes valides et non valides, filtrées comme clics ou hits par l'algorithme de nettoyage ;
- Le taux de faux positif ($FPR = \frac{FP}{FP+TN}$) mesure le taux de bruit ;
- Le taux de faux négatif ($FNR = \frac{FN}{FN+TP}$) mesure le taux de silence ;
- La précision ($ACC = \frac{TP+TN}{P+N}$) mesure la pertinence du nettoyage.

La référence de validation (label/nombre de P et N) est composée de ressources valides pouvant être demandées par un utilisateur final sur les sites Web des FAL concernés. Cette référence de validation est définie par un logiciel de contrôle de site Web (Xenu's Link Sleuth 1.3.8) qui rapporte, entre autres, les URI incorporant des ressources cliquables destinées aux utilisateurs finaux.

Notez que le nombre des requêtes labélisées, par le logiciel, comme clics ou hits, peut changer en fonction du temps écoulé entre la date du fichier ALF et l'extraction des URI, car le contenu et la structure Web peuvent changer. Les tests de méthodes doivent donc être effectués sur la base de référence de validation obtenue à partir de la même extraction d'URI.

7) Évaluation et discussion des résultats

Les résultats du nettoyage sont présentés dans la **Table III.8**. L'analyse du taux de bruit (*FPR*) de l'algorithme CON.CLE montre que 53% se compose d'images sans extension de type de fichier, en plus des extensions de type de fichier adaptatif inattendues. 47% du taux de silence concerne différents formats de contenu multimédia cliquable (ressources non html) demandés par les utilisateurs finaux.

Les résultats en termes de FPR du CON.CLE illustrent l'affirmation de son inadéquation pour le Web Adaptatif/Dynamique basé sur des frames sans extension de type de fichier. Les résultats en termes de FNR démontrent le défi de l'exhaustivité lié à la construction apriori d'une FLT.KDB exhaustive.

Globalement, le CON.CLE est très bruyant (*ACC 88%*). Le test du ADV.CLE donne des taux sans bruit (*FPR*) et sans silence (*FNR*). Cependant, l'ADV.CLE est saturant pour les serveurs Web, en particulier dans le contexte du Web dynamique. L'ADV.CLE est basé sur une VLD.KDB construite lors de l'extraction des URI du site Web. Ainsi, pour les sites Web à contenu et structure dynamiques, la mise en place d'une VLD.KDB à jour et synchronisée se transforme en un processus complexe, qui sature le serveur, et voire même impossible, le cas échéant d'un contenu personnalisé [24], [25], [27], [28]. En outre, il peut manquer les contenus créés/annulés automatiquement s'il n'est pas synchronisé en temps réel avec le processus de nettoyage.

Table III-8 Résultats du nettoyage par partitionnement génétique

ALF	Size	P	N	TP	FP	TN	FN	FPR	FNR	ACC
ADV.CLE										
AL1.GOV	26 168	2 564	23 604	2 564	0	23 604	0	0%	0%	100%
AL2.COM	31 433	13 108	18 325	13 108	0	18 325	0	0%	0%	100%
AL3.COM	5 945	987	4 958	987	0	4 958	0	0%	0%	100%
AL4.COM	17 676	3 164	14 512	3 164	0	14 512	0	0%	0%	100%
AL5.COM	680	48	632	48	0	632	0	0%	0%	100%
AL6.LAB	61 288	8 266	52 962	8 266	0	52 962	0	0%	0%	100%
SUM/AVG	143 190	28 137	114 993	28 137	0	114 993	0	0%	0%	100%
GEN.CLE										
AL1.GOV	26 168	2 564	23 604	2 564	0	23 604	0	0%	0%	100%
AL2.COM	31 433	13 108	18 325	13 108	0	18 325	0	0%	0%	100%
AL3.COM	5 945	987	4 958	910	77	4 958	0	8%	0%	99%
AL4.COM	17 676	3 164	14 512	2 740	424	14 512	0	13%	0%	98%
AL5.COM	680	48	632	48	0	632	0	0%	0%	100%
AL6.LAB	61 288	8 266	52 962	8 021	245	52 962	0	3%	0%	99%
SUM/AVG	143 190	28 137	114 993	27 391	746	114 993	0	4%	0%	99%
CON.CLE										
AL1.GOV	26 168	2 564	23 604	2 252	312	22 947	657	12%	18%	96%
AL2.COM	31 433	13 108	18 325	12 725	383	6 474	11 851	3%	65%	61%
AL3.COM	5 945	987	4 958	795	192	4 852	106	19%	2%	95%
AL4.COM	17 676	3 164	14 512	3 040	124	11 715	2 797	4%	19%	83%
AL5.COM	680	48	632	43	5	627	5	10%	1%	99%
AL6.LAB	61 288	8 266	52 962	3 987	4 279	52 688	274	51%	1%	92%
SUM/AVG	143 190	28 137	114 993	22 842	5 295	99 303	15 690	17%	18%	88%

La méthode GEN.CLE proposée offre une précision très proche du ADV.CLE. L'analyse de son FPR montre qu'il s'agit de requêtes qui portent sur les icônes de logos des sites Web expérimentés. Donc, il ne s'agit pas forcément de hits mal référencés par

notre algorithme GEN.CLE du moment que certains utilisateurs finaux peuvent avoir cliqué sur les logos des sites pour les télécharger, e.g., l'utilisation du logo d'un site dans un exposé.

Ainsi, notre algorithme peut avoir mal référencé ces requêtes pour deux raisons :

- Le seuil bas de leur fréquence relative est proche de la fréquence des CLICKS ;
- Le logiciel utilisé dans la construction de la référence de validation les a reconnus comme logos de sites Web (*favicon*) et les a labélisés comme HITS même s'ils ont été demandés par des utilisateurs finaux via CLICKS.

Notre méthode GEN.CLE est la plus avantageuse en termes de pertinence, rapportée aux contraintes d'applicabilité et de coûts. Elle fournit des sorties de haute qualité avec un seul facteur de coût (*le processus de nettoyage*). Elle n'a pas besoin de données extra-journalisation, ce qui la rend insensible aux répercussions du Web Dynamiques/Adaptatif sur le contenu de la journalisation. De plus, puisqu'elle ne nécessite aucune connaissance préalable, elle est applicable pour les deux perspectives d'analyse, i.e., propriétaire de site et/ou serveur Web.

L'ADV.CLE, qui offre également des sorties de haute qualité, est saturant pour les serveurs et s'avère inutilisable dans le cas de changement automatique continu du contenu et de la structure Web, et en particulier dans le cas de contenu Web personnalisé [24], [25], [27], [28].

Le CON.CLE est très bruyante et s'avère obsolète dans le cas du Web à base de frames. Les nettoyages avancé et conventionnel nécessitent une connaissance exhaustive et préalable du site Web concerné, ce qui les limite au point de vue analytique des propriétaires de site Web.

III.4. Limites et perspectives

Cette contribution présente une analyse critique du nettoyage des données de l'usage du Web dites données de journalisation serveur. Les méthodes analysées, telles que le nettoyage conventionnel et celui avancé, sont centrées sur le contenu et souffrent de plusieurs limitations dans le contexte du Web Dynamique/Adaptatif.

A cet égard, deux méthodes de nettoyage centrées sur la structure sont proposées pour surmonter les limites des méthodes analysées dans le contexte en question. Étant

donné que les méthodes proposées sont centrées sur la structure des données de journalisation, elles présentent des avantages significatifs par rapport aux méthodes de nettoyage centrées sur le contenu, et ceci en termes de pertinence, rapportée aux contraintes d'applicabilité et de coûts.

En plus des avantages démontrés à l'échelle expérimentale de nos deux méthodes, leur intérêt applicatif peut être perçu de différentes manières, à savoir :

- Leur capacité à apporter plus de précision aux étapes de fouille des données en aval du nettoyage, i.e., découverte de motifs d'usage et leur analyse dans un contexte de Web Dynamic/Adaptatif.
- Leur capacité à capturer, en plus des clics d'utilisateurs finaux sur les URI des pages Web, ceux sur les différents contenus d'une page qui sont supposés par les méthodes actuelles comme hits d'agents utilisateurs.
- Leur capacité à capturer les clics sur les URI et contenu créés/supprimés automatiquement sans avoir besoin d'une base de connaissance a priori sur les URI et les contenus d'un site Web et/ou une synchronisation en temps réel.
- Elles sont applicables tant pour les propriétaires de sites que ceux de serveurs Web.

Enfin, notez que la méthode heuristique est destinée au nettoyage des données d'usage pour les besoins d'une analyse en temps réel, tandis que celle basée sur les techniques de partitionnement génétique sera très utile pour l'analyse hors ligne/différée des données de journalisation de masse. Nos deux méthodes ont été testées à une échelle expérimentale sur un échantillon de fichiers de journalisation limité et nécessitent une validation à une échelle plus large.

Chapitre IV

Deuxième Contribution

Structuration des Données De l'Usage du Web

IV.1. Contexte, problème et contributions visées

Aux termes de notre deuxième contribution nous abordons la tâche de la structuration des données de journalisation Web dans le cadre du prétraitement des données au titre de la fouille des données de l'utilisation du Web. Nous mettons l'accent sur les limites des méthodes de structuration actuelles, orientées « Agents », en matière de pertinence dans le contexte de réticence à l'authentification et à l'acceptation des cookies, pour proposer une méthode orientée « flux de clics » plus pertinente.

Les données d'utilisation Web, appelées données Log, font référence aux requêtes des utilisateurs finaux et agents enregistrées par les serveurs Web dans des fichiers journaux (*Access Log File – ALF*). Les données Log reflètent l'activité d'utilisation du Web et sont exploitées à des fins différentes, e.g., l'optimisation du système ; recommandation et personnalisation ; publicité et sponsoring. La fouille des données de l'usage du Web (*Web Usage Mining – WUM*) étant intéressée par le comportement des utilisateurs finaux, les requêtes d'agent sont à identifier et à nettoyer, pour garder celles des utilisateurs finaux destinées à être structurées en utilisateurs et/ou sessions uniques/singulières avant l'identification et l'extraction des transactions utiles à l'analyse.[21], [30], [31], [35], [37]. Noter que la finalité de la structuration est, donc, l'identification (*construction/reconstruction*) de sessions de visites de l'utilisateur final. Le terme « *construction de session* » est utilisé dans le cas de méthodes proactives. Cependant, le terme « *reconstruction de session* » est plus approprié aux méthodes de structuration réactive [29].

Étant donné que les serveurs Web enregistrent les requêtes (*utilisateurs finaux et agents*) entrelacées dans un ordre séquentiel, indépendamment de leur source ou de leur type ; il n'est pas évident de les identifier par utilisateur unique dans le cas d'une réticence de l'utilisateur final face aux systèmes de structuration proactives, i.e., authentification, cookies, Java applets. Par conséquent, les méthodes de structuration réactive utilisant des méthodes heuristiques, basées sur les paramètres adresse IP, agent, topologie du site Web et le temps, souffrent de problèmes de qualité qui affectent la pertinence du processus de fouille en aval et la fiabilité des modèles découverts.

À cet égard, malgré la large utilisation des méthodes de structuration proactive, les méthodes réactives restent d'intérêt pour plusieurs raisons, à savoir :

- Ils capturent l'activité des utilisateurs non couverte par les systèmes de structuration proactive ;
- Ils ne soulèvent pas de problèmes de confidentialité.
- L'acceptation/imposition des systèmes proactives dépend de la position ou de l'utilité du site Web et de la perception des utilisateurs quant aux avantages de leur acceptation.

Dans ce contexte, notre contribution souligne d'abord les limites des méthodes de structuration réactive en termes de qualité. Ensuite, elle introduit une méthode de structuration centrée sur le flux pour améliorer la qualité de la construction. Les méthodes réactives actuelles, centrées sur l'agent, traitent, indifféremment, l'ensemble des requêtes, et structure les données sur la base des références manquantes ou de seuils empiriques de temps de visite identifiés par adresse IP et agent. Cependant, la méthode que nous proposons, et qui est centrée sur les flux, contraint le processus de structuration par une fonction-objectif, de sorte que la structuration repose principalement sur l'identification de flux de clics pertinents avant de les assigner par agent et adresses IP.

Notre méthode a été testée sur un échantillon de fichier Log et a été évaluée sur la base d'un ensemble de critères proposés dans la littérature correspondante. Les résultats démontrent son efficacité en matière de structuration en termes de qualité de sessions construites.

IV.2. Analyse des méthodes de structuration – Travaux connexes

IV.2.1. Concepts et problèmes de la structuration

1) Concepts

Un exemple représentatif d'ALF tiré d'un tutoriel Apache Server est présenté dans la Figure IV.1. Les entrées du fichier Log notées entre crochets peuvent être formatées en attributs utiles en fonction des objectifs de l'analyse.

La **Figure IV.1** représente un utilisateur final anonyme (entrée/attribut authentification et cookie vides [-]) qui demande/clic une page/ressource (*REQ.RES*) [.../apache.org/example.html/...] à partir d'une page/ressource précédente/référente (*REF.RES*) [http://apache.org]. L'agent utilisé/navigateur (*User-agent – USE.AGE*) [Mozilla/4.08 (Win*; Chrome*)] charge/pointe les composants de l'aperçu/affichage de la page [...apache_pb.gif...] et l'affiche avec les spécifications/frame appropriées

[...style.css...]. Les requêtes et scripts de d'administration du site Web ainsi que celles des robots d'indexation et autres agents sont enregistrées dans le même fichier journal/Log dans le même format.

```
[127.0.0.1],      [-],      [-],      [10/Oct/2000:13:55:36      -0700],
[GET/apache.org/example.htmlHTTP/1.0],[200], [2326],      [http://apache.org],
[Mozilla/4.08 (Win*; Chrome*)], [-]
```

```
[127.0.0.1],  [-], [-], [10/Oct/2000:13:55:36 -0700],[GET/style.css      HTTP/1.0],
[200], [2326],[http://apache.org/example.html], [Mozilla/4.08 (Win*; Chrome*)], [-]
```

```
[127.0.0.1], [-],[-], [10/Oct/2000:13:55:36 -0700], [GET/apache_pb.gif HTTP/1.0],
[200], [2326],[http://apache.org/example.html], [Mozilla/4.08 (Win*; Chrome*)], [-]
```

Figure IV-1 Echantillon de fichier Log

Étant donné que la fouille des données d'usage Web s'intéresse au comportement des utilisateurs finaux, les requêtes d'agent sont référées comme bruit à nettoyer avant la fouille. Seuls les URI sur lesquels les utilisateurs finaux ont cliqué, tels que les liens html et DN, sont retenus pour l'analyse. Ensuite, les requêtes doivent être structurées en utilisateurs uniques/singulier puis en sessions serveurs/visites. Une session serveur est la figure/donnée de base servant à l'identification des transactions extraites pour la découverte de modèles d'utilisation.[21], [27], [28].

La structuration des données de journalisation Web consiste en l'identification d'utilisateurs uniques/singuliers, puis la reconstruction de session serveur, et enfin, l'identification de transaction servant la fouille et l'extraction des connaissances utiles pour l'optimisation, la personnalisation/recommandation, et le placement publicitaire [21], [34], [37]. L'identification des utilisateurs est destinée à regrouper les requêtes par utilisateur unique/singulier sur la base de certains attributs utiles/entrées ALF, i.e., IP, Login, User-Agent, etc. Lorsque les informations d'identification par authentification/cookie sont disponibles, des utilisateurs uniques/singuliers peuvent être identifiés. Sinon, seuls des utilisateurs uniques peuvent être identifiés sur la base d'heuristique centrée sur des paires IP-USE.AGE distinctes supposées représenter des requêtes émanant d'un seul utilisateur. Une session serveur consiste en une séquence de

pages consultées par un seul utilisateur au cours d'une seule visite sur le site Web. Une transaction ou un épisode est tout sous-ensemble significatif d'une session serveur qui représente un intérêt particulier pour la découverte de motifs d'usage. Ainsi, la session serveur/visite est la figure/information de base pour la fouille de l'utilisation du Web. [21].

2) Inconvénients

Les serveurs enregistrent les requêtes entrelacées dans un ordre séquentiel, indépendamment de leur source ou de leur type. Lorsque les informations d'authentification ou de cookie sont disponibles, la structuration des données de journalisation en utilisateurs uniques/singulier et sessions serveurs est simple. Les demandes sont triées par identifiant d'authentification et/ou cookies pour grouper les requêtes de chaque utilisateur. Selon certains seuils de temps mort (*timeout*) ou d'intervalle de navigation (*time gap*) de navigation, les termes d'une séquence de requêtes d'utilisateur unique/singulier sont divisés en sessions serveur qui représentent des visites uniques/singulières. L'identification de session basée sur de tels systèmes d'identification (*authentification/cookies*) est appelée méthode proactive et fournit des sessions presque réelles. Le terme proactif désigne la coopération de l'utilisateur (*authentification*) ou l'acceptation (*cookies*) à être identifiée. En cas de réticence des utilisateurs envers ces systèmes, il est recouru aux méthodes réactive pour l'identification d'utilisateurs individuels (*singulier*) via l'heuristique IP/Agent/Temps.

L'analyse des travaux connexes significatifs identifiés [21], [22], [29], [33], [35], [37], [45], [49], [87]–[99], au titre de notre revue de littérature, conclue que les termes identification d'utilisateur individuel (*unique/singulier*), de session (*serveur/utilisateur*) et de transaction, ainsi que ceux de méthodes d'identification de session (*construction/reconstruction*) et d'implémentation des solutions y afférentes sont utilisés de manière interchangeable dans les titres des travaux identifiés et leurs contenus. Toutefois, les contributions de ces travaux se limitent à l'un de ces termes/concepts. À cet égard, il est à noter que notre contribution porte sur la méthode générique et les concepts canoniques de la structuration réactive des données de journalisation, i.e., l'identification d'un utilisateur individuel singulier et la reconstruction de session serveur [21], [29], [30], [32], [37], [45].

Lorsque l'identification d'utilisateurs individuels uniques par leurs identifiants ou cookies associés n'est pas possible, la séquence de navigation des pages Web provenant de la même IP et agent (*paire IP/Agent*) est supposée être un flux de clics appartenant à un utilisateur individuel singulier (*flux de clics individuel*). Contrairement aux flux de clics d'utilisateurs individuels uniques, les flux de clics d'utilisateurs individuels singuliers n'appartiennent pas nécessairement à des utilisateurs individuels uniques, mais indique seulement que la séquence de navigation est générée à par des utilisateurs distincts. Dans ce contexte, seule l'identification de sessions serveurs individuelles singulière est possible sur la base de certaines heuristiques.

L'identification de session serveur basées sur ces heuristiques est appelée méthode réactive et fournit une session reconstruite censée être aussi proche que possible de la session réelle en termes de contenu, de taille et de durée. [27], [29], [30], [45]. Ainsi, une reconstruction de session réactive consiste à diviser chaque flux de clics individuels singuliers en différentes sessions/visites sur la base d'heuristiques dédiées, i.e., centrée temps, topologie réelle ou graph du site Web.

En raison de la journalisation séquentielle conjuguée aux contraintes du Web Dynamique/Adaptatif ; il n'est pas évident d'identifier les utilisateurs et les sessions individuels. Ainsi, la présente contribution porte sur les méthodes réactives et leurs limites qui seront et analysées en détail dans le cadre de la présente contribution.

IV.2.2. Méthode réactives de structuration

1) Reconstruction générique

Le flux de clics utilisateur est découpé en sessions distinctes si deux pages successives ne sont pas liées. L'information sur les liens entre pages sont vérifiées sur la base de la topologie/structure du site. Les sessions identifiées dont la durée dépasse certaines seuils empiriques d'inactivité (*intervalle de temps*) ou d'activité (*temps de visite*) sont scindées en plusieurs sessions distinctes [21], [22], [24], [26], [32], [34], [35], [37], [43].

Notation. (S) un ensemble de (n) sessions, (P) un ensemble de (m) pages, (L) un vecteur logique (*True/False*) relatif aux liens inter pages, (A) un ensemble de paires distinctes (*IP & User-agent*), (T) un horodatage de (t_n).

Définition 1. Une session (s_n) est une séquence de pages (p_m) contrainte par le tuple (A, L, T) . Les termes de la séquence/session sont obtenus par la fonction-objectif (*Session Reconstruction Function*) ci-dessous décrite par le **processus 1**.

Processus 1.

- (I.1) L'espace de reconstruction des sessions individuelles (*Single Session*) est limité au sein de paires distinctes (IP-Agent) supposées représenter des utilisateurs individuels singuliers ;

$$s_n := p_m \Sigma m [A.p_m = A.p_{m+1}] \quad (1.1)$$

$$\left[\begin{array}{l} \arg. \max \Rightarrow p_m \Sigma m [l(p_m, p_{m+1}) = True] \quad (1.2) \\ \left\{ \begin{array}{l} \arg. \max \Rightarrow p_m \Sigma m [(t(p_1, p_m) \vee t(p_1, p_{m+1})) <> \Phi] \quad (1.3) \end{array} \right. \end{array} \right.$$

$$p_m \Sigma m \notin s_n \Sigma n \rightarrow s_n = p_m \quad (1.4)$$

- (I.2) La reconstruction de session individuelles dans l'espace défini est contrainte pour maximiser les sessions/séquences de pages (*l'entrée REQ.RES dans l'ALF*) sont supposées être liées selon la topologie du site Web ou leurs pages référentes (*l'entrée REF.RES dans l'ALF*), e.g., $page_n$ est la référente de la $page_{n+1}$;
- (I.3) Le processus de reconstruction est contraint de générer des sessions/séquences selon des seuils d'horodatage temporel prédéfinis (*inactivité/activité*) ;
- (I.4) Les pages qui n'ont pas été assignées sont supposées être des utilisateurs individuels singuliers qui n'ont accédé qu'à une seule page (*single page access*).

2) Reconstruction orientée sur le temps

En cas d'indisponibilité d'informations sur la topologie du site Web, e.g., entrée REF.RES non disponible (*CLF*), site Web dynamique (*site Web avec système de personnalisation*); la séquence de pages naviguées par le flux de clic utilisateur individuel singulier est divisée en sessions distinctes/singulière sur la base de seuils de temps empiriques, i.e., temps d'activité ou inactivité. Le seuil empirique d'activité/inactivité le plus utilisé est de 30 minutes d'activité pour une seule session. Le seuil le plus utilisé pour l'inactivité égal à un intervalle de temps de 10 minutes entre deux pages Web

successives naviguées par un même flux de clic. La fonction objectif, **Process 2**, ci-dessous décrit le processus de reconstruction de la session temporelle [22], [29], [38], [44], [45], [45], [73].

3) Reconstruction orientée sur la topologie du site Web – Topologie réelle

Le flux de clics utilisateur singulier est découpé en sessions distinctes/ singulière s'il contient des séquences de deux pages successives qui ne sont pas liées. L'information sur les liens inter pages est vérifiée sur la base de la topologie/structure du site Web lorsque cette information est disponible, e.g., propriétaires du site Web, extraction des URI du site Web [22], [29], [38], [44], [45], [45], [73].

4) Reconstruction orientée sur la topologie inférée du site Web – Graph du site Web

De même que la reconstruction orientée sur la topologie réelle du site Web, le flux de clic utilisateur singulier est découpé en sessions singulières distinctes s'il contient des séquences de deux pages successives qui ne sont pas liées. Par contre, L'information sur les liens inter pages est vérifiée sur la base de l'entrée REF.RES (*page référente*) de l'ALF quand il s'agit d'une journalisation de formats ECLF. Ainsi, une séquence de deux pages successives, où la précédente ne fait pas référence à la suivante, sont considérées comme fin/début de deux sessions distinctes, respectivement [22], [29], [38], [44], [45], [45], [73].

La définition de session et le processus de mise en œuvre des méthodes réactives orientées temps et navigation (*topologie réelle et inférée*) sont présentés aux termes de la fonction-objectif (*Session Reconstruction Function*) ci-dessous.

Process 2.

- (2.1) L'espace de reconstruction des sessions singulières par méthode orientée temps et navigation est limité au sein de paires distinctes (IP-Agent) supposées représenter des utilisateurs individuels singuliers ;
- (2.2) La reconstruction orientée temps est contrainte de maximiser les sessions/séquences à partir de flux singuliers (*paires distinctes IP-Agent*) de clics correspondant aux seuils temps définis ;

$$\left[\begin{array}{l}
 s_n := p_m \Sigma m [A.p_m = A.p_{m+1}] \quad (2.1) \\
 \left. \begin{array}{l}
 \text{arg. max} \left\{ \begin{array}{l}
 p_m \Sigma m [(t(p_1, p_m) \vee t(p_1, p_{m+1})) \langle \rangle \Phi] \quad (2.2) \\
 p_m \Sigma m [l(p_m, p_{m+1}) = True] \quad (2.3)
 \end{array} \right. \\
 p_m \Sigma m \notin s_n \Sigma n \rightarrow s_n = p_m \quad (2.4)
 \end{array} \right.
 \end{array}
 \right.$$

- (2.3) La reconstruction orientée navigation est contrainte de maximiser les sessions/séquences à partir de flux de clics singuliers (*paires distinctes IP-Agent*) dont les pages sont liées sur la base de topologie (*réelle/inférée*) du site Web ;
- (2.4) Les pages qui n’ont pas été assignées sont supposées être des utilisateurs individuels singuliers qui n’ont accédé qu’à une seule page (*single page Access*).

IV.2.3. Analyse des limites et proposition

Les méthodes réactives pour la reconstruction de session sont centrées sur l’agent. Ils supposent que les utilisateurs singuliers sont identifiables par des paires distinctes d’adresses IP et d’agents dont les sessions singulières consistent en une séquence de navigation de pages Web liées. La référence de liaison d’inter pages est vérifiée sur la base de la topologie du site. Cette approche peut être biaisée par les caractéristiques du système Web, telles que la mise en cache, l’hébergement multiple, le Web dynamique, en plus du problème principal de la journalisation, i.e., l’entrelacement des requêtes de plusieurs utilisateurs.

1) Facteurs de limite

- *Système de mise en cache*

Les systèmes de mise en cache exécutés sur des serveurs intermédiaires, à savoir les proxys, sont des algorithmes de prédiction de navigation des pages Web pour les stocker dans le proxy à des fins d’optimisation du trafic. Ainsi, les pages Web prédites sont délivrées à l’utilisateurs à partir du proxy, et par conséquent, l’activité de leur navigation n’est pas journalisée sur les serveurs du site Web. Notez que les agents

(*navigateurs Web*) exécutent également une activité de mise en cache dans laquelle chaque page Web demandée est stockée pendant un certain temps.

- ***L'hébergement multiple***

L'hébergement multiple fait référence aux agents qui convoient les requêtes de l'utilisateur finaux, de sorte que l'adresse IP enregistrée sur le serveur du site Web est celle du dernier proxy qui a convoyé la requête et non l'adresse IP de la machine de l'utilisateur final. Ainsi, dans le cas d'un proxy à IP unique ou multiple, une requête d'utilisateur final peut être enregistrée avec une adresse IP unique pour plusieurs utilisateurs finaux ou plusieurs adresses IP pour un utilisateur final unique/singulier. En outre, l'hébergement multiple désigne le fait qu'un utilisateur final peut naviguer sur le Web à travers différents agents-navigateurs et que plusieurs utilisateurs finaux peuvent naviguer avec le même agent-navigateur.

- ***Le Web dynamique***

La conception Web dynamique s'appuie sur un algorithme dédié à la personnalisation automatique du contenu et à l'adaptation de la structure du site Web, respectivement, en fonction du comportement de chaque utilisateur/groupe d'utilisateurs finaux.

2) Analyse et proposition

- ***Analyse des limites***

Les facteurs ci-dessus peuvent affecter la qualité des sessions reconstruites par méthodes réactives.

En premier lieu, notez que la méthode orientée temps à besoin de la récupération du cache pour compléter les références manquantes dans la séquence de navigation des sessions singulières à reconstruire en fonction des seuils de temps prédéfinis. Un tel processus peut conduire à une affectation erronée des références manquantes récupérées. En outre, la mise en œuvre d'un tel processus est complexe en raison du besoin de connaissances préalables sur les horodatages des requêtes récupérées aux différents niveaux de cache, i.e., agent-navigateur, proxy, serveur Web. Globalement, la récupération des données des caches pour la reconstruction de session est considérée comme une source d'impertinence qui affecte le reste du processus de fouille.

Deuxièmement, la méthode orientée topologie recourt à la topologie du site Web pour vérifier les liens inter pages. Toutefois, dans le cas d'un contenu et d'une structure Web dynamiques où le contenu et la structure Web changent constamment et automatiquement ; il n'est pas évident de définir et de mettre à jour une référence de vérification pertinente. Cela devient même impossible dans le cas d'un contenu personnalisé par utilisateur unique. De plus, les trois méthodes, i.e., orientée temps, topologie réelle et inférée, peuvent être biaisées par le facteur d'hébergement multiple puisqu'elles supposent que les paires distinctes IP-Agent navigateur représentent des utilisateurs distincts (*unique ou singulier*).

Enfin, étant donné que les requêtes sont enregistrées entrelacées quelle que soit leur source, le regroupement des requêtes par paires distinctes IP-Agent peut conduire à une affectation erronée des requêtes lorsque deux utilisateurs distincts ont le même identifiant d'agent (*IP-Agent*). Dans ce cas, les perspectives d'analyse sont limitées car seul le chemin de séquence de session reconstruite peut être utile mais pas le temps de navigation.

- **Proposition**

Les méthodes citées ci-dessus affectent la qualité de la structuration en termes de trois dimensions, i.e., le contenu et la taille de la session, le comportement de navigation représenté et la qualité des modèles découverts [29], [45]. Étant donné que la méthode générique est représentative des limites de celles orientées temps et topologie, qui sont centrées sur des attributs-agent (*IP-Agent-navigateur*) la présente contribution sera, donc, centrée sur l'amélioration de la qualité de la structuration réactive à travers le cas de la méthode générique.

Contrairement à l'approche centrée sur les attributs-agent, la méthode proposée procède à la structuration réactive à partir d'une approche centrée sur le flux de clic-agent. La méthode proposée, axée sur les flux, contraint le processus de structuration à travers une fonction-objectif, de sorte que les sessions reconstruites reposent principalement sur l'identification de flux singuliers pertinents. Le processus centré sur le flux dirige la reconstruction de sessions pour identifier des flux singuliers pertinents avant de les assigner par paires distinctes IP-Agent-navigateur.

Étant donné que l'objet de base de la structuration réactive, qui reflète les limites

de structuration, est la session singulière ; notre approche est censée atténuer les répercussions des facteurs limitants en termes de reconstruction de session sur la base des contraintes de pertinence devant minimiser leur impact. La méthode proposée centrée sur les flux est expérimentée et comparée à la méthode axée sur les agents sur la base des métriques de qualité appropriées proposées dans la littérature y afférente.

IV.3. Contribution 2 : Structuration centrée sur les flux de clics

IV.3.1. Concept et approche

1) Flux de clics pertinents

Contrairement aux méthodes centrées sur l'agent qui contraignent la construction de sessions aux paires distincts IP-Agent-navigateur définis comme des utilisateurs singuliers (voir *Définition 1 et 2*) comme suit : $(s_n = p_m \Sigma m [A.p_m = A.p_{m+1}])$; notre méthode est centrée sur les flux de clics qui consistent en des requêtes de pages successives liées $(p_m \Sigma m [l(p_m, p_{m+1}) = 1])$ contraintes par des paramètres de pertinence analytique avant de les assigner aux utilisateurs et sessions singuliers.

2) Paramètres de pertinence

Notre méthode groupe les flux de clics en sessions singulières par paires distinctes IP-Agent-navigateurs sous la contrainte de maximiser les requêtes qui satisfont à des paramètres de pertinence analytique sensés neutraliser les facteurs limitants. L'objet des paramètres de pertinence en question sont présentés ci-dessous.

- ***Accès pertinent***

Il est entendu par un accès pertinent, les requêtes dont les paires distinctes d'IP-Agent-navigateur représentent un premier accès évident au site Web. Cette information peut être vérifiée sur la base de la racine (*REF.RES.ROO*) de l'attribut REF.RES des requêtes enregistrées dans l'ALF. Une requête dont l'attribut REF.RES.ROO est différent du nom de domaine du site Web représente un premier accès d'un 'utilisateur final au site Web à partir d'un lien externe.

- ***Clic pertinent***

Les requêtes dont les paires d'IP-agent-navigateur appartiennent à l'ensemble des paires d'accès pertinent doivent être prises en compte en priorité lors de l'assignation des requêtes par utilisateur et session singulière.

- **Horodatage pertinent**

Seules les requêtes dont l'horodatage correspond à un utilisateur final (*humain*) doivent être assignées au même utilisateur et session singulière. À cet égard, un intervalle d'horodatage inférieur à un certain seuil peut être qualifié d'assignation erronée qu'il faut minimiser autant que possible, sinon considérer comme un flux de requêtes de robots.

IV.3.2. Méthode d'application

La pertinence d'une session singulière est traitée comme un problème d'optimisation contraint par une hiérarchie des paramètres de pertinence. Ainsi, le processus de reconstruction est dirigé pour maximiser les sessions singulières sur la base des flux de clics identifiés ($s_n = p_m \Sigma m [l(p_m, p_{m+1}) = True]$) pour :

En premier lieu, assignation des requêtes correspondant aux accès pertinents ($p_1 \Sigma s_n \in Req_n [REF. ROO \neq DN]$) et les clics ($\Sigma_2^m P_m \in Req_n \Sigma n [A.p_m \in p_1 \Sigma s_n]$) dont l'horodatage satisfait à la contrainte de seuil de navigation humaine ($p_m \Sigma m [t(p_m, p_{m+1}) \geq \Phi]$) et celle d'utilisateur singulier ($p_m \Sigma m [A.p_m = A.p_{m+1}]$) ; hors des requêtes qui ne satisfont pas à ces contraintes.

En deuxième lieu, le processus itère sur les requêtes non assignées pour assigner celles qui correspondent au maximum des contraintes de pertinence possible jusqu'à ce qu'il n'y ait plus de requêtes compatibles avec les contraintes du processus.

Enfin, les requêtes restantes seront référées comme étant des utilisateurs singuliers qui n'ont accédé qu'à une seule page (*single page Access*).

Le processus de reconstruction est décrit comme une fonction-objectif (*Session Reconstruction Function*).

Process 3.

- (3.1) L'espace de reconstruction de sessions singulières couvre toutes les requêtes qui représente un flux de clics (*une succession de clics sur des pages liées*) ;
- (3.2) Identifier les attributs agents (*IP-Agents navigateurs*) des premiers accès évidents au site Web (*accès pertinent*) par rapport à ceux qui ne le sont pas ;
- (3.3) Maximisez les sessions singulières contenant des requêtes correspondant à des accès pertinents sur la base de leurs identifiants IP-Agent-navigateur appartenant à

l'ensemble d'accès pertinents identifiés (*clics pertinents*), et ;

- (3.4) Maximisez les sessions singulières contenant des requêtes dont l'horodatage correspond à un seuil défini (*horodatages pertinents*) ; ayant le même identifiant d'agent (*IP-Agent-navigateur*).

L'impact ciblé sur la qualité de la reconstruction de session centrée sur le flux (*processus 3*) consiste à atténuer les facteurs limitants en termes de :

- Comme la reconstruction de session est centrée sur les flux ($s_n = p_m \Sigma m [l(p_m, p_{m+1}) = True]$) au lieu des agents ($p_m \Sigma m [A.p_m = A.p_{m+1}]$) (3.5), permet de découvrir des sessions singulières effectuées via un hébergement multiple avant de renvoyer les requêtes restantes à un accès unique.
- Le traitement, en premier lieu, des accès ($p_1 \Sigma s_n \in Req_n [REF.ROO \neq DN]$) et des clics ($\sum_2^m P_m \in Req_n \Sigma n [A.p_m \in p_1 \Sigma s_n]$) pertinents hors de ceux qui ne le sont pas, en plus de la contrainte du seuil d'horodatage pertinent ($p_m \Sigma m [t(p_m, p_{m+1}) > \Phi]$) permet de réduire les affectations erronées dues au facteur d'entrelacement.
- La vérification des liens inter pages sur la base de la topologie inférée ($l(p_m, p_{m+1}) = True \Rightarrow req_n(req.res) = req_{n+1}(ref.res)$) est destinée à surmonter le facteur Web dynamique limitant où la topologie réelle est complexe à obtenir et même impossible en cas de structure et de contenu personnalisés.

$$s_n := p_m \Sigma m [l(p_m, p_{m+1}) = True] ; \quad (3.1)$$

$$l(p_m, p_{m+1}) = True \Rightarrow req_n(req.res) = req_{n+1}(ref.res)$$

arg. max

$$p_1 \Sigma s_n \in Req_n [REF.ROO \neq DN] \quad (3.2)$$

$$\sum_2^m P_m \in Req_n \Sigma n [A.p_m \in p_1 \Sigma s_n] \quad (3.3)$$

$$p_m \Sigma m [t(p_m, p_{m+1}) > \Phi] \quad (3.4)$$

$$p_m \Sigma m [A.p_m = A.p_{m+1}] \quad (3.5)$$

$$p_m \Sigma m \notin s_n \Sigma n \rightarrow s_n = p_m \quad (3.6)$$

IV.3.3. Méthode d'implémentation et d'évaluation

1) Méthode d'implémentation

Afin de comparer notre méthode centrée sur les flux de clics et celles centrées sur l'agent ; les deux méthodes ont été implémentées sous R IDE via une programmation fonctionnelle orientée objet. Les sessions singulières sont reconstruites par un code de fonction (*Objective Function*) et incrémentées dans un objet-liste de (n) sessions (*séquences*) et leurs (m) pages naviguées (*termes de séquence*). La fonction de reconstruction de session a recours à une fonction d'évaluation binaire qui étiquette les requêtes en fonction de leur compliance avec les contraintes de la fonction-objectif, i.e., (1) pour compliance, sinon (0).

L'algorithme IV.1 décrit les étapes d'implémentation des processus de structuration, i.e., centrées flux de clics et agent.

Algorithme IV-1 Reconstruction de session centrée Flux de Clics & Agent

```

01  DATA INPUT & HANDLING
02  C.WLDB.1 ← reqnΣn [TIM, IP, UA, PAG, REF, ROO] ; C.WLDB.2 ← C.WLDB.1
03  OBJECTS
04  Stream-centric Sessions Object (SSO) ⇒ SSO = list (ssonΣ1n n , ssomΣ1m n )
05  Agent-centric Sessions Object (ASO) ⇒ ASO = list (asonΣ1n n , saomΣ1m n )
06  FITNESS FUNCTIONS
07  Agent Vector (AV) ⇒ reqn [IP & UA] ∩ reqn+x [IP & UA] ≠ ∅ ← 1 else 0
08  Stream Vector (SV) ⇒ reqnΣn [PAG ∩ REF ≠ ∅] ← 1 else 0
09  Relevant Access Vector (RAV) ⇒ reqnΣn [REF.ROO ≠ DN] ← 1 else 0
10  Relevant Click Vector (RCV) ⇒ reqnΣn [(IP&UA) ∈ reqn(IP&UA)Σn[REF.ROO ≠ DN]
11  Relevant Timestamps Vector (RTV) ⇒ reqnΣn
12  OBJECTIVE FUNCTION (Stream-centric reconstruction)
13  Sessions Reconstruction ⇒ SCAN C.WLDB
14  WHILE reqn [SV & RAV & RCV & RTV & AV = 1] WEND
15      SSO [sson=1, ssom=1] ← reqn [RAV = 1] & C.WLDB = reqnΣn [-reqnΣn ∈ SSO ]
16      SSO [ssonΣ2n n , ssomΣ2m n] ← reqn [SV = 1] & C.WLDB = reqnΣn [-reqnΣn ∈ SSO ]
17  WHILE reqn [SV & RTV = 1] WEND
18      SSO [ssonΣn=∅ n , ssomΣm=∅ n] ← reqn [SV & RTV = 1] & C.WLDB = reqnΣn [-reqnΣn ∈ SSO ]
19  WHILE reqnΣn ∈ C.WLDB ≠ ∅ WEND
20      SSO [ssonΣn=∅ n , ssomΣm=∅ n] ← reqnΣn
21  DATA OUTPUT 1 (Stream-centric Single Session Object)
22  SSO (ssonΣ1n n , ssomΣ1m n)
23  OBJECTIVE FUNCTION (Stream-centric reconstruction)
24  Sessions Reconstruction ⇒ SCAN C.WLDB.2
25  WHILE reqn [AV & SV = 1] WEND
26      ASO [asonΣ1n n , asomΣ1m n] ← reqn [AV & SV = 1] & C.WLDB = reqnΣn [-reqnΣn ∈ ASO ]
27  WHILE reqnΣn ∈ C.WLDB.2 ≠ ∅ WEND
28      ASO [asonΣn=∅ n , asomΣm=∅ n] ← reqnΣn
29  DATA OUTPUT 2 (Agent-centric Single Session Object)
30  ASO (asonΣ1n n , asomΣ1m n)
    
```

- **(01) Données d’entrée et prétraitement (*Data Input and Handling*) :**
- **(02)** Comme noté en **section 2.2.2**, les données Log brutes enregistrées dans un ALF de format ECLF sont nettoyées pour obtenir une base de données sans bruits (*C. WLDB*). Cette *C. WLDB* contient la journalisation des clics des utilisateurs finaux et les attributs utiles à la structuration en sessions singulières. On a dupliqué la *C. WLDB* pour implémenter les deux méthodes, i.e., centrée flux et agents dans le même algorithme.

- **(03) Objet (*Object*) :**
- **(04 & 05)** L’algorithme de reconstruction est appelé à incrémenter les sessions reconstruites dans objet-liste (*Sessions Object – SO*) de (*n*) sessions (*séquences*) et leur (*m*) pages naviguées (*termes de séquence*). Les sorties de la méthode centrées sur les flux et celle sur les agents, i.e., les sessions, sont notées SSO (*Stream-centered Session Object*) et ASO (*Agent-centered Session Object*), respectivement.

- **(06) Fonction d’évaluation (*Fitness Functions*) :**
- **(07)** Le Vecteur d’Agent (*Agent Vector –AV*) reçoit les valeurs labélisant l’espace de construction de session de la fonction-objectif de la méthode centrée sur les agents, i.e., **Processus 1.1 & 2.1** ; et correspond aussi à la compliance avec la cinquième contrainte de la fonction-objectif de notre méthode centrée sur les flux, i.e., **Processus 3.5**. Ceci est l’une des différences majeures entre notre méthode centrée sur les flux des agents et celle centrée sur les agents, la contrainte définissant l’espace de construction de cette dernière devient seulement une contrainte secondaire de reconstruction pour notre méthode.
- **(08)** Le Vecteur de Flux (*Stream Vector SV*) reçoit les valeurs de compliance avec les contraintes de la méthode centrée sur les agents, i.e., **Process 1.2 & 2.3**, et correspond aussi à la labélisation de l’espace de construction de session de la fonction-objectif de notre méthode centrée sur les flux, i.e., **Process 3.1**. Ceci est l’une des différences majeures entre notre méthode centrée sur les flux des agents et celle centrée sur les agents, la contrainte définissant notre espace de construction était une contrainte secondaire de reconstruction pour la méthode centrée sur les agents.
- **(09), (10), and (11)** évaluent la compliance avec les contraintes de notre méthode

centrée sur les flux, i.e., **Process 3.2, 3.3 & 3.4**, respectivement.

- **(12) Fonction-Objectif/de reconstruction centrée sur les flux (*Stream-centric Objective/Reconstruction Function*) :**
- **(13)** L’algorithme analyse la base de données C. WLDB.1 contenant (n) requêtes $(req_n \Sigma n)$;
- **(14)** tant que les valeurs des index des SV, RAV, RCV, RTV et AV des requêtes analysées $(req_n \Sigma n)$ égalent à 1 ;
- **(15)** incrémente les requêtes $(req_n \Sigma n)$ qui représentent des accès pertinents « *Relevant Access* » [RAV = 1] comme premier terme de la session singulière à reconstruire $([ss_{o_n=1}, ss_{o_m=1}])$ dans l’objet-liste/session SSO; et retire-les de la base C. WLDB.1 $(C.WLDB.1 = req_n \Sigma n [-req_n \Sigma n \in SSO])$;
- **(16)** incrémente les requêtes $(req_n \Sigma n)$ qui représentent des flux de clics [SV = 1] comme termes suivants de la session singulière à reconstruire $([ss_{o_n} \Sigma_2^n n, ss_{o_m} \Sigma_2^m n])$ dans l’objet-liste/session SSO ; et retire-les de la base C. WLDB.1 $(C.WLDB.1 = req_n \Sigma n [-req_n \Sigma n \in SSO])$.
- **(17)** Tant que, seulement, les valeurs des index correspondant de SV et RTV des requêtes analysées $(req_n \Sigma n)$ égalent à 1 ;
- **(18)** incrémente les requêtes $(req_n \Sigma n)$ qui représentent des flux de clics [SV = 1] comme la sessions singulière suivante à reconstruire $([ss_{o_n} \Sigma_{n=\emptyset}^n n, ss_{o_m} \Sigma_{m=\emptyset}^m n])$ dans l’objet-liste/session SSO; et retire-les de la base C. WLDB.1 $(C.WLDB.1 = req_n \Sigma n [-req_n \Sigma n \in SSO])$.
- **(19)** tant que la base C. WLDB.1 analysée n’est pas vide de requêtes $(req_n \Sigma n)$;
- **(20)** incrémente les requêtes $(req_n \Sigma n)$ comme sessions singulière, à reconstruire, d’utilisateurs individuels singuliers qui n’ont accédé qu’à une seule page (*single page Access*), SSO $([ss_{o_n} \Sigma_{n=\emptyset}^n n, ss_{o_m} \Sigma_{m=\emptyset}^m n])$.
- **(21) Sortie de reconstruction de sessions singulière centrée sur les flux (*Stream-centric Single Session Output –SSO*) :**
- **(22)** Un objet-liste incrémenté de (n) sessions et (m) termes chacune, SSO $(ss_{o_n} \Sigma_1^n n, so_m \Sigma_1^m n)$.

- **(23) Fonction-Objectif/de reconstruction centrée sur les agents (*Agent-centric Objective/Reconstruction Function*) :**
- (24) l’algorithme analyse la base de données C. WLDB.2 contenant (n) requêtes $(req_n \Sigma n)$;
- (25) tant que les valeurs des index de AV et SV des requêtes analysées $(req_n \Sigma n)$ égalent à 1 ;
- (26) incrémente les requêtes $(req_n \Sigma n)$ de paires distinctes IP & UA et de pages liées $[SV = 1]$ comme termes des sessions singulières à reconstruire $([aso_{n=1}, aso_{m=1}])$ dans l’objet-liste/session ASO; et retire-les de la base C. WLDB.2 $(C.WLDB.2 = req_n \Sigma n [-req_n \Sigma n \in ASO])$;
- (27) tant que la base de données C. WLDB.2 analysées n’est pas vide de requêtes $(req_n \Sigma n)$;
- (28) incrémente les requêtes $(req_n \Sigma n)$ comme sessions singulières, à reconstruire, d’utilisateurs individuels singuliers qui n’ont accédé qu’à une seule page (*single page Access*), ASO $([sso_n \sum_{n=\emptyset}^n n, sso_m \sum_{m=\emptyset}^m n])$.
- **(29) Sortie de reconstruction de sessions singulière centrée sur les flux (*Agent-centric Single Session Output –ASO*) :**
- (30) Un objet-liste incrémente de (n) sessions et (m) termes chacune, ASO $(aso_n \sum_1^n n, aso_m \sum_1^m n)$.

2) Méthode d’évaluation

Les travaux de revue de littérature et les contributions relatifs à la fouille des données de l’usage du Web et leur structuration ne fournissent pas de cadre unifié pour l’évaluation de la reconstruction de session réactive. Cependant, certaines contributions empiriques et expérimentales ont abordé ce sujet de manière explicite ou implicite sous différents termes, objectifs et contextes, e.g., cadre expérimental pour l’évaluation de la reconstruction de session et la mesure de sa précision, des recherches empiriques approfondies sur les régularités dans l’utilisation du Web, la pertinence de la structuration et son impact sur la découverte de motifs d’usage [24], [29], [45], [92], [92], [100]–[104].

Néanmoins, l’évaluation de la reconstruction de session réactive peut être caractérisée à travers trois paramètres de qualité, i.e., qualité de la session, de la navigation reflétée et de la découverte de motifs d’usage.

- ***Paramètre de qualité de reconstruction de sessions***

Paramètre de la qualité de session : Dans le cas de disponibilité de connaissances à priori antérieure sur des sessions réelles (*structuration proactive*), les résultats de la méthode réactive sont comparés aux résultats proactifs sur la base de la précision et du rappel en termes de mesures liées au nombre, à la taille et à la durée de la session. Sinon, des recherches expérimentales ont démontré que les sessions reconstruites (*méthodes réactives*) par rapport aux sessions réelles (*méthodes proactives*) souffrent d’effets de sous-dimensionnement ou de surdimensionnement. La reconstruction de session réactive tend à fournir un nombre élevé (*surdimensionnement*) de sessions courtes (*sous-dimensionnement*) en termes de nombre de requêtes et temps de navigation.

Paramètre de qualité de navigation reflétée : Certaines recherches empiriques approfondies fournissent des informations sur les fortes régularités liées à la navigation des utilisateurs en termes de longueur de chemin et de contenu parcouru. Par contre, d’autres recherches démontrent que les régularités de navigation diffèrent en fonction du profil des utilisateurs et du contenu consulté.

Paramètre de qualité de découverte de motifs d’usage : Il est mentionné de manière récurrente dans la littérature connexe que la qualité du prétraitement et de la reconstruction de session a un impact sur la qualité de la découverte de motifs d’usage, e.g., nombre de motifs découverts et qualité des règles de prédiction. A cet égard, un large consensus stipule que des sessions reconstruites pertinentes fournissent des modèles et des règles prédictifs plus précis.

- ***Indicateurs de qualité et interprétation***

Indicateurs de qualité : Les paramètres de qualité de la navigation n’étant pas génériques, notre contribution sera évaluée sur la base de la qualité des sessions et la découverte de motifs d’usage (*règles d’association*). À cet égard, douze indicateurs de qualité (*Q.I*) ont été sélectionnés pour comparer la qualité de notre méthode de reconstruction centrée sur les flux et celle centrée sur les agents.

- (QI.1) Nombre de sessions reconstruites.
- (QI.2) Nombre de sessions d'accès singulier (*utilisateurs ayant consulté une seule page*).
- (QI.3) La moyenne de la taille des sessions reconstruites en termes du nombre de pages.
- (QI.4) La moyenne de la durée des sessions reconstruites en minutes.
- (QI.5) La moyenne de visualisation des pages consultées en minutes.
- (QI.6) Nombre de sessions reconstruites reflétant un hébergement multiple.
- (QI.7) Nombre des intervalles de temps de navigation inter pages égalant 0.
- (QI.8) Nombre de règles découvertes avec un Support Maximum (0.01).
- (QI.9) Support Minimum pour découvrir au moins une règle.
- (QI.10) Indicateur de Confiance des règles découvertes avec le Support Minimum.
- (QI.11) Support Minimum pour découvrir au moins une règle avec un lift supérieur à 1.
- (QI.12) Nombre des règles découvertes ayant un lift supérieur à 1.

Interprétation : Les indicateurs **QI.1** et **QI.2** concernent l'effet de surdimensionnement. Comme indiqué ci-dessus, un indicateur bas correspond à une qualité supérieure. Cependant, les **QI.3** à **QI.5** concernent l'effet de sous-dimensionnement. Un indicateur élevé correspond à une qualité élevée. Le **QI.6** correspond au nombre de sessions composées de requêtes provenant de différentes adresses IP et/ou UA (*hébergement multiple*). Cet indicateur reflète la capacité de la méthode à surmonter le facteur limitant de l'hébergement multiple, comme indiqué dans la **section 3.2**. Le **QI.7** concerne le nombre de cas de navigation où l'intervalle de temps entre deux pages naviguées est égal à 0. Ce cas reflète la sensibilité de la méthode au facteur limitant de l'entrelacement, comme indiqué à la **section 3.2**. Les **QI.8** à **QI.12** reflètent les paramètres de qualité de découverte de modèles. Une méthode de reconstruction de la qualité est censée : fournir un nombre plus élevé de règles avec le support maximal (**QI.8**), détecter les règles avec un support minimal plus élevé (**QI.9**), ayant un indicateur de confiance élevé (**QI.10**). Les **QI.11** et **QI.12** portent sur le support minimum permettant la découverte de règles avec un lift supérieur à 1. Notez qu'un tel lift démontre qu'une prévision basée sur un modèle est possible en plus de celle basée sur des règles. Ainsi, **QI.11** et **QI.12** reflètent la qualité prédictive de la méthode en termes

du support minimum nécessaire à la découverte des règles avec un lift supérieur à 1 et leur nombre.

IV.3.4. Expérimentation et évaluation

1) Données et paramètres expérimentaux

Les données expérimentales sont présentées dans la **Table IV.1**. La méthode centrée sur les flux ainsi que la méthode centrée sur les agents ont été testées sur un échantillon de quatre ALF. Les ALF présentés dans cette table sont censés être représentatif des différentes pratiques en termes de conception et de contenu Web. Les ALF concernés représentent :

Table IV-1 Données expérimentales

ALF	DN du Site Web	Source de l'ALF	Clics
ALF 1	www.ai.univ-paris8.fr/	Université	8 266
ALF 2	www.khanyounis.mun.ps	https://www.mosa.gov.ps/khanyounis.mun.ps/log/access.log	2 564
ALF 3	www.megapeloteros.com	http://salablanda.com.ar/megapeloteros.com.access.log	48
ALF 4	www.facades.fr	http://igm.univ-mlv.fr/~cherrier/download/L1/access.log	987

- Le site Web de notre laboratoire (*ALF 1*) ;
- Une administration publique (*ALF 2*) ;
- Deux sites Web commerciaux (*ALF 3 & 4*).
- La taille des ALF est donnée en nombre de requêtes.

Notez que les fichiers journaux (*ALF*) ont été nettoyés des requêtes agents comme expliqué dans la **section 2.2.2**. À cet égard, nous avons eu recours aux méthodes appropriées proposées dans la littérature connexe [23], [27], [28], [37], [44]. Ainsi, la taille des fichiers journaux représente le nombre de requêtes clics hors celles des hits des agents nettoyés. Notez que le rapport entre le nombre des clics des utilisateurs finaux et celui des agents dans un fichier journal brut peut dépasser 1/10 [24], [27], [29]. L'algorithme de reconstruction de session a été implémenté sous l'API SparkR.

2) Evaluation des résultats

Les résultats de qualité de reconstruction de session sont présentés dans la **Table IV.2**. En termes de nombre de sessions, i.e., nombre de sessions (**QL.1**), nombre de sessions à accès unique (**QL.2**) ; notre méthode centrée sur les flux génère un nombre inférieur à celle centrée sur les agents. Ainsi, notre méthode est moins affectée par l'effet de surdimensionnement.

Table IV-2 Evaluation de la qualité de reconstruction

ALF	QL.1	QL.2	QL.3	QL.4	QL.5	QL.6	QL.7	QL.8	QL.9	QL.10	QL.11	QL.12
Reconstruction Centrée sur les Flux												
ALF 1	820	367	2,58	13,92	6,02	2	32	173	0,4	0,47	0,08	2
ALF 2	805	380	3,03	6,87	3,48	36	4	115	0,5	0,56	0,1	10
ALF 3	20	10	2,4	5,57	0,95	41	0	55	0,7	0,8	0,05	40
ALF 4	149	48	3,04	4,3	1,9	5	10	247	0,6	0,8	0,4	2
<i>SUM/AVE</i>	<i>1794</i>	<i>805</i>	2,76	7,67	3,09	<i>84</i>	<i>46</i>	<i>590</i>	0,55	0,66	0,16	<i>54</i>
Reconstruction Centrée sur les Agents												
ALF 1	1 577	1 303	1,34	5,17	3,18	0	59	27	0,21	0,2	0,01	2
ALF 2	1 961	1 688	1,24	0,34	1,09	0	46	26	0,29	0,2	0,06	2
ALF 3	31	23	1,54	4,85	0,84	0	0	20	0,6	0,7	0,03	9
ALF 4	541	467	1,66	1,58	0,4	0	25	47	0,23	0,2	0,01	2
<i>SUM/AVE</i>	<i>4110</i>	<i>3481</i>	1,45	2,99	1,38	<i>0</i>	<i>130</i>	<i>120</i>	0,33	0,33	0,03	<i>15</i>

En ce qui concerne la taille moyenne de session (**QL.3**), la durée moyenne (**QL.4**) et le temps moyen de visualisation des pages (**QL.5**) ; notre méthode fournit une taille de session, une durée et un temps de visualisation plus longs. Cela démontre sa capacité à atténuer l'effet de sous-dimensionnement. Le fait que notre méthode centrée sur les flux offre reconstruit un certain nombre de sessions avec plusieurs adresses IP et UA (**QL.6**) prouve sa capacité à surmonter le facteur limitant de l'hébergement multiple présenté à la **section 3**.

Notre méthode a reconstruit un nombre inférieur (**QL.7**) de sessions avec une intervalle de temps de navigation inter pages nul (0). Cela démontre son efficacité pour atténuer les facteurs limitants d’entrelacement et l’assignation erronée des requêtes d’utilisateur finaux lorsque deux utilisateurs possédant les mêmes adresse IP et UA parcourent les mêmes pages dans le même intervalle de temps.

Enfin, en termes de qualité de découverte de motifs d’usage, notre méthode centrée sur les flux offre une meilleure qualité en termes de :

- Règles découvertes avec le support maximal (**QL.8**) ;
- Support minimal requis pour la découverte de règles (**QL.9**) ;
- Indicateur de confiance des règles avec le support minimum (**QL.10**) ;
- Le support minimum requise pour la découverte de règles avec un lift significatif (**QL.11**) et leur nombre (**QL.12**), et donc la capacité de permettre, au-delà de la règle prédiction, une prédiction basée sur un modèle, e.g., Chaîne de Markov, Séries Temporelles.

Notre méthode centrée sur les flux surpasse la méthode centrée sur les agents en termes de qualité des sessions reconstruites en termes de capacité à atténuer les effets de surdimensionnement (**QL.1 & 2**) et de sous-dimensionnement (**QL.3,4 & 5**). Cependant, les performances en termes d’hébergement multiple (**QL.6**) et d’entrelacement (**QL.7**) nécessitent une validation externe basée sur la disponibilité d’étiquetage à priori de sessions réelles/de méthodes proactives, obtenues à partir de cookies et/ou d’authentification. Enfin, notre méthode présente des avantages significatifs en termes de qualité de découverte de motifs d’usage, à savoir une prédiction basée sur des règles et des modèles.

IV.4. Limites et perspectives

Cette contribution dresse une analyse critique de la structuration des données de l’usage du Web/Log. Elle décrit les limites des méthodes réactives de reconstruction de session en termes de session reconstruites, de qualité de navigation reflétée et de découverte de motifs d’usage. Les limites des méthodes analysées sont dues aux répercussions des caractéristiques du Web dynamiques et des propriétés du système de journalisation Web sur la qualité de la reconstruction de session. Les facteurs limitants affectent principalement les séquences de navigation, le contenu et les identifiants des

utilisateurs finaux, i.e., l'entrelacement, la topologie dynamique, les identifiants multiples. Les méthodes analysées sont centrées sur les agents et s'avèrent inappropriées dans un tel contexte.

Étant donné que la méthode proposée est centrée sur les flux de clics de pages liées sur la base de la topologie réelle du site Web ; elle présente des avantages significatifs par rapport à la méthode de structuration centrée sur les agents en termes de session, de qualité de navigation reflétée et de découverte de motifs d'usage. Enfin, les performances de la méthode centrée sur les flux doivent être validées sur la base de références réelles, à savoir des sessions pré labélisées obtenu dans un contexte de construction de sessions par méthode proactive.

Chapitre V
Troisième Contribution
Optimisation de l’Usage du Web

V.1. Contexte, problème et contribution visée

Ce chapitre présente les résultats préliminaires de nos travaux sur l'optimisation de l'usage du Web basée sur les données de journalisation. Il présente le contexte et le problème de l'optimisation par dimension, illustre ses avantages et ses intérêts, analyse ses aspects contradictoires (*parasitique*) et évoque le problème du contrôle du produit de l'optimisation globale. A cet égard, nous proposons à titre de prospection une méthode de classification semi-supervisée pour le contrôle d'une optimisation équilibrée (*symbiotique*). Le potentiel de la méthode prospectée est évalué par une démonstration sur des données de journalisation du site Web de notre université.

Apprendre des données d'utilisation pour optimiser les temps de réponse, adapter la structure du site et raccourcir les chemins de navigation sont des défis critiques pour les sites Web, notamment ceux commerciaux. Ces optimisations permettent d'attirer suffisamment de visiteurs à démarcher dans la perspective de les convertir en clients et les fidéliser.

Une structure adaptée qui met en évidence les raccourcis vers les contenus d'intérêt pour les utilisateurs dans les meilleurs délais, conduirait probablement à les garder le plus longtemps sur le site Web. Un visiteur attiré atteignant ses objectifs par un chemin court, naviguant sur le site Web dans une structure adaptée, dans des temps raisonnables ; est susceptible d'être conservé suffisamment sur le site Web pour être abordé tout en augmentant la possibilité de le convertir en un client fidèle.

Pour ce faire, l'optimisation de l'utilisation du Web basée sur les données de journalisation aborde séparément trois dimensions, i.e., l'optimisation du trafic, de la structure, et des chemins traversés. Ces optimisations servent l'intérêt de différents acteurs, à savoir [13], [21], [25]–[27], [30], [31], [33], [35], [36], [69]–[72] :

- L'artefact, en l'occurrence le site Web, le serveur, et le réseau dont l'intérêt réside dans la fluidité du trafic et l'optimisation de la consommation des ressources, la consommation de la bande passante, etc. ;
- Les utilisateurs-finaux dont l'intérêt réside dans l'optimisation des temps de réponse, des structures adaptées à leur habitude de navigation ;
- L'analyste qui peut être le concepteur, le propriétaire du serveur ou du commerce dont l'intérêt porte sur la disponibilité de l'information de journalisation pour l'analyse

devant servir la conception de services futures, marketing, publicité, etc.

Les dimensions objet d’optimisation (*trafic, chemins, structure*) dépendent de trois facteurs, à savoir, le temps d’accès aux ressources, leur taille, et la longueur du chemin parcouru [13], [30], [31]. Vu l’absence de tendance de corrélation liant ces facteurs, une démarche de traitement par dimension peut conduire à des optimisations parasitiques (*conflituelles et/ou contradictoires*), en l’occurrence, l’optimisation d’une dimension au détriment d’une autre, et ainsi, servir l’intérêt d’une partie au détriment d’une autre, e.g. :

- La mise en cache pour l’intérêt de l’artefact réduit la portée et la pertinence de l’analyse car les pages mises en cache ne sont pas journalisées sur le serveur ;
- Le raccourci des chemins de navigation dans l’intérêt des utilisateurs peut générer un fort flux sur le réseau, e.g., effet de l’aiguillage de la navigation et sa concentration dans le temps.
- L’adaptation de la structure peut influencer sur le trafic de par la taille et la longueur des nouveaux chemins de navigation générés.

Les indicateurs sur les facteurs déterminants peuvent donner un aperçu explicatif de la valeur du produit symbiotique de l’optimisation globale.

Par contre, vu :

- L’absence de tendance de corrélation liant ces facteurs déterminants ;
- La dépendance, du facteur « Temps » de sollicitation des ressources/pages Web, de l’utilisateur le rendant, ainsi, hors de contrôle ;

La **prédiction pour le contrôle** et le filtrage des optimisations distinctes par produit symbiotique de l’optimisation globale devient problématique n’ayant pas de solution comme problème de fonction de régression.

Notre contribution entreprend la question comme problème de classification hors modèles de fonction de régression, et ceci pour proposer une méthode d’apprentissage semi-supervisé pour filtrer et contrôler la valeur symbiotique de l’interaction. Une démonstration de notre méthode est menée sur les données de journalisation du site Web de notre université.

L’objectif étant le **contrôle** de l’optimisation de plusieurs dimensions à la fois

dans un contexte de Web dynamique ; et face à ces contraintes, i.e., structures, chemins de navigation et tailles des ressources/pages uniques dynamiques conduits par des agents-robots ; impose la prospection d'une méthode de prédiction qui :

- Ne dépend pas des facteurs hors contrôle, en l'occurrence le temps de visite ;
- Soit applicable dans un contexte de données massives accentuées par le caractère dynamique automatisé et virale des modifications des chemins, des structures et des tailles de pages/URI unique ;
- Est capable de produire des résultats pertinents et fiables sans contraintes de coûts invalidantes pour son applicabilité et intégration dans un écosystème de fouille de données massives d'usage Web.

Il s'agit d'une approche qui permet de filtrer les propositions d'optimisations (*raccourcissement des chemins de navigation, modification de la structure du site Web, mise en cache/temps d'expiration*) d'un système d'adaptation automatique sur la base de la prédiction de leur produit symbiotique/parasitique.

V.2. Illustration des méthodes et des techniques d'optimisation

V.2.1. Contexte expérimental et organisation des données

1) Contexte expérimental

Les données expérimentales utilisées pour l'illustration et la validation de notre approche sont les données de journalisation du site Web de l'université Paris 8 fourni avec IP anonymisées (*cryptées*). Il s'agit de 3 Giga de données de journalisation, contenant 15 540 897 observations/requêtes, couvrant une période d'activité utilisateurs d'une année allant du 23/07/2017 au 04/07/2018.

Les travaux et résultats préliminaires de notre recherche ont été conduits sur une échèle expérimentale et ont porté sur un échantillon de 15 jours d'activité utilisateurs. Le descriptif des données expérimentales est présenté en **Table V.1**.

Table V-1 Données expérimentales

Période D'Activité	Nbr. Requêtes (Clics & Hits)	Nbr. Clics (Après nettoyage)	Nbr. Utilisateurs (Global)	Nbr. Utilisateurs (2 clics et plus)
10-24/05/2018	1 171 731	284 374	189 967	36 065

Les ratios clics/hits est de 0,24, à savoir 1/4. C'est un ratio qui demeure modéré par rapport aux données expérimentales utilisées au titre de nos expérimentations des contributions de nettoyage et structuration des données Log. L'échantillon a été nettoyé et structuré via nos méthodes de contributions. Ainsi, le nettoyage des données à réduit le volume à traiter de 75%. Le nombre moyen de sessions utilisateurs ayant navigué plus d'une page par rapport au nombre global des utilisateurs est de 19%.

Concernant le taux important de sessions utilisateurs ayant accédé à une seule page, il est à préciser ce qui suit :

- S'agissant d'une structuration réactive, le nombre d'utilisateurs représente le nombre de sessions reconstruites. En structuration réactive les concepts de sessions et utilisateurs sont permutable [29], [33], [37], [45]. Comme indiqué dans la contribution sur la structuration, la reconstruction réactive tend à générer un nombre important de sessions courtes.
- Les systèmes de mise en cache (*proxy et navigateur*) sont un facteur qui réduit forcément les sessions de deux clics à un seul, le cas échéant où l'un d'eux porte sur des pages fréquentes ayant été prédites et mises en cache pour l'optimisation du trafic.
- La nature du site et l'indexation externe directe (*sur les moteurs de recherche*) de liens internes parmi les résultats de recherches qui porte sur le site Web, ce qui est le cas pour le site Web de l'université Paris 8, e.g., pages Formation, Master, Licence, etc.
- Etant donnée que les données Log objet de notre expérimentation n'ont pas fait l'objet de « complètement » par la récupération du cache ; et que la navigation sur le site Web de l'université Paris 8 ne fait l'objet d'aucun système de filature et d'identification des utilisateurs ; aucune inférence de données manquantes mises en cache n'a été entreprise, pour des raisons de pertinence. Aussi, il est à signaler qu'éventuellement certaines sessions courtes n'ont pu être journalisées sur le serveur, le cas échéant de sessions portant, dans leur totalité, sur des pages fréquentes à fort probabilité de prédiction et mise en cache.

La préparation, l'organisation des données ainsi que les expérimentations tant pour l'illustration des méthodes et techniques d'optimisation et leurs limites que pour la validation de la méthode proposée sont conduites sous R via l'IDE R Studio et l'API Apache Spark R.

2) Organisation et formatage des données

Pour les besoins d'illustration et expérimentation, trois tables ont été utilisées pour l'organisation et la présentation des données structurées en entrée des différents traitements, i.e., une table de format transactionnel des sessions utilisateurs (*USE_TAB*), une table descriptive statistique des transactions/sessions (*STA_TAB*) et une table descriptive statistique de la structure de navigation par page de fin (*suffixe*) de sessions (*STR_TAB*).

La table utilisateurs (*USE_TAB*) présente la structuration en sessions utilisateurs sous le format transactionnel vertical, elle représente notre base de données de journalisation nettoyées et structurées. Un échantillon de dix (10) observations contenant les sessions de trois utilisateurs est présenté en **Table V.2-USE_TAB**. Cette table présente les sessions utilisateurs au format transactionnel par identifiant utilisateur/session (*USE_ID*), les pages visitées codifiées (*ACC_PAG*), l'ordre de visite (*CLI_ORD*), la page référente (*REF_PAG*), l'horodatage d'accès (*ACC_TIM*), et le temps de visite au format Posix (*Horodatage Unix*), et enfin la taille de la page visitée en octet (*volume transféré serveur-client*).

Table V-2 Table Sessions Utilisateurs [USE_TAB]

USE_ID	CLI_ORD	ACC_PAG	REF_PAG	ACC_TIM	VIE_TIM	SIZ
1	1	680	1596	1525966277	6	4853
1	2	699	680	1525966283	51	1069
27	1	3761	1552	1525980356	7	349
27	2	28	3761	1525980363	6	4755
27	3	2254	28	1525980369	51	2532
33	1	3761	639	1525966495	2	428
33	2	28	3761	1525966497	4	4755
33	3	2254	28	1525966501	5	2532
33	4	2524	2253	1525966506	9	426
33	5	2348	2524	1525966515	185	851

La **Table V.3-STA_TAB** présente un échantillon aléatoire de dix (10) observation de la table statistique (*STA_TAB*). Cette table donne un descriptif d'agrégations statistiques par utilisateur/session (*USE_ID*), i.e., la longueur de la session en nombre de clicks (*NBR_CLI*), le nombre de pages uniques de la session (*NBR_UNI_PAG*) comme indicateur de structure, la durée de la session (*SUM_TIM*) en

secondes, le durée moyenne par page visitée (*MEA_TIM*), la taille globale des pages visitées en octet de la session (*SUM_SIZ*), la taille moyenne en octet par page (*MEA_SIZ*), et enfin le flux généré en octet/seconde par session (*FLU*). Il s’agit d’un descriptif général de la navigation des utilisateurs sur le site.

Table V-3 Table Statistiques Sessions [STA_TAB]

USER_ID	NBR_CLI	NBR_UNI_PAG	SUM_TIM	MEA_TIM	SUM_SIZ	MEA_SIZ	FLU
1	2	2	57	28,50	5922,00	2961,00	103,89
16	2	2	106	53,00	2747,00	1373,50	25,92
24	2	2	55	27,50	4610,00	2305,00	83,82
25	2	2	55	27,50	4629,00	2314,50	84,16
26	2	2	64	32,00	5485,00	2742,50	85,70
27	3	3	64	21,33	7636,00	2545,33	119,31
33	6	6	256	42,67	13175,00	2195,83	51,46
34	5	5	82	16,40	8992,00	1798,40	109,66
41	2	2	915	457,50	2881,00	1440,50	3,15
43	3	3	108	36,00	1591,00	530,33	14,73

La **Table V.4-STA_STR** présente un descriptif d’agrégations statistiques des sessions de navigation des utilisateurs par pages de fin (*suffixe*) de sessions (*PAG_SUF_ID*), i.e., le nombre de chemins uniques navigués pour atteindre la page (*NBR_UNI_PAT*), le nombre de longueurs uniques des chemins navigués (*NBR_UNI_LEN*), le nombre de pages uniques naviguées ayant servi les chemins vers le suffixe (*NBR_UNI_PAG*), le temps moyens en secondes des sessions du suffixe (*MEA_TIM_NAV*), le volume moyen généré en octet des sessions du suffixe (*MEA_SIZ_NAV*), la vitesse moyenne en pages/secondes de navigation des pages des sessions du suffixe (*MEA_SPE_NAV*), le flux moyen du trafic généré des sessions du suffixe (*MEA_FLU*).

Cette table de données est utile pour l’analyse de la structure de la navigation vers les contenus d’intérêts, vu que la dernière page d’une session représente l’arrivée à expiration de l’intérêt de navigation de l’utilisateur, i.e., atteinte du contenu d’intérêt ciblé ou l’intérêt pour sa recherche [24], [29], [45], [92], [92], [100]–[104].

Table V-4 Table Statistiques Page [STA_STR]

PAG_SUF_ID	NBR_UNI_PAT	NBR_UNI_LEN	NBR_UNI_PAG	MEA_TIM_NAV	MEA_SIZ_NAV	MEA_SPE_NAV	MEA_FLU
1045	98	0	11	187	12 145	60	88
1047	35	1	10	223	13 168	56	71
1057	1	0	5	167	25 515	28	153
1058	3	1	16	438	36 052	44	174
1059	11	0	8	194	21 386	38	162
1061	2	0	10	128	29 472	19	222
1062	11	0	7	146	16 670	40	143
1063	7	0	5	87	12 839	33	170
1064	2	0	4	120	19 690	28	187
1065	1	0	4	122	18 700	31	153

V.2.2. Optimisation de la structure

1) Objectifs

La fouille des données de journalisation pour l'optimisation de la structure d'un site Web vise son adaptation aux habitudes de navigation des utilisateurs. La structure initiale d'un site Web étant conçue selon la perception du concepteur, de son contenu, et des utilisateurs ciblés ; peut s'avérer inadaptée à celle des utilisateurs ayant réellement navigué sur le site. Ainsi, la navigation des utilisateurs sur le site est analysée sur la base des données de journalisation dans la perspective d'adapter sa structure via l'identification des pages fréquemment visitées ensemble lors d'une même session, et ceci dans le but de les lier le cas échéant ou elles ne le sont pas, ou vice versa.

2) Techniques

Les techniques de fouille utilisées à cet effet sont les règles d'association. Une règle d'association est une implication de l'ordre de : $X \Rightarrow Y$.

Définition 1. Etant donnée :

- $I = \{ i_1, i_2, \dots, i_m \}$, où (I) est un ensemble de (m) éléments/pages Web (i) ;
- $T = \{ t_1, t_2, \dots, t_n \}$, où (T) est un ensemble de (n) transactions/sessions (t) ;
- $X = Kt_n$, où (X) est un sous-ensemble (K) de (m) éléments/pages Web (i) de (I) objet de (n) transactions/sessions de (T) ;
- $Y = K't_n$, où (Y) est un sous-ensemble (K') de (m) éléments/pages Web (i) de (I) objet de (n) transactions/sessions de (T) ;

Une règle d’association de l’ordre de $X_{Pages\ Web} \Rightarrow Y_{Pages\ Web}$ mesure la fréquence d’occurrence de (X, Y) dans un ensemble (T) de (n) de transaction/sessions (t) à travers les indicateurs ci-après :

- L’indicateur de Support (S) qui mesure la fréquence relative des transactions de co-occurrence $S(X, Y)$ par rapport au nombre total des transactions, estimant ainsi la probabilité de co-occurrence $P(X \cap Y)$ dans une transaction, comme suit :

$$S(X, Y) = \frac{\sum n t_n (X \cap Y)}{\sum n t_n} = P(X \cap Y)$$

- L’indicateur de Confiance (C) qui mesure la fréquence relative des transactions de co-occurrence conditionnelle $C(X \rightarrow Y)$ par rapport au nombre des transactions d’occurrence de (X) , estimant ainsi la probabilité de co-occurrence conditionnelle $P(X|Y)$ dans une transaction, comme suit :

$$C(X \rightarrow Y) = \frac{\sum n t_n (X \cap Y)}{\sum n t_n (X)} = P(X|Y)$$

- L’indicateur Lift (L) qui mesure l’interdépendance $L(X \Rightarrow Y)$, estimant ainsi la corrélation d’occurrence $Cor.(X, Y)$, comme indiqué ci-dessous. Un Lift différent de 1 confirme la dépendance de (X) et (Y) et valide le potentiel prédictif de la règle.

$$L(X \Rightarrow Y) = \frac{C(X \rightarrow Y)}{S(y)} = \frac{S(X, Y)}{S(x) \times S(Y)} = P(X|Y)$$

Une variété d’algorithmes et méthodes d’extraction de règles d’association sont proposés dans la littérature [8], [36], [39], [105]–[112], e.g., Apriori, FP-Growth, Eclat, SaM, SETM, AIS, etc. La plupart des algorithmes constituent des optimisations de la méthode Apriori en termes de pertinence et fiabilité en fonction de la nature des données, leur taille, et de l’objectif de l’analyse.

En fonction de la nature des données, leurs volumes, l’objectifs et le contexte de l’analyse ; les seuils d’intérêt sont définis sur la base de supports, confiance, lift minimum ou maximum. Contraindre l’extraction par un seuil minimum évitera un nombre important de règles inutiles et/ou évidentes. Une règle évidente de la structure de navigation est celle perceptible, sans analyse approfondie, de par la nature du contenu et

la structure du site. Par contre, la définition de seuils minimum élevés est souvent le cas de contextes dont on dispose de connaissances apriori par rapport à l'objectif de l'analyse. Par contre, l'utilisation de seuils maximum est souvent le cas d'analyse exploratoire dont on ne dispose pas des connaissances apriori. Aussi, la définition de seuils maximum est utile pour la découverte des cas atypique. Enfin, il est recommandé de procéder à l'échantillonnage des données en fonction de l'objectif de l'analyse ; et d'appliquer les seuils par échantillon d'intérêt [38]–[40], [42], [105]–[107], [113]–[115].

Il s'agit là de définir une stratégie de définition des seuils et des échantillons d'intérêt car dans le cas de données massives et de grande diversité, il est peu probable d'obtenir des règles à fort seuils d'intérêt au-delà des règles évidentes. Les échantillons d'intérêt peuvent être définis d'une manière objective (*partitionnement, paramètre de la statistique descriptive, nature des données, etc.*) ou subjective le cas échéant de connaissances apriori sur l'objet représenté par les données et/ou son contexte. Comme aussi, il est à préciser que dans un contexte de Web dynamique dirigé par des agents-robots sur la base des règles de motifs fréquents, les motifs de faibles supports ont un intérêt particulier, étant données que ce ne sont pas, probablement, le résultat évident de l'adaptation dynamique[34], [35], [40], [41], [105], [114], [115].

3) Illustration

- *Paramètre d'extraction des règles*

Pour l'extraction des règles d'association de notre échantillon, la méthode Apriori a été appliquée à la base de données transactionnelles y afférente, i.e., **Table V.2-USE_TAB**. On a utilisé un support maximum de 0,01 au vu de notre perspective exploratoire et l'absence de connaissance apriori particulière sur la structure du site Web et les habitudes de navigation de ses utilisateurs.

Dans un premier temps on a procédé à une extraction générique des règles sans stratégie particulière, puis une deuxième extraction stratégique, et ceci pour illustrer -en plus des règles utiles à l'adaptation de la structure- le gain en information utile en la matière via une application par sous-échantillon d'intérêt (*extraction stratégique*). Les résultats de l'extraction générique et celle stratégique sont présentés dans les **Tables V.4 & 5**. Les tables présentent les 5 premières règles par support (**SUP**) ou confiance (**CONF**), et le cas échéant d'autres règles pour le besoin d'illustration. L'identifiant des

règles (**ID**) représente son classement parmi celles de l’extraction concernée. Les indicateurs de Lift (**LIF**) et le nombre d’utilisateurs ayant supporté la règles (**COU**) sont aussi présentés.

Les seuils, échantillons, sous-échantillons, les résultats et les informations utiles pour une adaptation de la structure sont commentés et analysé par table et sous-table et comparés ci-dessous, puis une interprétation au titre des adaptations les plus utiles à retenir est donnée.

- *Analyse des résultats et interprétations*

Extraction Générique : La **Table V.5** présente les résultats d’extraction générique, des règles d’association, appliquée à notre base transactionnelle (*Table V.2-USE_TAB*). Il s’agit d’extraction portant sur les transactions (*pages externes incluse*) triée par confiance (**Sous-Table V.5.i**), puis par support (**Sous-Table V.5.ii**).

Parmi les connaissances utiles (*règles et interprétation*) et/ou actions à prévoir au titre d’une optimisation de la structure contenues dans les sous-tables ci-dessus, on peut citer :

- Un aperçu global reflète l’information évidente sur la navigation des utilisateurs du site, i.e., pages d’accueil, formation, master, licence. Il s’agit soit de de pages indexées directement avec la page d’accueil sur les moteurs de recherche, soit figurant (*ayant un lien*) sur la page d’accueil.

Table V-5 Résultats d’extraction générique des règles navigation

ID	REGLES	SUP	CONF	LIF	COU
(i) REGLES GENERALES SANS CONTRAINTES TRI PAR CONF					
1	{UFR-instituts-et-departements} => {-FORMATIONS-}	0,018	0,874	3,053	661
2	{Licences , Masters,univ-paris8.fr} => {-FORMATIONS-}	0,015	0,870	3,039	550
3	{Licences , univ-paris8.fr} => {-FORMATIONS-}	0,027	0,842	2,941	961
4	{Formation-continue} => {-FORMATIONS-}	0,012	0,839	2,930	443
5	{Licences , Masters} => {-FORMATIONS-}	0,035	0,724	2,529	1 248
10	{google.fr , Licences} => {-FORMATIONS-}	0,015	0,539	1,882	532
(ii) REGLES GENERALES SANS CONTRAINTES TRI PAR SUP					
1	{ } => {univ-paris8.fr}	0,360	0,360	1,000	12 973
2	{ } => {google.fr}	0,308	0,308	1,000	11 101
3	{ } => {-FORMATIONS-}	0,286	0,286	1,000	10 329
4	{ } => {google.com}	0,196	0,196	1,000	7 068
5	{ } => {Masters}	0,157	0,157	1,000	5 649
10	{google.fr} => {univ-paris8.fr}	0,103	0,336	0,933	3 727

- La vérification de l'existence de liens entre les pages objet des règles a fort indicateurs de confiance dont les supports sont jugés objectivement ou subjectivement significatifs, e.g., { Licences , Masters } => { FORMATIONS } (**Sous-Table V.5.i – ID.5**). Cette vérification peut s'effectuer sur la base de topologie réelle du site (*connaissance a priori*) ou construite (génération du graph du site, extraction de motif séquentiel).
- Au-delà des règles évidentes qu'on suppose que le concepteur du site a prévu et au vu de l'indicateurs de confiance des règles {UFR-instituts-et-departements}=>{FORMATIONS} (**Sous-Table V.5.i – ID.1**), { Licences , Masters,univ-paris8.fr }=>{ FORMATIONS } (**Sous-Table V.5.i – ID.2**); la structure du site gagnerai en adaptabilité en mettant directement des liens qui pointent vers les pages licence et master à partir de la page d'accueil et/ou particulièrement de celle de formation, ce qui n'est pas le cas actuellement. Enfin, étant donné que la règle {UFR-instituts-et-departements}=>{FORMATION} (**Sous-Table V.5.i – ID.1**) a l'indicateur de confiance le plus élevé peut donner lieu à une adaptation qui consisterai à son indexation à côté de la page d'accueil sur les moteurs de recherches ou à la rigueur sa mise en avant via un lien sur la page d'accueil.

Par conséquent, il importe de signaler les limites des règles et les éventuels biais à l'interprétation d'une telle extraction générique, à savoir :

- Les règles dont l'indicateur de confiance est élevé ont un faible support car les pages externes ayant pointé vers le site sont prises en compte (**Sous-Tables V.5.i & ii – ID's.10**). La prise en compte de ces pages, en plus de celles qu'elles ont pointé (*pages de premier accès au site*), augmente le support des pages comme item/ensemble de longueur 1 sans indicateurs utile de confiance et de lift (**Sous-Table V.5.ii**). Ceci intervient au détriment des indicateurs de supports et de confiance des items/pages de navigation interne qui sont plus utiles pour une adaptation de la structure (**Sous-Table V.5.i**). Cet effet est dû au fait que plus le site a des utilisateurs et de pages d'accès directes hors page d'accueil, plus les items de longueur 1 portant sur ces pages internes et celles les ayant pointés sont nombreux, ce qui réduit la proportion du reste des items.

Pour illustrer les limites et les biais à l'interprétation d'une adaptation basée sur une extraction générique des règles d'associations, nous avons procéder à une extraction

stratégique qui repose sur l'application des seuils d'intérêt maximum par échantillon d'intérêt en fonction de la nature des items/pages et l'information utiles visée.

Extraction Stratégique : La **Table V.6** présente notre stratégie et les résultats y afférents. Nous expliquerons notre stratégie et interprétons ses résultats pour démontrer les biais des règles génériques, d'une part, et d'autre part, montrer la significativité de celles obtenues via une démarche stratégique, et ceci en commentant les résultats par sous-tables comparées à ceux de l'extraction générique.

La **Sous-Table V.5.i** représente une extraction de règles limitée à l'ensemble des pages internes, les pages externes étant utiles uniquement pour une connaissance sur les sites pointant vers le nôtre. A titre de cette sous-table, il importe de mettre le point sur ce qui suit :

- Comparé aux règles d'extraction générique sur l'ensemble des pages (**Sous-Table V.5.i**), nous obtenons de meilleurs indicateurs de confiance dont deux avec de meilleurs supports (**Sous-Table V.6.i. IDs.2 & 4**), i.e., { Licences-2015-2016,univ-paris8.fr }=>{ -FORMATIONS- } et { UFR-instituts-et-departements,univ-paris8.fr }=>{ -FORMATIONS- }.
- Aussi, la première règle de l'extraction générique (**Sous-Table V.6.i. ID.1**) ayant un indicateur de confiance plus important que la première règle de l'extraction générique (**Sous-Table V.5.i. ID.1**), et avec un support légèrement inférieur ; est d'autant plus significative du fait qu'elle porte sur une implication entre trois items/pages, alors que la première règle de l'extraction générique porte seulement sur une implication de deux pages. Aussi, cette extraction identifie les items { Premiere-inscription-a-Paris-8,3290,univ-paris8.fr & INSCRIPTIONS-91 } (**Sous-Table V.6.i. ID.3**) ce qui n'était pas le cas pour l'extraction générique.
- Au titre d'une adaptation de structure, l'utilité d'indexer la page { UFR-instituts-et-departements } à côté de la page d'accueil sur les moteurs de recherche ou de mettre des liens vers cette page dans la page d'accueil est plus affirmée du fait qu'elle fait partie d'une implication portant sur plus d'un item/page avec un indicateur de confiance proche de 100%, i.e., 0,98 & 0,97.

Table V-6 Résultats d'extraction stratégique des règles

ID	REGLES	SUP	CONF	LIF	CON	
(i) REGLES GENERALES TRI PAR CONF (CONTRAINTE → PAGES EXTERNE EXCLUES)						
1	{Licences-2015-2016,UFR-instituts-et-departements} => {-FORMATIONS-}	0,010	0,982	3,361	377	
2	{Licences-2015-2016,univ-paris8.fr} => {-FORMATIONS-}	0,032	0,977	3,345	1 143	
3	{Premiere-inscription-a-Paris-8,3290,univ-paris8.fr} => {-INSCRIPTIONS-91-}	0,012	0,975	4,684	428	
4	{UFR-instituts-et-departements,univ-paris8.fr} => {-FORMATIONS-}	0,028	0,974	3,336	993	
5	{Licences-2015-2016,Masters,univ-paris8.fr} => {-FORMATIONS-}	0,011	0,955	3,270	405	
(ii) REGLES GENERALES TRI PAR SUP (CONTRAINTE → PAGES EXTERNE EXCLUES)						
1	{}	=> {univ-paris8.fr}	0,302	0,302	1,000	10 905
2	{}	=> {-FORMATIONS-}	0,292	0,292	1,000	10 534
3	{}	=> {-INSCRIPTIONS-91-}	0,208	0,208	1,000	7 507
4	{}	=> {Masters}	0,155	0,155	1,000	5 582
5	{univ-paris8.fr}	=> {-FORMATIONS-}	0,098	0,325	1,113	3 545
(iii) REGLES GENERALES TRI PAR CONF (CONTRAINTE → ENSEMBLE DE LONGUEUR 2 ET PLUS & PAGES EXTERNE EXCLUES)						
1	{Licences-2015-2016,UFR-instituts-et-departements} => {-FORMATIONS-}	0,010	0,982	3,361	377	
2	{Licences-2015-2016,univ-paris8.fr} => {-FORMATIONS-}	0,032	0,977	3,345	1 143	
3	{Premiere-inscription-a-Paris-8,3290,univ-paris8.fr} => {-INSCRIPTIONS-91-}	0,012	0,975	4,684	428	
4	{UFR-instituts-et-departements,univ-paris8.fr} => {-FORMATIONS-}	0,028	0,974	3,336	993	
5	{Licences-2015-2016,Masters,univ-paris8.fr} => {-FORMATIONS-}	0,011	0,955	3,270	405	
(iv) REGLES PAGES FREQUENTES TRI PAR CONF (CONTRAINTE → ENSEMBLE DE LONGUEUR 1 & PAGES EXTERNE EXCLUES)						
1	{}	=> {univ-paris8.fr}	0,302	0,302	1,000	10 905
2	{}	=> {-FORMATIONS-}	0,292	0,292	1,000	10 534
3	{}	=> {-INSCRIPTIONS-91-}	0,208	0,208	1,000	7 507
4	{}	=> {Masters}	0,155	0,155	1,000	5 582
5	{}	=> {Licences-2015-2016}	0,095	0,095	1,000	3 426
(v) REGLES SITES REFERENTS FREQUENTS TRI PAR CONF (CONTRAINTE → ENSEMBLE PAGES REFERENTES & DIFFERENT DU DN & LONGUEUR 1)						
1	{}	=> {google.fr}	0,453	0,453	1,000	11 568
2	{}	=> {google.com}	0,284	0,284	1,000	7 254
3	{}	=> {bing.com}	0,041	0,041	1,000	1 045
4	{}	=> {android-app:}	0,017	0,017	1,000	440
5	{}	=> {fp.univ-paris8.fr}	0,015	0,015	1,000	377
6	{}	=> {google.dz}	0,011	0,011	1,000	278
(vi) REGLES SITES REFERENTS FREQUENTS HORS MOTEURS DE RECHERCHE TRI PAR CONF (CONTRAINTE → ENSEMBLE PAGES REFERENTES & DIFFERENT DU DN & LONGUEUR 1)						
1	{}	=> {com.google.android.googlequicksearchbox}	0,064	0,064	1,000	344
2	{}	=> {ecosia.org}	0,043	0,043	1,000	231
5	{}	=> {qwant.com}	0,033	0,033	1,000	178
8	{}	=> {duckduckgo.com}	0,022	0,022	1,000	119
13	{}	=> {m.facebook.com}	0,018	0,018	1,000	95
14	{}	=> {Gmain.jhtml}	0,017	0,017	1,000	94
17	{}	=> {www-artweb.univ-paris8.fr}	0,016	0,016	1,000	85
24	{}	=> {link}	0,010	0,010	1,000	56
(vii) REGLES PAGES DE PREMIER ACCES FREQUENT TRI PAR CONF (CONTRAINTE → ENSEMBLE PAGES PREMIER ACCES & LONGUEUR 1)						
1	{}	=> {univ-paris8.fr}	0,240	0,240	1,000	8 640
2	{}	=> {-FORMATIONS-}	0,106	0,106	1,000	3 812
3	{}	=> {-INSCRIPTIONS-91-}	0,060	0,060	1,000	2 159
4	{}	=> {Masters}	0,048	0,048	1,000	1 727
5	{}	=> {Licences}	0,020	0,020	1,000	735
(viii) REGLES PAGES FREQUENTE DE NAVIGATION INTERNETRI PAR CONF (CONTRAINTE → ENSEMBLE DE LONGUEUR 2 ET PLUS)						
1	{Premiere-inscription-a-Paris-8,3290}	=> {-INSCRIPTIONS-91-}	0,013	0,830	5,981	230
2	{Formation-continue}	=> {-FORMATIONS-}	0,016	0,732	2,204	292
3	{Licences-2015-2016}	=> {-FORMATIONS-}	0,059	0,717	2,160	1 061
4	{Licences-2015-2016,Masters}	=> {-FORMATIONS-}	0,034	0,687	2,069	615
5	{UFR-instituts-et-departements}	=> {-FORMATIONS-}	0,029	0,682	2,054	526
(ix) REGLES PAGES FREQUENTES DE FIN DE SESSIONS TRI PAR CONF (CONTRAINTE → ENSEMBLE PAGES FIN SESSIONS & LONGUEUR 1)						
1	{}	=> {-INSCRIPTIONS-91-}	0,097	0,097	1,000	3 502
2	{}	=> {UFR-instituts-et-departements}	0,067	0,067	1,000	2 421
3	{}	=> {Informations-sur-la-mobilisation-en-cours-a-l-Universite}	0,046	0,046	1,000	1 648
4	{}	=> {Licences-2015-2016}	0,044	0,044	1,000	1 601
5	{}	=> {Premiere-inscription-a-Paris-8,3290}	0,043	0,043	1,000	1 564

La **Sous-Table V.6.ii** présente un tri par support de la sous-table précédente (**Sous-Table V.5**), i.e., extraction de règles pages externes exclus. Elle est plus utile pour une adaptation de la structure du fait qu'une implication entre deux items/pages internes avec un indicateur de confiance important comparé à l'extraction générique dont la seule règle impliquant deux items contient une page externe avec un indicateur de confiance faible (**Sous-Table V.5.ii**).

La **Sous-Table V.6.iii** qui présente une extraction de règles sur les ensembles de plus de deux items/pages sans prise en compte de celles externes exclues ; confirme d'une part, l'utilité de la mise en avant de la page {UFR-instituts-et-departements}, et d'autre part, fait ressortir, au-delà des règles évidentes (*page d'accueil, formation, master, licence*), une nouvelle règle a l'indicateur de support très proche de ceux de l'extraction générique avec un indicateur de confiance plus important, i.e., { Premiere-inscription-a-Paris-8,3290,univ-paris8.fr }=>{-INSCRIPTIONS-91-} (**Sous-Table V.5.iii. ID.4**).

La **Sous-Table V.6.iv** qui porte sur les résultats en règles afférentes aux pages fréquemment visitées par sessions (*à ne pas confondre avec les fréquences absolus et relative des pages visitées*) montre que cette extraction stratégique sur l'échantillon des items de longueur 1 est plus fiable que celle de l'extraction générique, et ceci du fait que l'ensemble des items de navigation évidente figure parmi les 5 première règles y compris un item évident qui n'était pas identifiable lors de l'extraction générique tant par confiance que par support {-INSCRIPTIONS-91-} (**Sous-Table V.6.iv.ID.3**). L'ensemble des 5 règles sont portées par des indicateurs de support et de confiance acceptables. Ceci permet au concepteur de vérifier si sa perception de la navigation des utilisateurs à travers la structure conçue est adaptée aux habitudes de navigation identifiées par les règles extraites, et de ne pas être induit en erreur par la considération de l'item {-INSCRIPTIONS-91-} comme peut évident du fait que l'extraction générique ne le ressorte pas parmi les premières règles.

La **Sous-Table V.6.v** présente l'extraction de règles triées par confiance et limitée à l'ensemble de longueur 1 contenant seulement les URIs externes référent à notre site. Cette information permet d'avoir un aperçu sur la visibilité sur le Web des URIs structurant le site. Cette information est utile au titre de l'adaptation de la structure du site pour deux raisons :

- Elle permet de voir que le principal référent au site sont les moteurs de recherche. La limitation de l'extraction aux seuls URIs externe au site permet, aussi, de voir plus que le moteur de recherche dominant comme c'est le cas dans l'extraction générique. Aussi, elle nous renseigne sur le fait que seuls les sites relevant de Paris 8 figurent parmi les référents au site, e.g., {fp.univ-paris8.fr}.
- Les informations de cette extraction sont utiles pour l'interprétation des extraction suivantes, i.e., les pages externes référents au site en dehors des moteurs des recherches, les pages de premier accès fréquent au site.

La **Sous-Table V.6.vi** présente l'extraction de règles triées par confiance et limitée à l'ensemble de longueur 1 contenant seulement les URIs externes référent à notre site hors moteur de recherche connus. Cela vise à identifier les sites référents à notre site pour avoir un aperçu sur sa position au sein de la structure du Web, i.e., hub ou autorité. Cette information peut être utile le cas échéant de besoin d'adapter la structure interne des pages du site en termes de pages hub et autorité. On a procédé à plusieurs itérations d'extraction après chaque exclusion de moteurs de recherche identifiés. Reste que les résultats d'extraction font ressortir toujours d'autre engins qui sont souvent des moteurs, portails de recherche et/ou moteur, crawler d'indexation. Il s'agit de résultats à analyser au cas par cas dans la perspective d'identification d'intrusion et autre malware. Le deuxième site référent hors moteurs de recherche identifié est un site relevant de Paris 8, i.e., {www-artweb.univ-paris8.fr} (**Sous-Table V.5.vi.ID.17**).

La **Sous-Table V.6.vii** présentant l'extraction de règles limitées à l'ensemble d'items de longueur 1 relatives aux pages de premiers accès au site, confirme la fiabilité de l'extraction stratégique car les pages évidentes de premier accès figurent parmi les 5 premières règles triées par indicateur de confiance, ce qui n'était pas le cas pour l'extraction générique ou l'item { -INSCRIPTIONS-91- } n'y figurait pas. Elle confirme, aussi, la nécessité d'un nettoyage des URIs du site car l'item {Licences} étant le même que celui { Licences-2015-2016} en termes de page désignée fait que les indicateurs de confiance d'un seul item soient partagés, ce qui réduit son classement au sein des règles découvertes.

La **Sous-Table V.6.viii** présentant les résultats de l'extraction limité aux ensembles d'items de longueur 2 et plus triés par confiance vise l'obtention de règles

quant à la navigation interne indépendamment des pages de premier accès et celle de site référent. C'est cette navigation qui est le plus représentative de la structure du site et son adaptation aux habitudes de navigation des utilisateurs. Les résultats de cette extraction nous renseignent d'une structure de navigation plus pertinente que celle évidente (*page d'accueil, formation, licences, masters*) mise en avant par l'extraction générique. Dans la structure identifiée par l'extraction stratégique les items {Premiere-inscription-a-Paris-8,3290 ; INSCRIPTIONS-91 ; Formation-continue ; UFR-instituts-et-departements ; Formation ; Master ; Licences} dominent les 5 première règles. Ainsi, en excluant les pages de premier accès fréquent on obtient une information pertinente quant à la structure de navigation interne par l'identification de deux nouveaux items, i.e., {Formation-continue & Premiere-inscription-a-Paris-8,3290}, et ceci en plus de la confirmation de ceux de {UFR-instituts-et-departements et INSCRIPTIONS-91}, qui ont été aussi mises en avant par l'extraction stratégique (**Sous-Table V.6.i. ID.3**).

La **Sous-Table V.6.ix** présente les résultats de l'extraction de règles appliqué aux ensembles d'items de longueur 1 portant sur les pages de fins de sessions. Elle vise l'obtention de connaissance utile sur la structure de navigation des utilisateurs pour affiner celles déjà obtenu, à savoir obtenir une information la plus affinées que possible sur la structure de navigation. Un item fréquent en amont, au sein et en aval de la structure confirmera la nécessité de son indexation à côté de la page d'accueil sur le moteur de recherche, sur la page d'accueil, ou à la rigueur sur la page d'accueil sans avoir besoin de le chercher dans un menu déroulant. A cet égard, il s'agit des items {INSCRIPTION ; UFR-instituts-et-departements ; Premiere-inscription-a-Paris-8,3290 ; Licences} qui figure, encore une autre fois, parmi les cinq premières règles avec des indicateurs de support et de confiance très significatives.

Des deux extractions : De l'analyse et les interprétations sur la structure de navigation et la structure du site, il est à signaler une structure de navigation particulière qui ne semble pas être évidente, i.e., l'ordre d'implication entre les items/pages naviguées tant sur le plan des règles évidentes {page d'accueil, formation, master, licence, Inscription} que ceux découverte grâce à notre extraction stratégique impliquant, en plus, les items/pages {Instituts et IFR et département, Première Inscription, Formation continue}.

Au contraire d'un ordre évident respectif, e.g., {Formation} => {Licence | Master | Inscription | ...}, i.e., pages génériques/branches vers pages spécifiques/feuilles en termes de contenu ; la structure de navigation dans sa globalité affiche un ordre inverse, e.g., {Licence | Master | Inscription | ...} => {Formation}. Une telle structure de navigation démontre que :

- Il s'agit d'utilisateurs ayant accédé au site via les liens des pages {Licence | Master | Inscription | ...} indexées directement avec le lien de du DN/Page d'accueil du site sur les moteurs de recherche ; puis au à la page {Formation}.
- Les pages indexées directement sur le moteur de recherche, notamment, celles de {Licence | Master | Inscription | ...} servent, indépendamment de la nature de leur contenu, comme page d'accès au site, puis à la page {Formation}.
- La tendance de la structure de navigation des pages de premier accès, d'accès fréquent, et celles de navigation interne ; est centrée sur la page {Formation} qui figure toujours dans les deux premières règles dont les extractions ayant porté sur les sous-échantillons d'ensemble incluant cette page.

- ***Recommandation pour l'optimisation de la structure***

Pour une adaptation de la structure du site à la lumière de la connaissance extraite des règles de la structure de navigation des utilisateurs, il est à retenir ce qui suit :

- En premier lieu, au titre d'une adaptation sans remettre en cause la structure générale du site, nous recommandons le placement de liens vers les pages Licences, Master, Instituts UFR et Département, Inscription, Formation Continue sur la page d'accueil ou sur celle de Formation.
- En deuxième lieu, prévoir une restructuration générale du site, en sorte que les liens des pages indexées sur le moteur de recherche à côté de la page d'accueil ne figurent pas sur la page d'accueil.
- Enfin, il est prendre en considération que tout modification de la structure est à effectuer en cohérence avec le contenu des pages sur lesquelles les liens seront placés. S'agissant de recommandations d'adaptation basées sur une fouille et extraction de connaissance de données de journalisation, la validité et la fiabilité dépendent aussi de la fouille et des connaissance à priori sur le contenu [26], [30], [31], [33]–[35], [69], [70], [116].

V.2.3. Optimisation des chemins de navigation

1) Objectifs

La fouille des données de journalisation pour l'optimisation de la navigation des utilisateurs sur le site vise à leur permettre de trouver et atteindre leurs contenus d'intérêt via les chemins les plus courts possible. A cet égard, raccourcir des chemins de navigation réduit les abandons/sorties des utilisateurs du site sans atteinte de leurs objectifs, et augmente ainsi les possibilités de réaliser des transactions. Ainsi, optimiser les chemins de navigation sur un site, c'est permettre à l'utilisateur d'atteindre son contenu d'intérêt via le moindre nombre de clics possible, et ceci pour qu'il ne se disperse pas, ne perde pas son temps et n'abandonne pas le site pour aller voir ailleurs.

2) Techniques

Les techniques de fouille utilisées pour l'optimisation des chemins de navigation sont les règles/motifs séquentielles. Différemment des règles d'association portant sur les motifs de co-occurrences fréquents, un(e) motif/règle séquentielle est une règle d'association portant sur l'ordre d'occurrence, i.e., les motifs séquentiels fréquents.

Une règle séquentielle est une implication de l'ordre de : $X_{t_n} \rightarrow Y_{t_{n+1}}$

Définition 1. Etant donnée :

- $I = \{ i_1, i_2, \dots, i_m \}$, où (I) est un ensemble de (m) éléments/pages Web (i) ;
- $T = \{ t_1, t_2, \dots, t_n \}$, où (T) est un ensemble de (n) transactions/sessions (t) , horodatées (t_n) ;
- $X = Kt_n$, où (X) est un sous-ensemble (K) de (m) éléments/pages Web (i) de (I) objet de (n) transactions/sessions de (T) ;
- $Y = K't_n$, où (Y) est un sous-ensemble (K') de (m) éléments/pages Web (i) de (I) objet de (n) transactions/sessions de (T) ;

Une règle séquentielle de l'ordre de : $Page X_{t_n} \rightarrow Page Y_{t_{n+1}}$ mesure la fréquence de l'ordre d'occurrence de la séquence $\langle (X_{t_n}) - (Y_{t_{n+1}}) \rangle$ dans un ensemble (T) de (n) de transaction/sessions (t) , horodatées (t_n) à travers l'indicateur de support.

L'indicateur de Support (S) qui mesure la fréquence relative des transactions portant sur des séquences $S = \langle (X_{t_n}) - (Y_{t_{n+1}}) \rangle$ par rapport au nombre total des

transactions/sessions, estimant ainsi la probabilité séquence $P(< (X_{t_n}) - (Y_{t_{n+1}}) >)$ dans une transaction, comme suit :

$$S < (X_{t_n}) - (Y_{t_{n+1}}) > = \frac{\sum n t_n (< (X_{t_n}) - (Y_{t_{n+1}}) >)}{\sum n t_n} = P(< (X_{t_n}) - (Y_{t_{n+1}}) >)$$

Une variété d'algorithmes et méthodes d'extraction de règles d'association sont proposés dans la littérature [97], [97], [117]–[124], dont on peut citer des exemples par méthodes :

- Méthodes de hachage en largeur, e.g., GSP, SPAD.
- Méthodes de hachage en profondeur, e.g., PSP, PREFIXSPAN, SPAM.
- Méthodes pour motifs fermés, e.g., CLOSPAN, BID.
- Méthodes pour motifs sous contraintes d'horodatage, d'items, de longueur.
- Méthodes pour motifs incrémentaux, e.g., ISE, ISM, IUS, FASTUP, KISP.

Comme pour les seuils d'intérêt en règles d'association, les seuils des règles de motifs séquentiels sont définis en fonction de du contexte et de l'objectifs de l'analyse, de la nature des données. Aussi, définir une stratégie d'extraction au-delà de celle générique permettra l'extraction de règles plus représentative dans le cas de données massive et fortement hétérogène de par la diversité de l'objet représenté par les données.

3) Illustration

- **Paramètre d'extraction des motifs**

Notre stratégie d'extraction utilisant l'algorithme SPAD a été déroulée via les étapes et paramètres suivants :

- Une extraction générique sur items avec un support maximum de 0.01, appliquée au notre base de données (*USE_TAB*), pour avoir un aperçu global sur les motifs séquentiels fréquent de navigation.
- Identifier les motifs séquentiels fréquentes de la longueur à optimiser en les ramenant au seuil prédéfini de longueur optimale de navigation.
- Définition des raccourcis possibles en fonction du contenu des pages à soustraire de la séquence et celle à lier directement.

- *Analyse des résultats et interprétation*

La **Table V.7** présente les résultats de l’extraction de motifs avec un support maximum de 0,01 ayant généré 75 règles. La table contient un échantillon illustratif qui montre que les motifs les plus fréquents de supports relativement signifiant (*Support > 10%*) sont les sous séquences de longueur 1. Le premier motif d’une longueur de plus de 2 intervient à la sixième position parmi le classement de l’ensemble des règles avec un support faible (*Support < 10%*). Les autres motifs de la même longueur sont plus loin. Le premier motif d’une longueur de plus de 2 est à la 25^{ème} position avec un très faible support (*Support = 2,9%*).

Table V-7 Résultats d’extraction de motifs séquentiels de navigation

ID	SOUS-SEQUENCE	SUPPORT	EFFECTIF
01	(univ-paris8.fr)	0,302	10 905
02	(-FORMATIONS-)	0,292	10 534
03	(-INSCRIPTIONS-91-)	0,208	7 507
04	(Masters)	0,154	5 582
05	(Licences-2015-2016)	0,094	3 426
06	(univ-paris8.fr)(-FORMATIONS-)	0,093	3 358
25	(univ-paris8.fr)(-FORMATIONS-)(Licences-2015-2016)	0,029	1 051

La **Table V.8** présente les fréquences relatives par utilisateur des longueurs de chemins de navigation, à ne pas confondre avec le paramètre de la statistique descriptive. La fréquence calculée a été obtenue en appliquant l’algorithme d’extraction aux longueurs des séquences de navigation, ce qui considère, le cas échéant d’utilisateurs authentifiés la parcourt de la même longueur par le même utilisateur plusieurs fois comme donnée redondante, à considérer une seule fois dans le calcul du support. Les longueurs les plus fréquentes sont celles, respectivement, de navigation de 2 (*Support = 53%*), 3 (*Support = 25%*) et 4 pages (*Support = 10%*). Il est à rappeler qu’au vu de l’objectif de l’analyse que l’optimisation des chemins de navigation, les sessions de longueur 1 ne sont pas prises en compte.

Ainsi, la fréquence en termes de longueur de chemins n’est pas reflétée par celles des séquences de motifs car les chemins portent sur des motifs différents. Comparé avec les analyses de la structure sur la base des règles d’association, on peut affirmer que le

fait d'indexer des liens de pages de contenu à côté de la page d'accueil a aussi un impact d'optimisation de la navigation par raccourcissement des chemins.

Table V-8 Fréquence relative des longueurs des chemins de navigation

ID	LONGUEURS	SUPPORTS	EFFECTIF
1	(2)	0,533	19 226
2	(3)	0,253	9 137
3	(4)	0,104	3 761
4	(5)	0,053	1 903
5	(6)	0,027	989
6	(7)	0,014	504

- ***Recommandation pour l'optimisation de la navigation***

De ce qui précède, nous pouvons conclure que la proportion des séquences longues de navigation n'est pas significative d'autant plus que même celles de longueurs plus de 2 portent sur des motifs différents. La significativité faible des motifs longs peut donner lieu à la conclusion que les chemins de navigation sont optimisés. Comme aussi, dans une perspective perfectionniste pour un site commercial, il est envisageable de se baser sur une évaluation relative des supports et des effectifs et considérer que certains chemins de longueur plus de 2 sont à optimiser/raccourcir sur la base de connaissance a priori et en fonction de la nature des contenus du site.

Pour raccourcir un chemin de navigation, il faut en premier lieu vérifier s'il s'agit réellement d'absence de liens entre les pages à lier de la séquence à raccourcir, puis prendre en considération le contenu des deux pages à lier. Le cas échéant de l'existence de liens mais qui ne sont pas utilisés, la question devient un problème de stratégie de recherche et significativité de l'information autour du liens « Information scent foraging » [33].

Au titre de notre site, et sur la base de la connaissance a priori issue des règles d'associations, nous pouvons recommander de raccourcir les sous-séquences liant la page d'accueil aux formations Licence ou Master passant par la page formation, e.g., (univ-paris8.fr) - (-FORMATIONS-) - (Licences-2015-2016) à raccourcir pour devenir → (univ-paris8.fr) - (Licences-2015-2016) (**Table V.7.ID.25**).

V.2.4. Optimisation du trafic

1) Objectifs

L'objectif de l'optimisation du trafic est d'économiser la bande passante du serveur et le réseau en général afin de réduire le temps de réponse aux requêtes des utilisateurs. C'est le rôle des systèmes de mise en cache sur les serveurs intermédiaires (*proxys*) et les clients (*caches navigateurs*) des pages Web fréquemment demandées. La mise en cache sur les clients est effectuée par le navigateur pour les pages déjà consultées. Ainsi, les pages mises en cache sont délivrées à partir des serveurs intermédiaires et des agents-navigateurs, ce qui répartie le trafic sur le réseau et optimise les flux.

2) Techniques

La mise en cache sur les clients par les agents-navigateurs consiste en la conservation dans le cache navigateur des pages déjà visitées jusqu'à leur expiration. Le temps d'expiration est défini par le concepteur en fonction de l'importance de la page pour les besoins d'analyse, la capacité du serveur du site et la fréquence de sa consultation et son impact sur le trafic. La mise en cache sur proxy est basée sur des techniques de prédictions de la navigation des utilisateurs, i.e., Chaîne de Markov, Règles d'Association, Règles Séquentielles, etc. Les pages prédites sont donc stockées et chargées à partir des proxys. Par conséquent, les pages à contenu et structure dynamique sont moins exposées et complexes à mettre en cache[26], [30], [31], [33], [125].

Les pages et séquences de navigation dont les règles d'association et séquentielles sont portées par des indicateurs de support, de confiance et de lift significatifs sont utilisées pour la prédiction et la mise en cache. En plus de la prédiction basée sur les règles de motifs fréquents (*Règles d'association et séquentielles*), les techniques de modélisation de dépendance, à l'exemple de la Chaîne de Markov, sont utilisées pour la modélisation du comportement de navigation des utilisateurs sur la base de l'historique de navigation et servir de base de prédiction, notamment au niveau des proxys [125]–[127].

Etant donnée (S) un espace d'état, (A) matrice de probabilités de transitions et (λ) la probabilité de distribution initiale. Une chaîne de Markov est la probabilité de transition (A) d'un état ($S_{\lambda_{t-1}}$) de distribution (S_{λ_t}), dénotée $P(S_{\lambda_t}) = (S_{\lambda_{t-1}})A$.

3) Illustration

- *Données et paramètres*

Notre illustration consiste en la simulation de mise en cache sur nos données de journalisation expérimentale, i.e., journalisation du site Web Paris 8. La simulation de mise en cache consiste en l'élimination des requêtes de certaines pages, objet de prédiction, de notre base de données (USE_TAB) pour obtenir cinq tables, à savoir :

- USE_TAB_BIS, du même format et attributs que USE_TAB, i.e., sessions utilisateurs au format transactionnel ; mais dont les transactions contiennent seulement les **données** de sessions non **éliminées** par la simulation de la mise en cache.
- STA_TAB_BIS, table d'agrégation de USE_TAB_BIS, du même format et attributs que STA_TAB, i.e., descriptif d'agrégations statistique des sessions utilisateurs ; mais qui porte seulement sur les **données** de sessions **non éliminées** par la simulation de mise en cache.
- USE_CAC_TAB, du même format et attributs que USE_TAB, i.e., sessions utilisateurs au format transactionnel ; mais dont les transactions contiennent seulement les **données** de sessions **éliminées** par la simulation de mise en cache.
- STA_CAC_TAB, table d'agrégations de USE_CAC_TAB, du même format et attributs que STA_TAB, i.e., descriptif d'agrégation statistique des sessions utilisateurs ; mais qui porte seulement sur les **données** de sessions **éliminées** par la simulation de mise en cache.
- Les données des tables d'agrégation statistiques serviront la présentation chiffrée de l'impact de mise en cache sous forme d'une table synthèse (CAC_IMP_TAB) sur son impact sur l'activité de navigation avant et après la simulation de la mise en cache.

Vu, d'une part, que nos données de journalisation n'ont pas fait l'objet de récupération et complètement des données des requêtes mise en cache réellement du fichier log source ; et d'autre part, qu'il s'agit d'illustration ; nous avons opté pour la simulation d'une mise en cache peu conséquente. Ainsi, notre simulation de mise en cache a été basée sur une seule règle d'association dont le support est relativement faible, i.e., { Premiere-inscription-a-Paris-8,3290,univ-paris8.fr }=>{-INSCRIPTIONS-91-} et sans prise en compte de l'item de la page d'accueil { univ-paris8.fr }.

- *Résultats et interprétations*

La **Table V.9** présente un descriptif chiffré avant et après la mise en cache en termes de flux de données global exprimé en octet/seconde (GLO_FLU), volume global de données (GLO_SIZ) en octet, temps de visite global (GLO_VIE_TIM) en seconde, nombre global de clics (NBR_GLO_CLI), nombre de clics par utilisateur (NBR_CLI_USE), et nombre d’utilisateurs (NBR_USE). Le descriptif « Avant mise cache » reflète l’activité avant l’élimination des requêtes des pages de navigation prédites sur la base de notre règle d’association { Premiere-inscription-a-Paris-8,3290,univ-paris8.fr }=>{-INSCRIPTIONS-91-}. Par contre, le descriptif « Mise en cache » traduit l’activité des requêtes mise en cache, à savoir celle portant sur les pages prédites, stockées et consultées à partir du proxy. Ainsi, le descriptif « Après mise en cache » quantifie le reste d’activité sur le serveur. Enfin, le descriptif « Optimisation » souligne le gain sur le serveur en activité transférée vers le proxy, à savoir l’écart d’activité sur le serveur avant et après, respectivement, la mise en cache.

Table V-9 Impact de Mise en Cache [CAC_IMP_TAB]

	NBR_USE	NBR_CLI_USE	NBR_GLO_CLI	GLO_VIE_TIM	GLO_SIZ	GLO_FLU
Avant mise cache	36 065	3	113 919	6 857 296	387 194 893	56
Mise en cache	6 261	2	10 012	462 685	35 085 302	76
Après mise en cache	29 804	2	103 907	6 394 611	352 109 591	55
Optimisation	- 6 261	- 2	- 10 012	- 462 685	- 35 085 302	- 1

Les résultats présentés montrent un gain (*Optimisation*) sur le serveur d’une seconde de flux global (GLO_FLU). Ce gain et le résultat de la réduction (*Optimisation*) du volume des données/pages (GLO_SIZ) de 35 085 302 octets. Ces données, initialement délivrées par le serveur, ont été mises en cache et délivrées par le proxy. De même, le temps afférent aux données délivrées par le serveur est réduit (*Optimisation*) du temps des visites (GLO_VIE_TIM) des pages concernées, i.e., 462 685 secondes. Ce gain d’activité implique une économie (*Optimisation*) sur le serveur de 10 012 clics d’utilisateurs (NBR_GLO_CLI) dont des requêtes de sessions de 6 261 utilisateurs (NBR_USE) ont été servies dans leur totalité par le proxy, ce qui a réduit (*Optimisation*) le nombre de clics par utilisateur (NBR_CLI_USE) sur le server de 3 à 2.

V.3. Limites, problème et proposition

V.3.1. Hypothèse des limites et définition du problème

Une optimisation par dimension porte sur :

- Le mise en cache par prédiction des pages fréquemment (*proxy*) ou déjà (*navigateur*) visitées, ce qui d'une part, réduit (*optimise*) le trafic/flux sur le serveur en le répartissant sur le réseau (proxys) et les clients (*navigateurs*), et d'autre part, améliore les temps de réponse ;
- Le raccourcissement des chemins de navigation en fonction des séquences de navigation fréquentes, ce qui permet aux utilisateurs d'atteindre leurs contenus d'intérêts via les chemins les plus courts possibles ;
- L'adaptation de la structure du site aux structures de navigation fréquentes des utilisateurs (*habitudes de navigation*), ce qui permet aux utilisateurs de naviguer plus de contenus d'intérêt lors d'une session.

Par contre :

- Les visites des utilisateurs des pages mises en cache ne seront plus journalisées sur le serveur. Il s'agit du transfert d'une partie importante de l'activité de navigation hors serveur, ce qui réduit la capacité, la pertinence et la significativité de l'analyse des activités des utilisateurs, vu que c'est les pages les plus fréquemment visitées, en l'occurrence les plus importantes, qui sont éligibles à la prédiction et la mise en cache ;
- La reconfiguration de la navigation en termes de longueurs de chemins et de structure autour des motifs fréquents y afférents n'est pas sans conséquences sur le trafic/flux. Cette reconfiguration aura un impact sur le flux en fonction des tailles des nouvelles pages et les temps de navigation objets des motifs fréquents, à savoir que dans le contexte de Web dynamique la tailles de pages uniques varient en fonction des adaptations et recommandations dynamiques.

A titre d'exemple :

- Le raccourci des chemins de navigation dans l'intérêt des utilisateurs peut générer un fort flux sur le réseau par l'effet de l'aiguillage de la navigation et sa concentration dans le temps (*Conflit d'optimisation*) ;

- L'adaptation de la structure peut influencer sur le flux du trafic de par la taille des ressources et la longueur des nouveaux chemins de navigation générés (*Conflit d'optimisation*).
- Le raccourci des chemins pour l'intérêt de l'utilisateur, le cas échéant d'influence vers le haut sur le flux, diminuera le temps de réponse, ce qui n'est pas dans l'intérêt de l'utilisateur (*Conflit & Contradiction d'optimisation*).

Ainsi, on peut en déduire que l'impact d'une optimisation par dimension peut s'avérer conflictuel et contradictoire par rapport aux intérêts des différentes parties prenantes, i.e., concepteur/propriétaire (*analyse*), l'artefact (*site web, serveur, réseau*), l'utilisateur. Une optimisation d'une dimension au détriment d'une autre est qualifiée dans la littérature relative à l'évaluation IHM d'optimisation/relation parasitique par opposition à une optimisation/relation symbiotique sensible aux à l'équilibre entre les différentes dimensions d'intérêts des parties prenantes [14]–[17], [128], [129].

Problème. En résumé, l'optimisation par dimension pose le problème du contrôle de son impact symbiotique/parasitique, en l'occurrence, prédire et contrôler/filtrer les conflits et contradictions d'optimisation.

V.3.2. Illustration des limites et proposition

1) Analyse générale orientée sessions de navigation

- *Mise en cache & impact sur la capacité d'analyse*

Pour ce qui est de l'optimisation du trafic au détriment de l'analyse, nous rappelons les résultats de la mise en cache présentés en **Table V.9** ou cette optimisation a fait en sorte qu'une partie conséquente de l'activité a été transférée hors serveur, i.e., 10 012 clics ayant engendré un manque d'activité à analyser de 6 261 sessions/utilisateurs ; représentant, respectivement, 8 et 17 % de l'activité en termes de nombre de clics et utilisateurs.

- *La reconfiguration de la navigation & impact sur le flux*

La **Figure V.1** présente un aperçu global sur l'activité de navigation par sessions utilisateurs en termes d'une sélection d'indicateurs quantitatifs descriptifs de cette activité, i.e., nombre de clics et flux par utilisateur (**Grapshe 1 & 2**), le flux par nombre de clics/longueur de chemin et temps moyen de navigation/visite

des sessions utilisateurs (**Graphe 3 & 4**).

Les utilisateurs qui ont parcouru différentes longueurs sur le site (**Graphe 1**) ont généré une variation croissante et décroissante de flux (**Graphe 2**), indépendamment de la longueur/nombre de clics qu'ils ont parcouru (**Graphe 3**), dont la seule tendance évidente de corrélation est celle négativement liée au temps de visite. A cet égard, nous retenons que des chemins courts peuvent générer des flux supérieurs à ceux de chemins plus longs. Ainsi, une longueur de chemin optimisée par raccourcissement peut augmenter le flux, à savoir parasiter la dimension « optimisation du flux ».

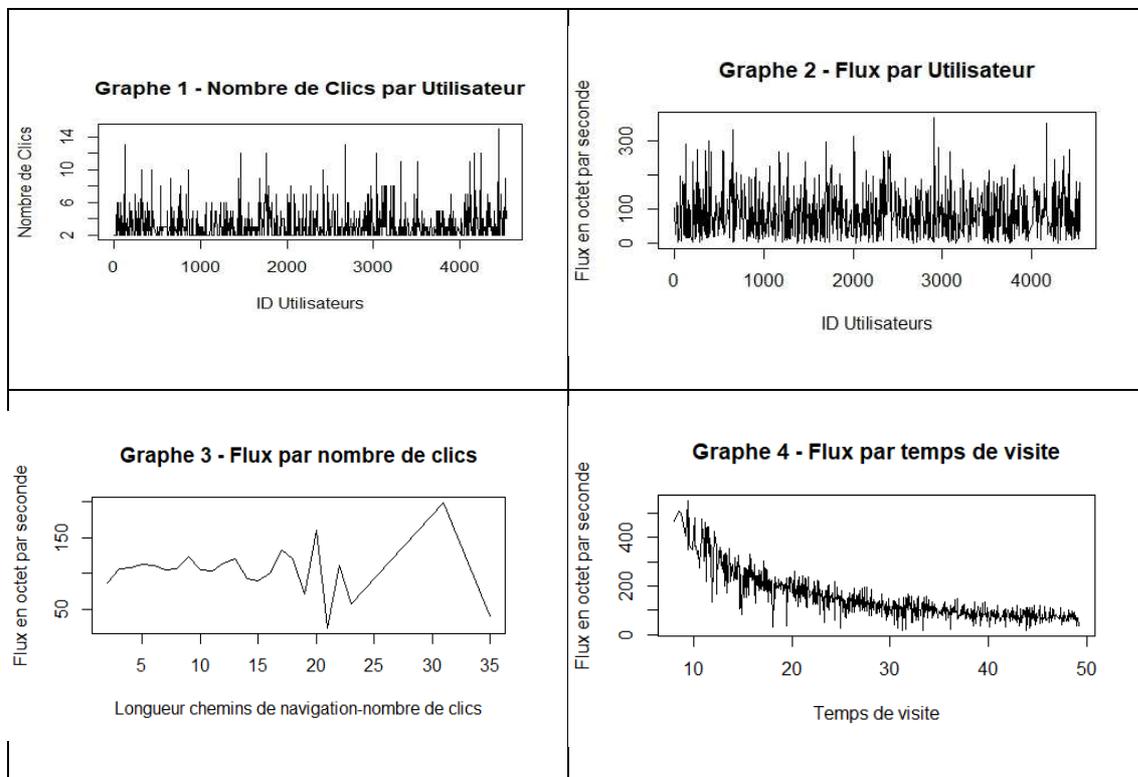


Figure V-1 Descriptif de l'activité de navigation

En résumé, des chemins courts de navigation peuvent tant réduire qu'augmenter le flux de trafic. Donc, le raccourcissement des chemins de navigation, pour l'intérêt des utilisateurs et l'artefact, n'est pas évident du fait qu'il peut induire une augmentation du flux de trafic et, aussi, par conséquent le temps de réponse.

Le temps de navigation étant le seul facteur avec tendance de corrélation en

rapport avec la taille des ressources, et vu qu’il dépend de l’utilisateur et qu’il n’est pas directement contrôlable ; l’impact contradictoire d’une optimisation (*chemins raccourci – flux et temps de réponse croissants*) ne peut être contrôlée.

La **Figure V.2** présente la matrice de corrélation (*coefficient de pearson*) des indicateurs flux (FLUX), temps des sessions (TIM), volume des données des sessions (SIZ) et la longueur des chemins parcourus (LEN_PAT) en nombre de clics. En couleur ;

- Les étoiles représentent l’estimation de la significativité en p-values ;

{“***” ; “**” ; “*” ; “.” ; “ ”} \Leftrightarrow {0, 0.001 ; 0.01 ; 0.05 ; 0.1 ; 1}

- Les courbes de tendance de distribution (*sur histogramme*) et de tendance de corrélation (*sur nuage de points*).

Une évaluation relative des coefficients et significativités des corrélations de la matrice traduit les constats de l’aperçu général.

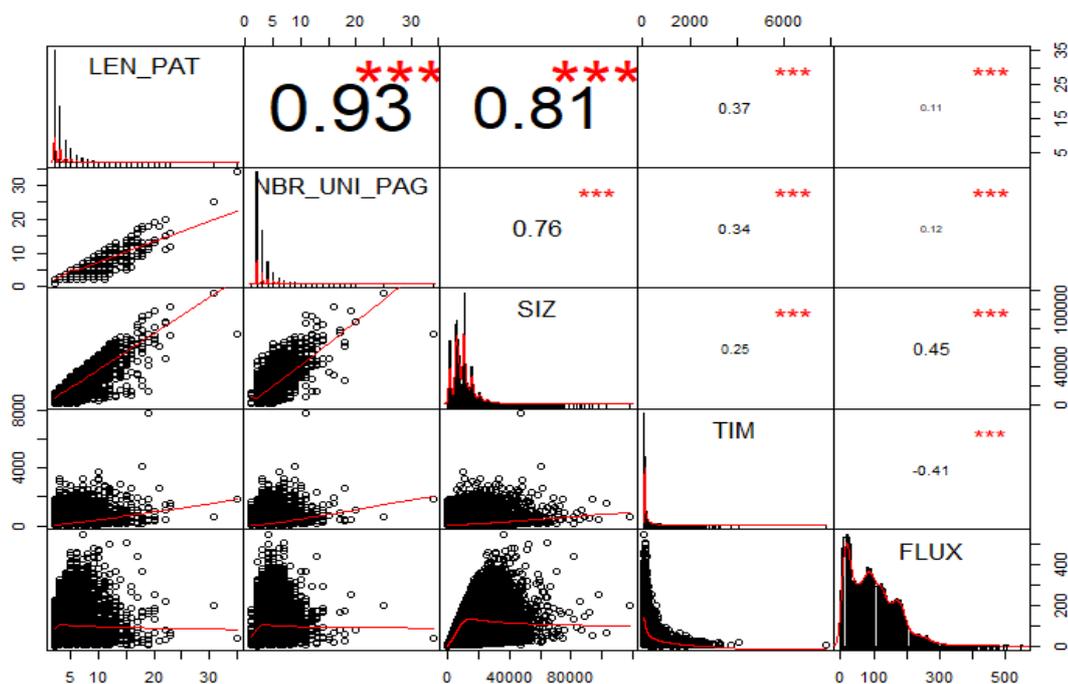


Figure V-2 Corrélations des indicateurs de l’activité de navigation par sessions

Les corrélations significatives en rapport direct avec le flux sont le temps (Corr ***: TIM, FLUX = -0,41) et le volume de données (Corr*** : SIZ , FLUX = 0,45), i.e., la croissance du flux quand les utilisateurs consultent des ressources/pages dans un

temps court. C'est pourquoi la forte corrélation de la longueur des chemins parcourus avec le volume des données (Corr ***: LEN_PAT , SIZ = 0,81) n'a pas de répercussion significative en terme de corrélation sur le flux (Corr ***: LEN_PAT , FLUX = 0,11). En effet cette forte corrélation reflète seulement la somme de la taille des sessions (*somme de la taille des pages consultées*) en fonction de sa longueur, i.e., l'évidence d'une session longue en termes du nombre de clics/pages consultées (*deux mesures différentes du même objet une continue et l'autre discrète*).

A ce stade d'analyse, le volume dans le temps des pages visitées demeure le seul indicateur significativement corrélé avec le flux. Puisque le facteur Temps de navigation ne peut être directement contrôlé ; l'éventualité de conflits d'optimisation et d'optimisation contradictoire, sa prédiction pour son **contrôle** ne peut être entreprise comme problème de fonction de régression.

2) Analyse détaillée orientée structures de navigation

L'analyse par sessions utilisateurs n'ayant pas fourni de solution pour le contrôle des conflits et contradictions d'optimisation ; nous avons procédé à une analyse détaillée centrée sur la configuration des chemins et structure de navigation. A cet égard, au lieu d'analyser la navigation sur la base des indicateurs statistiques des sessions utilisateurs (*Table V.3–Table Statistiques Page [STA_TAB]*), on a procédé à l'analyse de la configuration de la navigation sur la base des indicateurs statistiques par suffixe (*Table V.4–Table Statistiques Page [STA_STR]*), i.e., nombre de longueurs (NBR_UNI_LEN), de chemins (NBR_UNI_PAT), et de pages (NBR_UNI_PAG) uniques, volume moyen généré (MEA_SIZ), temps moyen de navigation (MEA_TIM_NAV), vitesse moyenne de navigation (MEA_SPE_NAV) et le flux moyen généré (MEA_FLU). Cette analyse vise la recherche d'indicateurs contrôlables de l'optimisation symbiotique via la perspective structure de navigation au lieu de la perspective session de navigation dont le seul facteur d'impact direct à forte corrélation significative était celui qui n'était pas contrôlable, i.e., le Temps.

La **Figure V.3**, au titre d'une évaluation relative des tendances, montre que les indicateurs d'impact direct sur le flux sont liés toujours au temps, en l'occurrence, l'évidence du volume dans le temps, i.e., le volume (Corr*** : MEA_SIZ_NAV , MEA_FLU = 0,67) et le temps (Corr*** : MEA_TIM_NAV , MEA_FLU = -0,32)

moyens par sessions, la vitesse de navigation (Corr*** : MEA_SPE_NAV , FLUX = -0,40).

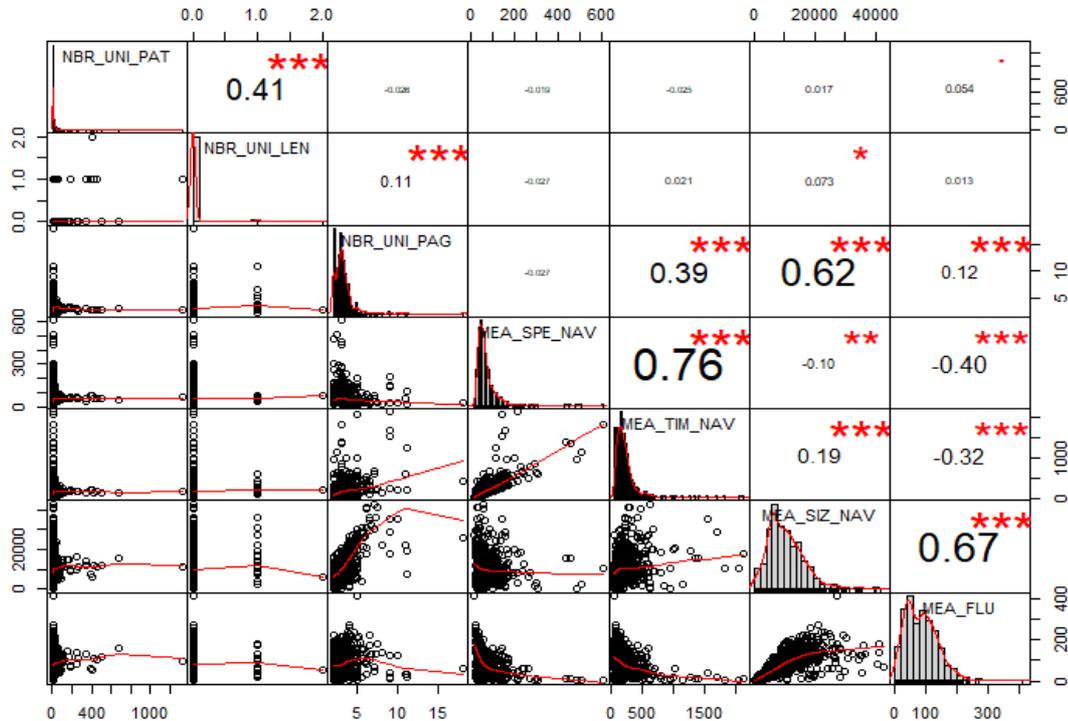


Figure V-3 Corrélation des indicateurs de l’activité de navigation par suffixe

Par contre deux indicateurs d’intérêt sont à prendre en considération au titre notre perspective :

- La corrélation du nombre de chemins et longueurs uniques (Corr*** : NBR_UNI_PAT , UNI_LEN = 0,41) qui nous renseigne sur la dépendance de ces deux descripteurs de structure d’une part, et d’autre part confirme l’existence de fortes régularités dans le Web autour de chemins de navigation courts [24], [29], [45], [92], [92], [100]–[104]. Il s’agit de régularités en termes de forte concentration autour des chemins courts et certains contenus parmi l’ensemble des contenus proposés par un site Web ; ce qui est le cas aussi de notre site, comme le montrent les distributions des longueurs de chemins et pages uniques (**Figure V.2 & Figure V.3**). L’absence de distribution gaussienne sera parmi les caractéristiques d’intérêt pour notre proposition.

- Le rapport entre le nombre de pages uniques et la moyenne de temps passé par page (Corr*** : NBR_UNI_PAG , $MEA_TIM_NAV = 0,39$), qui confirme notre hypothèse sur lien entre la diversité des structures de navigation et la répartition du flux dans le temps. Une diversité qui peut être compromise par une optimisation autour des structures et chemins de navigation fréquents et conduire à l'augmentation du flux au désavantage de l'utilisateur et de l'artefact.

Les corrélations des indicateurs de structure avec les indicateurs du temps ou du volume sont soit faibles sans significativité soit des évidences de deux mesures d'un seul objet de mesuré.

V.3.3. Proposition

L'objectif étant d'identifier des corrélations significatives entre les indicateurs des trois dimensions d'optimisation de l'usage Web, i.e., longueur des chemins de navigation (*nombre de clics par sessions*), structure de la navigation (*nombre de longueurs et pages uniques par fins de sessions*), et le flux du trafic (*le volume dans le temps*) ; les résultats d'analyse tant sur le plan de la navigation que sa structure n'ont pas révélé de corrélations d'intérêt au-delà de l'évidence « volume dans le temps », dont le facteur temps de sollicitation des ressources est hors de contrôle.

Les résultats obtenus sous cet angle, qui représente une formulation de la question comme problème de fonction de régression, certes permettent d'expliquer ou de suivre l'optimisation par dimension à même suivre et expliquer la convolution et le produit de l'optimisation des trois dimensions ; mais ne permettent pas de prédire pour contrôler/filtrer le produits symbiotique/parasitique de l'optimisation, pour les raisons suivantes :

- L'absence de tendance de corrélation liant l'ensemble des indicateurs ;
- Le facteurs « Temps », étant dépendant de l'utilisateur, ne peut être directement contrôlable ;
- Le volume (*la taille des pages*) étant variable dans un contexte de Web dynamique renvoie les faibles corrélations, à même relativement significatives, des indicateurs hors facteur « Temps » à des évidences de liens aléatoires de coïncidence d'une variation de volume avec une autre de longueur, éventuellement l'évidence de variables représentant deux mesures d'un seul objet.

Dans notre contribution nous proposons la reformulation de la question du contrôle de l'optimisation symbiotique/parasitique d'un problème de fonction de régression à un problème de classification, hors modèles de fonction de régression, pour une prédiction permettant le contrôle du produit d'optimisation.

Rappels sur la formulation du Problème et le choix de la solution prospectée.

- L'optimisation par dimension, i.e., chemins, structure et flux, pose le problème du contrôle de son impact symbiotique/parasitique, i.e., prédire pour contrôler/filtrer les propositions d'un système d'adaptation dynamique d'un site Web, susceptible d'engendrer des conflits et des contradictions d'optimisation par les raccourcis et adaptations des chemins et structure de navigation.
- L'exemple d'optimisation parasitique du cache sur la capacité d'analyse, des chemins de navigation sur le flux, d'une part, et d'autre part, l'absence de tendance de corrélation entre les indicateurs d'activité de navigation hormis le facteur hors contrôle « Temps », renvoie la question du contrôle de l'optimisation à un problème de classification hors modèles de régression.

Le contexte de Web dynamique et ces contraintes, i.e., structures, chemins de navigation et tailles des ressources/pages uniques dynamiques conduits par des agents-robots ; impose la prospection d'une méthode de prédiction, basée sur la classification hors modèles de fonction de régression, qui :

- Ne dépend pas des facteurs hors contrôle, en l'occurrence le temps de visite ;
- Soit applicable sans contraintes invalidantes d'applicabilité en termes de coûts dans un contexte de données massives, accentuées par le caractère dynamique automatisé et virale des modifications des chemins, des structures et des tailles de pages/URI uniques ;
- Est capable de produire des résultats pertinents et fiables.

A cet égard, la prospection d'une solution est orientée vers l'**apprentissage semi-supervisé**, hors modèles de fonction de régression, de par sa vocation de prédire à partir de peu de données labélisées, ce qui répond à la contrainte d'applicabilité en termes de coûts dans un contexte de données massives.

V.4. Une approche d'apprentissage semi-supervisé pour l'Optimisation Symbiotique

V.4.1. Données, indicateurs et instrument de mesure

Sur la base des données dont nous disposons, i.e., données de journalisation du site Web de notre université, et **les indicateurs** de navigation et de structure des sessions utilisateurs (**Table V.2 & 3**) - triés par leur ordre d'horodate dans le fichier de journalisation- **qui sont** :

- Les chemins parcourus (*les séquences de pages des sessions*) ; leurs préfix (*page de début de session*) et suffixes (*page de fin de session*) ; leur nombre de pages uniques et longueurs ;
- Les temps de visite, taille et flux moyens par page ;

Nous avons sélectionné :

- Comme attributs, d'observation statistique/individu, représentatifs de la navigation (*chemins et structures*), hors ceux liés à l'indicateur hors contrôle « le Temps », i.e., le chemin parcouru (*les séquences de pages des sessions*), son préfix (*page de début de session*), son suffixe (*page de fin de session*), le nombre de pages uniques du chemin, la longueur du chemin ;
- Pour chaque individu on a calculé trois indicateurs d'optimisation reflétant les trois dimensions de l'optimisation de l'usage Web (*chemin, structure, flux*) ; et avons défini un seuil objectif pour qu'ils soient labélisés **Optimisé (1)** ou **pas optimisé (0)**. Ainsi, si les trois dimensions sont labélisées **Optimisées (1)**, le chemin sera labélisé **Symbiotique (1)**, sinon **Parasitique (0)**.

La **Table V.10** représente un échantillon de notre jeu de données labélisées, qui comprend :

- Les attributs de classification des chemins de navigation, i.e., le chemin codifié et défini comme variable qualitative (PAT), son suffixe (SUFIX), son préfixe (PREX), le nombre de pages uniques qui le composent (UNI_PAG) et sa longueur (LEN_PAT).
- Les indicateurs d'optimisation de chaque dimension, i.e., la longueur du chemin (LEN_PAT) qui est repris comme indicateur de longueur (*le plus un chemin est long*

le moins il est optimisé) ; le rapport (*coefficient*) entre la longueur du chemin et le nombre de page uniques comme indicateur sur l’optimalité de la structure (STR_IND), i.e., plus de pages uniques signifierai que la structure permet à l’utilisateur de chercher son intérêt sans avoir à revoir des pages qu’il a déjà vu (*le moins est ce coefficient le plus la structure est optimisée*) ; et l’indicateur du flux (FLU_IND) (*le plus le flux est important, le moins il est optimisé*).

- La labélisation binaire de l’optimalité de chaque dimension, i.e., de la longueur (LEN_OL), la structure (STR_OL), le flux (FLU_OL) ; où le seuil d’optimalité a été défini sur la base des quartiles. Ainsi les valeurs inférieures/supérieures au deuxième/quatrième quartile sont considérées comme optimisées (**1**), sinon pas optimisées (**0**).
- La labélisation binaire de l’optimalité symbiotique du produit des trois dimensions qui indique si les trois sont optimisées (**1**) ou pas (**0**).

Table V-10 Données et instrument de mesure de l’optimalité symbiotique

PAT	SUFX	PREX	UNI_PAG	LEN_IND	STR_IND	FLU_IND	LEN_OL	STR_OL	FLU_OL	SYM_L
39,39, 590	590	39	2	3	0,66	235,62	1	0	0	0
28, 3716	3716	28	2	2	1,00	193,18	1	1	0	0
1040, 2524, 39	39	1040	3	3	1,00	99,90	1	1	1	1
2180, 2524, 2337	2337	2180	3	3	1,00	3,39	1	1	1	1
3716, 28, 2254	2254	3716	3	3	1,00	208,84	1	1	0	0

Il s’agit d’un instrument de mesure proposé à titre expérimental dont nous tenons à signaler que :

- Les seuils d’intérêts au titre d’un processus d’extraction, analyse et interprétation de connaissances dans une perspective exploratoire sont souvent objectifs. Par contre le cas échéant de connaissance a priori sur le domaine ou l’objet, ou le cas échéant de perspective ciblée, les seuils seront définis d’une manière subjective [40], [114], [115] ;

Il n’existe pas dans la littérature de modèles et instruments validés empiriquement concernant la symbiotique Homme-Machine[14]–[17].

V.4.2. Méthode, modèle et évaluation

La **Figure V.4** présente notre modèle pour le contrôle de l'optimisation de l'usage Web. Ce modèle traduit l'application de notre méthode d'apprentissage semi-supervisé pour la prédiction et le filtrage de l'optimisation symbiotique/parasitique. Le processus global se résume comme suit :

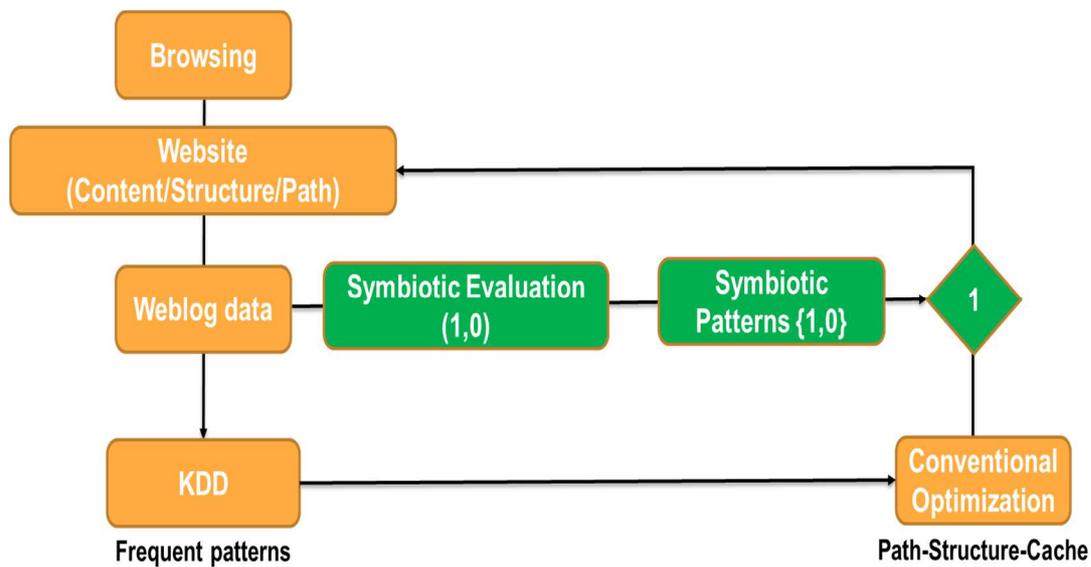


Figure V-4 Modèle de contrôle d'Optimisation Symbiotique/Parasitique

- Le processus d'optimisation par dimension (*conventionnelle*) fouille les données de la journalisation de l'activité de navigation (*Weblog data*) via les techniques appropriées (*Règles d'associations, motifs séquentiels, etc.*) pour l'extraction de règles et la découverte de motifs fréquents (*KDD*), et ceci dans la perspective d'optimiser les chemins et les structures de navigation. Dans le contexte d'un site Web dynamique, il s'agit, souvent, d'un processus automatisé d'optimisation dynamique dans un contexte de données massives.

- Notre processus de contrôle de l’optimisation fouille les mêmes données pour labéliser une partie des séquences de navigation (*peu de données*) sur la base de l’instrument de mesure (*seuils objectifs et/ou subjectifs*) ; et ceci pour construire un modèle de prédiction de la valeur symbiotique des séquences de navigation journalisées ayant servi le processus d’optimisation par dimension.
- Enfin, le processus de contrôle évalue les optimisations prévues par le processus d’optimisation conventionnelle (*beaucoup de données*), et ceci pour les filtrer en fonction du résultat de prédiction, afin de valider seulement les optimisations symbiotiques.

Nous tenons à signaler qu’il s’agit là de deux processus parallèles dans un contexte de Web dynamique générant des données massives, d’où l’intérêt de la méthode de classification semi-supervisée.

Pour tester la performance de prédiction de notre méthode on a choisi trois algorithmes à tester via la technique d’apprentissage supervisé collectif, i.e., NaiveBayes et BayesNet et J48 Tree via Collectif EM [130]–[135]. Les critères de notre choix se résument comme suit :

- Le rapport performance/coût et le caractère intelligible du processus et de ses sortants ;
- La compatibilité avec la nature de nos données, i.e., variables quantitatives et qualitatives, avec possibilité de traitement sans transformation, ce qui va de la réduction des coûts ;
- Sensibilité à l’interdépendance des variables et leurs distributions.

Concernant l’évaluation de notre méthode, et vu que nous ne disposons pas de référence d’évaluation, s’agissant d’une problématique n’ayant pas été traitée auparavant à partir de cette perspective ; nous avons adopté une démarche d’évaluation relative, et ceci en comparant les résultats avec ceux d’un classifieur basé sur :

- L’algorithme ZeroR de par sa vocation de référence de seuil d’invalidation éliminatoire des résultats de prédiction ;
- L’algorithme OneR de par sa vocation de référence de seuil minimum possible à prédire.

V.4.3. Démonstration, discussion et perspectives

La **Table V.11** présente les résultats du test, sous Weka 3.8.1, de notre méthode sur le même échantillon de journalisation que nous avons utilisé dans l'illustration des limites de l'optimisation conventionnelle (**Table V-1**). Les résultats sont exprimés en termes de précision (PRE), rappel (REC) et pertinence (ACC).

Table V-11 Résultats de prédiction et d'apprentissage

ALGORITHMES ET METHODE		PRE	REC	ACC
NaiveBayes	Classification semi-supervisée Collectif EM Validation croisée sur 25% Test sur 75%	0.783	0.750	75.031
BayesNet		0.749	0.750	74.968
J48		0.845	0.796	79.628
<i>OneR</i>		0.728	0.670	66.952
<i>ZeroR</i>		0.330	0.574	57.415

L'algorithme J48 Tree réalise les meilleurs résultats. Par contre, NaiveBayes et BayesNet sont moins coûteux en calcul. NaiveBayes, malgré son hypothèse de probabilité conditionnelle -face à la corrélation significative du nombre des pages uniques et les longueurs des sessions- donne de meilleurs résultats que BayesNet. Cela peut s'expliquer par le fait que le nombre de pages uniques, rapporté aux longueurs des sessions, dépend plutôt du cache que de la longueur des chemins, i.e., les chemins longs, le sont car leurs pages ne sont pas sujettes au cache de par leur temps d'expiration défini par le concepteur. Comme il peut s'agir, éventuellement, du paradoxe des performances de NaiveBayes malgré son principe de probabilité conditionnelle ; ce qui est sujet à argumentation dans la littérature [136], [137].

D'un point de vue général, les résultats montrent que notre méthode à travers les trois algorithmes est loin du seuil éliminatoire de ZeroR, et au-dessus du seuil minimum possible à prédire de OneR. Ainsi, sur la base de ces résultats obtenus à titre prospectif, nous pouvons envisager l'approfondissement de notre approche via trois perspectives susceptibles d'aboutir sur des résultats efficaces et fiables en termes de précision, rappel et pertinence de la prédiction du produit symbiotique, en l'occurrence :

- En premier lieu, exploiter la nature séquentielle de l'attribut chemin de navigation (*PAT*) en se basant sur les techniques de classification des séquences, dès lors que cet attribut représente une session, i.e., une séquence de navigation de pages Web. A cet égard, on envisage d'aboutir à une méthode de prédiction plus pertinente où le produit symbiotique sera prédit sur la base des classes des séquences de navigation (*PAT*). Le reste des variables, i.e., préfixe, suffixe, nombre de pages uniques et longueurs des chemins de navigation seront pris en compte comme attributs de profil de séquence. Il s'agit-là du même principe de classification des sessions **utilisateurs** en tenant compte de **leurs** profils ;
- En deuxième lieu, identifier le seuil optimal (*par rapport à ceux proposés dans la littérature*) de la proportion des données d'apprentissage et la sélection de l'échantillon (*représentativité par rapport à l'ensemble*) devant permettre la pertinence, la fiabilité et la stabilité des résultats d'un apprentissage semi-supervisé. Il s'agit-là d'exploiter le concept de fortes régularités Web [72], [102], i.e., la tendance en termes de longueurs et de contenus de navigation ; et ceci pour l'identification du seuil optimal et la sélection de l'échantillon par référence à l'indice d'entropie des modalités des attributs de profil et leurs fréquences. L'hypothèse étant **que sur un ensemble** de séquences de navigation dont l'ordre de classement chronologique du fichier Log source est préservé, à considérer comme classement neutre et aléatoire ; **un échantillon d'apprentissage** dont **la proportion** et **la sélection du contenu** sont tirées en fonction de **l'indice d'entropie** ; sera suffisamment **représentatif** de l'ensemble des données de journalisation ;
- Enfin, la prospection des algorithmes et des méthodes d'apprentissage semi-supervisé les plus adaptés à la nature de nos données et à nos objectifs en termes de pertinence rapportée aux contraintes d'applicabilité et de coûts dans un contexte de données massives.



Chapitre VII
Conclusion Générale

Les travaux de notre thèse ont porté sur l'analyse de l'interaction Homme-Machine sur la base des traces d'utilisation. Notre recherche a abordé l'interaction Homme-Machine à travers le cas de l'analyse de l'interaction avec le Web basée sur la fouille des données de l'usage, i.e., les données Log de serveurs Web dites données de journalisation. Nos contributions ont visé l'amélioration de la qualité et de la pertinence de trois tâches critiques, i.e., le nettoyage des données de l'utilisation du Web, leur structuration, et l'identification de motifs d'usage pour l'optimisation symbiotique de l'usage du Web.

Trois nouvelles approches liées aux tâches en question sont introduites, i.e., le nettoyage basé sur la structure de journalisation, la structuration centrée sur les flux de clics et l'optimisation basée sur les motifs d'usage identifiés par une classification semi-supervisée. Les résultats expérimentaux des méthodes proposées, comparés à ceux des méthodes actuelles, ont démontré des améliorations significatives, en termes de pertinence, rapportées aux contraintes d'applicabilité et de coûts.

Notre approche de nettoyage, basée sur la structure, et celles identifiées dans la littérature, basées sur le contenu de la journalisation, ont été testées sur un panel de fichiers Log pour démontrer la pertinence de notre méthode en termes de distinction entre les clics des utilisateurs finaux et les hits sous-jacents des agents-navigateurs (*bruits*). Les attributs utilisés sont la Ressource (*page*) consultée et celle référente.

Les tests comparatifs ont été menés sur des fichiers Log disponibles sur Internet en plus de fichiers générés par simulation de navigation sur le site Web Paris 8 et Wikipedia en hébergement local. Notre méthode de nettoyage a montré des avantages significatifs en termes de pertinence, de contraintes d'applicabilité et de coûts.

Pour ce qui est de notre approche de structuration basée sur les flux de clics, et celles basées sur les agents, identifiées dans la littérature ; elles ont été testées sur plusieurs fichiers Log pour démontrer la pertinence de notre méthode dans un contexte sans filature des utilisateurs (*authentication et cookies*), et ceci en termes d'identification et de construction de sessions d'utilisateurs singuliers face aux contraintes de la journalisation séquentielle, l'hébergement multiple et le Web dynamique. Les attributs utilisés sont les IPs anonymes, l'Agent-navigateur, le Temps d'Accès, la Ressource (*page*) sollicitée et celle référente.

Notre méthode fournie plus de pertinence dans la structuration des données Log en sessions d'utilisateurs singuliers sans passer par les techniques usuelles qui posent problème par rapport à la protection de la vie privée, i.e., cookies et authentification. Elle permet donc l'analyse de l'usage Web sans atteinte à la vie privée, notamment dans le cas de systèmes de recommandation et d'optimisation orientées sur le contenu.

Concernant les résultats préliminaires de notre approche d'optimisation symbiotique de l'usage du Web, à savoir l'optimisation de plusieurs dimensions à la fois, i.e., flux du trafic sur le serveur et le réseau, chemins et structures de navigation ; elle a traité la question d'optimisation symbiotique comme problème de classification, et ceci afin de pouvoir prédire, filtrer et contrôler l'optimisation en sorte d'éviter les conflits et contradictions d'une optimisation parasitique.

A cet égard, nous avons illustré, d'une part, les limites de l'optimisation par dimension, i.e., les conflits et contradictions d'optimisation, et d'autre part, l'inadéquation des techniques statistiques à base de fonction de régression face au besoin de prédiction et de contrôle de son produit symbiotique/parasitique. A ce titre, nous avons proposé une approche de classification semi-supervisée pour la prédiction et le contrôle/filtrage de l'optimisation symbiotique/parasitique. L'expérimentation sur un échantillon de journalisation du site Web de notre université, a démontré son potentiel de prédiction et de contrôle/filtrage, à la fois, des trois dimensions pour préserver la symbiose des différentes optimisations. Ces résultats prospectifs nous ont ouvert des perspectives pour l'approfondissement de notre méthode pour qu'elle puisse déboucher sur des résultats plus fiables.

Nous estimons que nos travaux et contributions ont eu la particularité de traiter de la pertinence de trois tâches successives fortement interdépendantes, en termes de pertinence, relatives à une application d'un processus de fouille de donnée, i.e., préparation (*nettoyage & structuration de données de journalisation*), extraction de motifs (*motifs d'usage Web*), analyse et interprétation (*prédiction, filtrage et contrôle de l'optimisation symbiotique*). Ainsi, la valeur ajoutée de chacune de nos contributions en termes de pertinence au titre d'une tâche en amont contribue à celle de la tâche en aval, permettant, ainsi, de renforcer la pertinence de l'ensemble du processus de fouille et de ses sortants.

Enfin, nous tenons à signaler les difficultés liées au contexte expérimental de nos contributions, i.e., la non disponibilité de données expérimentales préparées et labélisées, l'appropriation par la sphère commerciale du domaine de la fouille des données de l'usage Web et les sensibilités en la matière en rapport avec la protection de la vie privée. A cet égard, nous considérons que nos approches, en apportant de la pertinence basée sur le seul indicateur d'utilisateur singulier, dans un contexte de données de journalisation sans information sur les utilisateurs ; représentent une alternative digne d'intérêt pour tout ce qui est de la recommandation, l'adaptation et l'optimisation Web orientées sur le contenu.

VII
Bibliographie

-
- [1] F.-X. de Vaujany, *Les grandes approches théoriques du système d'information*. Paris: Hermès science publications-Lavoisier, 2009.
- [2] S. Desq, B. Fallery, R. Reix, and F. Rodhain, "La spécificité de la recherche francophone en systèmes d'information," *Revue française de gestion*, vol. 33, no. 176, pp. 63–80, Oct. 2007.
- [3] S. Desq, B. Fallery, R. Reix, and F. Rodhain, "25 ans de recherche en Systèmes d'Information," *Systèmes d'Information et Management*, vol. 7, no. 3, p. 1, 2016.
- [4] A. Baccini, S. Déjean, D. Kompaoré, and J. Mothe, "Analyse des critères d'évaluation des systèmes de recherche d'information.," *Technique et Science Informatiques*, vol. 29, no. 3, pp. 289–308, 2010.
- [5] M. Beaudouin-Lafon, "Interaction homme-machine," *Encyclopédie de l'Informatique et des Systèmes d'Information*. Vuibert, 2006.
- [6] M. Magnaudet and S. Chatty, "Quel cadre épistémologique pour une science de l'interaction homme-machine?," in *27ème conférence francophone sur l'Interaction Homme-Machine.*, 2015, p. a9.
- [7] S. Proulx and S. Michel, "L'interactivité technique, simulacre d'interaction sociale et démocratie?," *Technologie de l'Information et Société*, vol. 7, no. 2, pp. 239–255, May 2016.
- [8] N. L. Li and P. Zhang, "The intellectual development of human-computer interaction research: A critical assessment of the MIS literature (1990-2002)," *Journal of the Association for information Systems*, vol. 6, no. 11, p. 9, 2005.
- [9] P. Zhang and N. Li, "An assessment of human–computer interaction research in management information systems: topics and methods," *Computers in Human Behavior*, vol. 20, no. 2, pp. 125–147, Mar. 2004.
- [10] P. Brey, "The epistemology and ontology of human-computer interaction," *Minds and Machines*, vol. 15, no. 3–4, pp. 383–398, 2005.
- [11] C. Kolski, H. Ezzedine, M.-P. Gervais, K. M. Oliveira, and A. Seffah, "Evaluation des SI," *Besoins en méthodes et outils provenant de l'ergonomie et de l'IHM*,

- INFORSID*, pp. 395–410, 2012.
- [12] H. Ezzedine and C. Kolski, “Démarche d’évaluation d’IHM dans les systèmes complexes, application à un poste de supervision du trafic ferroviaire,” *Revue d’Interaction Homme-Machine*, vol. 5, no. 2, 2004.
- [13] D. I. Zahran, H. A. Al-Nuaim, M. J. Rutter, and D. Benyon, “A COMPARATIVE APPROACH TO WEB EVALUATION AND WEBSITE EVALUATION METHODS,” vol. 2014, p. 20.
- [14] E. Brangier, “Le concept de" symbiose homme-technologie-organisation",” *N. Delobbe, G. Karnas & Ch. Vandenberg. Évaluation et développement des compétences au travail. UCL: Presses universitaires de Louvain*, vol. 3, pp. 413–422, 2003.
- [15] E. Brangier, A. Dufresne, and S. Hammes-Adelé, “Approche symbiotique de la relation humain-technologie: perspectives pour l’ergonomie informatique,” *Le travail humain*, vol. 72, no. 4, pp. 333–353, 2010.
- [16] É. Brangier and S. Hammes, “Comment mesurer la relation humain-technologies-organisation?. Élaboration d’un questionnaire de mesure de la relation humain-technologie-organisation basée sur le modèle de la symbiose,” *Perspectives interdisciplinaires sur le travail et la santé*, no. 9–2, 2007.
- [17] É. Brangier and S. Hammes-Adelé, “Beyond the technology acceptance model: Elements to validate the human-technology symbiosis model,” in *International Conference on Ergonomics and Health Aspects of Work with Computers*, 2011, pp. 13–21.
- [18] B. Farbey, F. F. Land, and D. Targett, “A taxonomy of information systems applications: the benefits’ evaluation ladder,” *Eur J Inf Syst*, vol. 4, no. 1, pp. 41–50, Feb. 1995.
- [19] K. M. de Oliveira, V. Thion, S. Dupuy-Chessa, M.-P. Gervais, S. S.-S. Cherfi, and C. Kolski, “Limites de l’évaluation d’un Système d’Information: une analyse fondée sur l’expérience pratique..,” in *INFORSID*, 2012, pp. 411–428.
- [20] Q. Zhang and R. S. Segall, “Web mining: a survey of current research, techniques, and software,” *International Journal of Information Technology & Decision*

Making, vol. 7, no. 04, pp. 683–720, 2008.

- [21] J. Srivastava, R. Cooley, M. Deshpande, and P.-N. Tan, “Web usage mining: Discovery and applications of usage patterns from web data,” *Acm Sigkdd Explorations Newsletter*, vol. 1, no. 2, pp. 12–23, 2000.
- [22] M. Srivastava, R. Garg, and P. K. Mishra, “Preprocessing techniques in web usage mining: A survey,” *International Journal of Computer Applications*, vol. 97, no. 18, 2014.
- [23] A. V. Srinivas, “A Survey on Preprocessing of Web-Log Data in Web Usage Mining,” *International Journal for Modern Trends in Science and Technology*, vol. Vol. 03, no. Issue 02, 2017.
- [24] K. Ronny, “Mining e-commerce data: the good, the bad, and the ugly,” *Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 8–13, 2001.
- [25] B. Berendt, B. Mobasher, and M. Spiliopoulou, “Web usage mining for e-business applications,” in *Tutorial, ECML/PKDD Conference*, 2002.
- [26] J. Srivastava, P. Desikan, and V. Kumar, “Web mining: Accomplishments and future directions,” in *National Science Foundation Workshop on Next Generation Data Mining (NGDM’02)*, 2002, pp. 51–56.
- [27] Z. Pabarskaite and A. Raudys, “A process of knowledge discovery from web log data: Systematization and critical review,” *Journal of Intelligent Information Systems*, vol. 28, no. 1, pp. 79–104, Feb. 2007.
- [28] Z. Pabarskaite, “Implementing advanced cleaning and end-user interpretability technologies in web log mining,” in *Information Technology Interfaces, 2002. ITI 2002. Proceedings of the 24th International Conference on*, 2002, pp. 109–113.
- [29] M. Spiliopoulou, B. Mobasher, B. Berendt, and M. Nakagawa, “A framework for the evaluation of session reconstruction heuristics in web-usage analysis,” *Informatics journal on computing*, vol. 15, no. 2, pp. 171–190, 2003.
- [30] M. Spiliopoulou, “Web usage mining for Web site evaluation,” *Communications of the ACM*, vol. 43, no. 8, pp. 127–134, Aug. 2000.

- [31] M. Spiliopoulou and C. Pohle, “Data mining for measuring and improving the success of web sites,” in *Applications of Data Mining to Electronic Commerce*, Springer, 2001, pp. 85–114.
- [32] R. Cooley, B. Mobasher, and J. Srivastava, “Web mining: information and pattern discovery on the World Wide Web,” in *Proceedings Ninth IEEE International Conference on Tools with Artificial Intelligence*, 1997, pp. 558–567.
- [33] T. Srivastava, P. Desikan, and V. Kumar, “Web mining—concepts, applications and research directions,” in *Foundations and advances in data mining*, Springer, 2005, pp. 275–307.
- [34] R. Cooley, P.-N. Tan, and J. Srivastava, “Discovery of interesting usage patterns from web data,” in *International Workshop on Web Usage Analysis and User Profiling*, 1999, pp. 163–182.
- [35] R. Cooley, P.-N. Tan, and J. Srivastava, “Discovery of Interesting Usage Patterns from Web Data,” in *Web Usage Analysis and User Profiling*, B. Masand and M. Spiliopoulou, Eds. Springer Berlin Heidelberg, 2000, pp. 163–182.
- [36] M. H. A. Wahab, M. N. H. Mohd, H. F. Hanafi, and M. F. M. Mohsin, “Data pre-processing on web server logs for generalized association rules mining algorithm,” *World Academy of Science, Engineering and Technology*, vol. 48, p. 2008, 2008.
- [37] R. Cooley, B. Mobasher, and J. Srivastava, “Data preparation for mining world wide web browsing patterns,” *Knowledge and information systems*, vol. 1, no. 1, pp. 5–32, 1999.
- [38] F. M. Facca and P. L. Lanzi, “Mining interesting knowledge from weblogs: a survey,” *Data & Knowledge Engineering*, vol. 53, no. 3, pp. 225–241, Jun. 2005.
- [39] M. Dimitrijević, Z. Bošnjak, and S. Subotica, “Discovering interesting association rules in the web log usage data,” *Interdisciplinary Journal of Information, Knowledge, and Management*, vol. 5, pp. 191–207, 2010.
- [40] L. Geng and H. J. Hamilton, “Interestingness measures for data mining: A survey,” *ACM Computing Surveys*, vol. 38, no. 3, pp. 9-es, Sep. 2006.
- [41] S. Jaroszewicz and D. A. Simovici, “Interestingness of Frequent Itemsets Using

-
- Bayesian Networks As Background Knowledge,” in *Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, New York, NY, USA, 2004, pp. 178–186.
- [42] K. McGarry, “A survey of interestingness measures for knowledge discovery,” *The knowledge engineering review*, vol. 20, no. 01, pp. 39–61, 2005.
- [43] M. Srivastava, R. Garg, and P. K. Mishra, “Analysis of Data Extraction and Data Cleaning in Web Usage Mining,” 2015, pp. 1–6.
- [44] V. CHITRAA and D. A. S. THANAMANI, “Web Log Data Cleaning For Enhancing Mining Process,” vol. 01, no. 03, p. 7, 2012.
- [45] B. Berendt, B. Mobasher, M. Spiliopoulou, and J. Wiltshire, “Measuring the Accuracy of Sessionizers for Web Usage Analysis,” p. 8.
- [46] V. Chitraa, D. Davamani, and A. Selvdoss, “A survey on preprocessing methods for web usage data,” *arXiv preprint arXiv:1004.1257*, 2010.
- [47] “Log Files - Apache HTTP Server.” [Online]. Available: <https://httpd.apache.org/docs/1.3/logs.html>. [Accessed: 19-Feb-2018].
- [48] “La mission du W3C | W3C - Bureau France.” [Online]. Available: <https://www.w3c.fr/a-propos-du-w3c-france/la-mission-du-w3c/>. [Accessed: 17-Jul-2019].
- [49] M. Spiliopoulou, “The laborious way from data mining to web log mining,” *Computer Systems Science and Engineering*, vol. 14, no. 2, pp. 113–126, 1999.
- [50] A. Anitha, “A new web usage mining approach for next page access prediction,” *International Journal of Computer Applications*, vol. 8, no. 11, pp. 7–10, 2010.
- [51] K. Chaudhary and S. K. Gupta, “Web Usage Mining Tools & Techniques: A Survey,” *International Journal of Scientific & Engineering Research*, vol. 4, no. 6, p. 1762, 2013.
- [52] K. Agrawal and H. Makwana, “Review of Different Log Management Tools used for Data Analysis,” *Data Mining and Knowledge Engineering*, vol. 7, no. 4, pp. 161–163, 2015.
- [53] V. Bharanipriya and V. K. Prasad, “Web content mining tools: a comparative study,”

- International Journal of Information Technology and Knowledge Management*, vol. 4, no. 1, pp. 211–215, 2011.
- [54] S. Jayaprakash, “A Survey on Web Mining Tools and Techniques,” p. 5.
- [55] A. Kumar and R. K. Singh, “Web Mining Overview, Techniques, Tools and Applications: A Survey,” 2016.
- [56] A. L. Lemos, F. Daniel, and B. Benatallah, “Web Service Composition: A Survey of Techniques and Tools,” *ACM Computing Surveys*, vol. 48, no. 3, pp. 1–41, Dec. 2015.
- [57] “Data Analysis and Reporting using Different Log Management Tools,” 07-Dec-2016. [Online]. Available: <http://ijcsmc.com/docs/papers/July2015/V4I7201553.pdf>. [Accessed: 07-Dec-2016].
- [58] “A survey- web mining tools and technique,” *International Journal of Latest Trends in Engineering and Technology*, vol. 7, no. 4, 2016.
- [59] R. Kosala and H. Blockeel, “Web mining research: A survey,” *ACM Sigkdd Explorations Newsletter*, vol. 2, no. 1, pp. 1–15, 2000.
- [60] B. Bakariya, K. K. Mohbey, and G. S. Thakur, “An Inclusive Survey on Data Preprocessing Methods Used in Web Usage Mining,” in *Proceedings of Seventh International Conference on Bio-Inspired Computing: Theories and Applications (BIC-TA 2012)*, vol. 202, J. C. Bansal, P. Singh, K. Deep, M. Pant, and A. Nagar, Eds. India: Springer India, 2013, pp. 407–416.
- [61] T. T. Aye, “Web log cleaning for mining of web usage patterns,” in *Computer Research and Development (ICCRD), 2011 3rd International Conference on*, 2011, vol. 2, pp. 490–494.
- [62] J. Borges and M. Levene, “Data mining of user navigation patterns,” in *Web usage analysis and user profiling*, Springer, 2000, pp. 92–112.
- [63] P. Bari and P. M. Chawan, “Web usage mining,” *Journal of Engineering, Computers & Applied Sciences (JEC&AS)*, vol. 2, no. 6, pp. 34–38, 2013.
- [64] M. Hanoune and F. Benabbou, “Traitement et exploration du fichier Log du serveur

-
- web, pour l'extraction des connaissances: Web usage mining," *Afrique Science: Revue Internationale des Sciences et Technologie*, vol. 2, no. 3, 2006.
- [65] A. G. Büchner and M. D. Mulvenna, "Discovering Internet Marketing Intelligence Through Online Analytical Web Usage Mining," *SIGMOD Rec.*, vol. 27, no. 4, pp. 54–61, Dec. 1998.
- [66] M. Grislin and C. Kolski, "Evaluation des Interfaces Homme-Machine lors du développement des systèmes interactifs," *Technique et Science Informatiques (TSI)*, vol. 15, no. 3, pp. 265–296, 1996.
- [67] W. de-Abreu-Cybis, "UseMonitor: suivre l'évolution de l'utilisabilité des sites web à partir de l'analyse des fichiers de journalisation," in *Proceedings of the 18th International Conference of the Association Francophone d'Interaction Homme-Machine*, 2006, pp. 295–296.
- [68] A. Mille, "Des traces à l'ère du Web," *Intellectica*, vol. 1, no. 59, pp. 7–28, 2013.
- [69] J. Vanderdonckt and A. Beirekdar, "Automated Web Evaluation by Guideline Review.," *J. Web Eng.*, vol. 4, no. 2, pp. 102–117, 2005.
- [70] W.-C. Chiou, C.-C. Lin, and C. Perng, "A Strategic Framework for Website Evaluation Based on a Review of the Literature from 1995-2006," *Inf. Manage.*, vol. 47, no. 5–6, pp. 282–290, Aug. 2010.
- [71] A. Halfaker *et al.*, "User Session Identification Based on Strong Regularities in Inter-activity Time," *arXiv:1411.2878 [cs]*, Nov. 2014.
- [72] A. Halfaker *et al.*, "User Session Identification Based on Strong Regularities in Inter-activity Time," 2015, pp. 410–418.
- [73] S. Langhnoja, M. Barot, and D. Mehta, "Pre-Processing: Procedure on Web Log File for Web Usage Mining," vol. 2, no. 12, p. 5, 2012.
- [74] M. Dhandi and R. K. Chakrawarti, "A comprehensive study of web usage mining," 2016, pp. 1–5.
- [75] R. Omar, A. O. Md Tap, and Z. S. Abdullah, "Web usage mining: A review of recent works," 2014, pp. 1–5.
- [76] T. R. Shultz *et al.*, "Confusion Matrix," in *Encyclopedia of Machine Learning*, C.

- Sammut and G. I. Webb, Eds. Boston, MA: Springer US, 2011, pp. 209–209.
- [77] M. A. Beaumont, K. M. Ibrahim, P. Boursot, and M. W. Bruford, “Measuring Genetic Distance,” in *Molecular Tools for Screening Biodiversity*, A. Karp, P. G. Isaac, and D. S. Ingram, Eds. Dordrecht: Springer Netherlands, 1998, pp. 315–325.
- [78] B. J. B. Keats and S. L. Sherman, “Population Genetics,” in *Emery and Rimoin’s Principles and Practice of Medical Genetics*, Elsevier, 2013, pp. 1–12.
- [79] M. C. Naldi, A. C. de Carvalho, and R. J. G. B. Campell, “Genetic clustering for data mining,” in *Soft computing for knowledge discovery and data mining*, Springer, 2008, pp. 113–132.
- [80] C. Bird, S. Karl, P. Mouse, and R. Toonen, “Detecting and measuring genetic differentiation,” in *Phylogeography and Population Genetics in Crustacea*, vol. 20112046, C. Schubart, Ed. CRC Press, 2011, pp. 31–55.
- [81] D. J. Lawson and D. Falush, “Population Identification Using Genetic Data,” *Annual Review of Genomics and Human Genetics*, vol. 13, no. 1, pp. 337–361, Sep. 2012.
- [82] N. Philip, “Génétique des populations.”
- [83] Y.-C. Chiou and L. W. Lan, “Genetic clustering algorithms,” *European journal of operational research*, vol. 135, no. 2, pp. 413–427, 2001.
- [84] J. F. Jimenez, F. J. Cuevas, and J. M. Carpio, “Genetic algorithms applied to clustering problem and data mining,” in *Proceedings of the 7th WSEAS International Conference on Simulation, Modelling and Optimization*, 2007, pp. 219–224.
- [85] U. Maulik, S. Bandyopadhyay, and S. B, “Genetic Algorithm-Based Clustering Technique,” *Pattern Recognition*, vol. 33, pp. 1455–1465, 2000.
- [86] “Allele frequency & the gene pool,” *Khan Academy*. [Online]. Available: <https://www.khanacademy.org/science/biology/her/heredity-and-genetics/a/allele-frequency-the-gene-pool>. [Accessed: 01-Mar-2018].
- [87] R. Ivancsy and S. Juhasz, “Analysis of web user identification methods,” *World Academy of Science, Engineering and Technology*, vol. 2, no. 3, pp. 212–219, 2007.
- [88] T. Hussain, S. Asghar, and N. Masood, “Web usage mining: A survey on

-
- preprocessing of web log file,” in *Information and Emerging Technologies (ICIET), 2010 International Conference on*, 2010, pp. 1–6.
- [89] D. Tanasa and B. Trousse, “Advanced data preprocessing for intersites Web usage mining,” *IEEE Intelligent Systems*, vol. 19, no. 2, pp. 59–65, Mar. 2004.
- [90] R. F. Dell, P. E. Román, and J. D. Velásquez, “Web User Session Reconstruction Using Integer Programming,” 2008, pp. 385–388.
- [91] T. Arce, P. E. Román, J. Velásquez, and V. Parada, “Identifying web sessions with simulated annealing,” *Expert Systems with Applications*, vol. 41, no. 4, pp. 1593–1600, Mar. 2014.
- [92] P.-N. Tan and V. Kumar, “Discovery of Web Robot Sessions Based on Their Navigational Patterns,” in *Intelligent Technologies for Information Analysis*, Berlin, Heidelberg: Springer Berlin Heidelberg, 2004, pp. 193–222.
- [93] J. D. Velásquez, Ed., *Advanced techniques in web intelligence-2: web user browsing behaviour and preference analysis*. Berlin: Springer, 2012.
- [94] Z. Huiying and L. Wei, “An intelligent algorithm of data pre-processing in Web usage mining,” in *Intelligent Control and Automation, 2004. WCICA 2004. Fifth World Congress on*, 2004, vol. 4, pp. 3119–3123.
- [95] N. Sharma and P. Makhija, “Web usage Mining: Web user Session Construction using Map-Reduce,” p. 4, 2017.
- [96] N. Sharma and P. Makhija, “Web usage Mining: A Novel Approach for Web user Session Construction,” p. 5, 2015.
- [97] M. A. Bayir, I. H. Toroslu, M. Demirbas, and A. Cosar, “Discovering better navigation sequences for the session construction problem,” *Data & Knowledge Engineering*, vol. 73, pp. 58–72, Mar. 2012.
- [98] S. Chitra and B. Kalpana, “Hierarchical Directed Acyclic Graph (HDAG) Based Preprocessing Technique for Session Construction,” in *Advances in Computing and Information Technology*, vol. 177, N. Meghanathan, D. Nagamalai, and N. Chaki, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2013, pp. 611–621.
- [99] S. Chitra and B. Kalpana, “A Novel Preprocessing Technique for Session

- Construction using Propositional DAGs,” *International Journal of Computer Applications*, vol. 64, no. 16, pp. 8–12, Feb. 2013.
- [100] Z. Zheng, B. Padmanabhan, and S. O. Kimbrough, “On the Existence and Significance of Data Preprocessing Biases in Web-Usage Mining,” *INFORMS Journal on Computing*, vol. 15, no. 2, pp. 148–170, May 2003.
- [101] B. Padmanabhan, Z. Zheng, and S. O. Kimbrough, “Personalization from incomplete data: what you don’t know can hurt,” 2001, pp. 154–163.
- [102] B. A. Huberman, “Strong Regularities in World Wide Web Surfing,” *Science*, vol. 280, no. 5360, pp. 95–97, Apr. 1998.
- [103] L. D. Catledge and J. E. Pitkow, “Characterizing browsing strategies in the World-Wide web,” *Computer Networks and ISDN Systems*, vol. 27, no. 6, pp. 1065–1073, Apr. 1995.
- [104] Jiming Liu, Shiwu Zhang, and Jie Yang, “Characterizing web usage regularities with information foraging agents,” *IEEE Transactions on Knowledge and Data Engineering*, vol. 16, no. 5, pp. 566–584, May 2004.
- [105] B. Vo and B. Le, “Interestingness measures for association rules: Combination between lattice and hash tables,” *Expert Systems with Applications*, vol. 38, no. 9, pp. 11630–11640, Sep. 2011.
- [106] A. Merceron and K. Yacef, “Interestingness measures for association rules in educational data,” in *Educational Data Mining 2008*, 2008.
- [107] C. Tew, C. Giraud-Carrier, K. Tanner, and S. Burton, “Behavior-based clustering and analysis of interestingness measures for association rule mining,” *Data Mining and Knowledge Discovery*, vol. 28, no. 4, pp. 1004–1045, Jul. 2014.
- [108] R. Agrawal, T. Imielinski, A. Swami, H. Road, and S. Jose, “Mining Association Rules between Sets of Items in Large Databases,” p. 10.
- [109] S. Kotsiantis and D. Kanellopoulos, “Association Rules Mining: A Recent Overview,” p. 12.
- [110] V. Mishra, T. K. Mishra, and A. Mishra, “Algorithms for Association Rule Mining: A General Survey on Benefits And Drawbacks of Algorithms,”

International Journal of Advanced Research in Computer Science, p. 5, 2010.

- [111] X. Wang and C. Zhai, "Mining term association patterns from search logs for effective query reformulation," in *Proceedings of the 17th ACM conference on Information and knowledge management*, 2008, pp. 479–488.
- [112] H. Yang and S. Parthasarathy, "On the use of constrained associations for web log mining," in *International Workshop on Mining Web Data for Discovering Usage Patterns and Profiles*, 2002, pp. 100–118.
- [113] F. Masseglia, D. Tanasa, and B. Trousse, "Diviser pour Découvrir: une Méthode d'Analyse du Comportement de Tous les Utilisateurs d'un Site Web.," *Ingénierie des Systèmes d'Information*, vol. 9, no. 1, pp. 61–83, 2004.
- [114] E. Suzuki, "Interestingness measures-limits, desiderata, and recent results," *QIMIE/PAKDD*, 2009.
- [115] A. Zimmermann, "Objectively evaluating interestingness measures for frequent itemset mining," in *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, 2013, pp. 354–366.
- [116] M. Charrad, M. B. Ahmed, and Y. Lechevallier, "Extraction des connaissances à partir des fichiers logs," *Atelier fouille du Web EGC2006*, vol. 768, 2005.
- [117] R. Agrawal and R. Srikant, "Mining sequential patterns," in *Proceedings of the Eleventh International Conference on Data Engineering*, Taipei, Taiwan, 1995, pp. 3–14.
- [118] D. S. B. Bajaj and D. Garg, "SURVEY ON SEQUENCE MINING ALGORITHMS," p. 7.
- [119] R. Boghey and S. Singh, "Sequential Pattern Mining: A Survey on Approaches," in *2013 International Conference on Communication Systems and Network Technologies*, Gwalior, 2013, pp. 670–674.
- [120] P. Fournier-Viger and J. C.-W. Lin, "A Survey of Sequential Pattern Mining," p. 24.
- [121] V. S. Motegaonkar and M. V. Vaidya, "A Survey on Sequential Pattern Mining Algorithms," vol. 5, p. 7, 2014.

- [122] V. C. S. Rao, “Survey on Sequential Pattern Mining Algorithms,” *International Journal of Computer Applications*, vol. 76, p. 8.
- [123] Q. Zhao, “Sequential Pattern Mining: A Survey,” *Technical Report*, p. 27.
- [124] F. Maseglia, M. Teisseire, and P. Poncelet, “Extraction de motifs séquentiels. Problèmes et méthodes,” *Ingénierie des systèmes d’information*, vol. 9, no. 3–4, pp. 183–210, Aug. 2004.
- [125] Q. Yang and H. H. Zhang, “Web-log mining for predictive Web caching,” *IEEE Transactions on Knowledge and Data Engineering*, vol. 15, no. 4, pp. 1050–1053, Jul. 2003.
- [126] J. Zhu, J. Hong, and J. G. Hughes, “Using Markov Chains for Link Prediction in Adaptive Web Sites,” in *Soft-Ware 2002: Computing in an Imperfect World*, vol. 2311, D. Bustard, W. Liu, and R. Sterritt, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2002, pp. 60–73.
- [127] M. Scholz, “R Package **clickstream** : Analyzing Clickstream Data with Markov Chains,” *Journal of Statistical Software*, vol. 74, no. 4, 2016.
- [128] T. De Greef, K. van Dongen, M. Grootjen, and J. Lindenberg, “Augmenting cognition: reviewing the symbiotic relation between man and machine,” in *International Conference on Foundations of Augmented Cognition*, 2007, pp. 439–448.
- [129] D. Roy, “10\$times\$—human-machine symbiosis,” *BT Technology Journal*, vol. 22, no. 4, pp. 121–124, 2004.
- [130] P. Lemberger, M. Batty, M. Morel, J.-L. Raffaëlli, and A. Géron, *Big data et machine learning: les concepts et les outils de la data science*. Malakoff: Dunod, 2016.
- [131] E. Biernat, M. Lutz, and Y. LeCun, *Data science: fondamentaux et études de cas : machine learning avec Python et R*. Paris: Eyrolles, 2015.
- [132] X. Kong, X. Shi, and P. S. Yu, “Multi-Label Collective Classification,” in *Proceedings of the 2011 SIAM International Conference on Data Mining*, 2011, pp. 618–629.

-
- [133] P. Sen, G. Namata, M. Bilgic, L. Getoor, B. Galligher, and T. Eliassi-Rad, "Collective Classification in Network Data," *AI Magazine*, vol. 29, no. 3, p. 93, Sep. 2008.
- [134] G. Govaert, "Classification partiellement supervis´ee," p. 56.
- [135] Z. Xing, J. Pei, and E. Keogh, "A brief survey on sequence classification," *ACM SIGKDD Explorations Newsletter*, vol. 12, no. 1, p. 40, Nov. 2010.

VIII

Liste des Figures

Figure I-1 Contexte de la Fouille des données de l'usage du Web	4
Figure II-1 Taxonomie sommaire de la fouille des données Web	18
Figure II-2 Processus de fouille des données de l'usage Web	24
Figure II-3 Système d'activité Web	28
Figure II-4 Matrice des Attributs Utiles	31
Figure III-1 Objet et caractéristique génétiques	60
Figure IV-1 Echantillon de fichier Log.....	76
Figure V-1 Descriptif de l'activité de navigation.....	124
Figure V-2 Corrélations des indicateurs de l'activité de navigation par sessions.....	125
Figure V-3 Corrélations des indicateurs de l'activité de navigation par suffixe	127
Figure V-4 Modèle de contrôle d'Optimisation Symbiotique/Parasitique.....	132

IX

Liste des Tableaux

Table I-1 Problématique SI & IHM	2
Table II-1 Entrées et Formats d'un Fichier ALF.....	30
Table II-2 Attributs Utiles d'un Fichier Log.....	31
Table III-1 Entrées d'un fichier de journalisation ALF/ECLF.....	46
Table III-2 Liste pertinente de requêtes de navigation-utilisateur final	51
Table III-3 Motif régulier de structure de journalisation	52
Table III-4 Données expérimentales	55
Table III-5 Résultats des nettoyages	57
Table III-6 Motifs Génétiques de la structure de journalisation	62
Table III-7 Descriptif des données expérimentales.....	68
Table III-8 Résultats du nettoyage par partitionnement génétique	70
Table IV-1 Données expérimentales	93
Table IV-2 Evaluation de la qualité de reconstruction.....	94
Table V-1 Données expérimentales	100
Table V-2 Table Sessions Utilisateurs [USE_TAB]	102
Table V-3 Table Statistiques Sessions [STA_TAB].....	103
Table V-4 Table Statistiques Page [STA_STR].....	104
Table V-5 Résultats d'extraction générique des règles navigation.....	107
Table V-6 Résultats d'extraction stratégique des règles	110
Table V-7 Résultats d'extraction de motifs séquentiels de navigation	117
Table V-8 Fréquence relative des longueurs des chemins de navigation	118
Table V-9 Impact de Mise en Cache [CAC_IMP_TAB].....	121
Table V-10 Données et instrument de mesure de l'optimalité symbiotique	131
Table V-11 Résultats de prédiction et d'apprentissage	134

X

Liste des Algorithmes

Algorithme III-1 Nettoyage Conventionnel	48
Algorithme III-2 Nettoyage Avancé	48
Algorithme III-3 Nettoyage basé sur les règles de la structure	54
Algorithme III-4 Nettoyage basé sur le Partitionnement Génétique.....	68
Algorithme IV-1 Reconstruction de session centrée Flux de Clics & Agent.....	87

XI

Publications

-
- [1] Ganibardi and C. A. Ali, “Web Usage Data Cleaning: A Rule-Based Approach for Weblog Data Cleaning,” in *Big Data Analytics and Knowledge Discovery*, vol. 11031, C. Ordonez and L. Bellatreche, Eds. Cham: Springer International Publishing, 2018, pp. 193–203. [\[CORE2018 B\]](#)
- [2] Ganibardi and C. A. Ali, “Weblog Data Structuration: A Stream-centric approach for improving session reconstruction quality,” in *Proceedings of the 20th International Conference on Information Integration and Web-based Applications & Services - iiWAS2018*, Yogyakarta, Indonesia, 2018, pp. 263–271. [\[CORE2018 C\]](#)
- [3] Ganibardi and C. Arab Ali, “A Genetics Clustering-Based Approach for Weblog Data Cleaning,” in *2018 Sixth International Conference on Enterprise Systems ES(ICEIS)*, Limassol, 2018, pp. 75–81. [\[CORE2018 C\]](#)

