UNIVERSITE BADJI MOKHTAR-ANNABA

DEPARTEMENT D'INFORMATIQUE

Introduction à la fouille de web

Dr. Samira LAGRINI

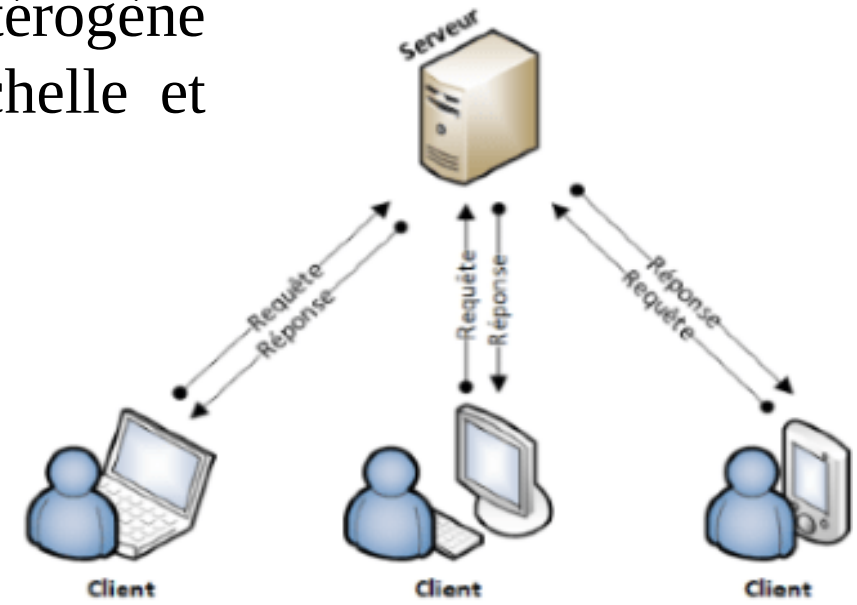
lagrini.samira83@gmail.com

Année universitaire:2024/2025

Qu’est ce que le world wide web?

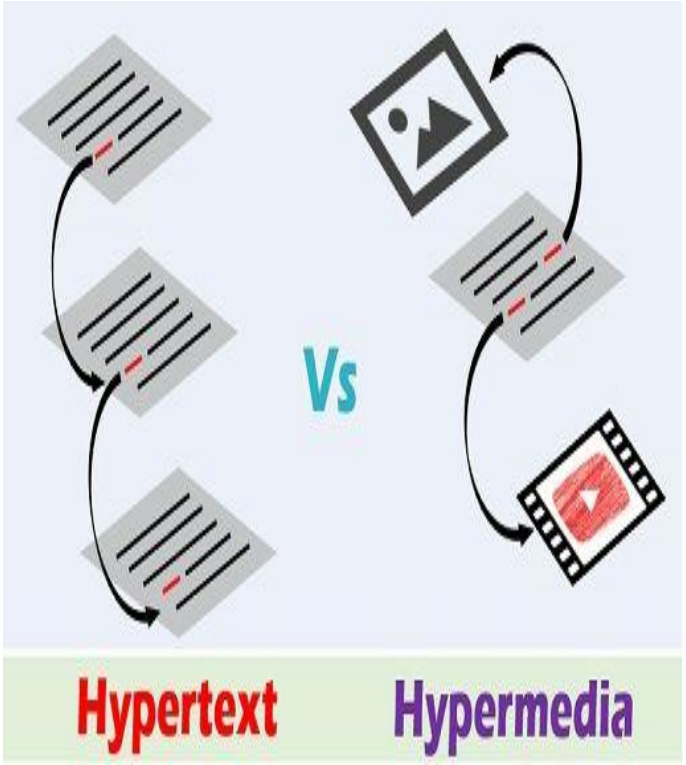
Le Web (*World Wide Web ou* WWW) est un réseau informatique basé sur Internet qui permet aux utilisateurs d'un ordinateur d'accéder à des donnée stockées sur **un autre** via le réseau mondial appelé **Internet**.

Le Web est un repertoire immense de donnée largement hétérogène (text, image, vedio, hypertexte…) distribué à grande échelle et fortement interconnecté



Hypertexte, Hyperliens, Hypermédia

- Le fonctionnement du Web repose sur la structure de ses documents hypertextes.
- Un document hypertexte est un texte qui contient des liens (hyperliens) vers d'autres blocs de texte,
- Hypermédia est une extension de l'hypertexte comprenant du texte, de l'audio, des images, de la vidéo et des images fixes ou animées.



Hypertexte, Hyperliens, Hypermédia

- L'hypertexte et l'hypermedia permettent aux auteurs de pages Web de lier leurs documents à d'autres documents connexes résidant sur des ordinateurs partout dans le monde. Pour visualiser ces documents, il suffit juste de suivre les hyperliens.
- YouTube est un bon exemple pour illustrer l'hypermédia. Il combine l'hypertexte avec les entrées vidéo. Dans les vidéos, il y a aussi des boutons d'hyperlien. En cliquant sur ces boutons, l'utilisateur accédera au blog des propriétaires de vidéos.

Particularités et défis du web

- **Un grand volume de données**
- **Le contenu du Web est dynamique**

Les données sur Internet sont rapidement mises à jour (*les actualités, les achats, les actualités financières, les sports, etc*).

- **Hétérogénéités de L'information**

plusieurs pages peuvent présenter des informations similaires en utilisant des mots et des formats complètement différents, (ce qui pose un problème pour l'intégration d'informations.

Particularités et défis du web-2-

➢ Hétérogénéité de format de données

Les données **structurées** (proviennent de bases de données), les **données semi-structurées** (pages Web), les données **non structurées** (texte et les fichiers multimédias).

➢ Le Bruit

Une page Web contient beaucoup d' informations (le contenu principal , les hyperliens, les publicités, les avis de droit d'auteur ..etc). Pour une application particulière, seule une partie de l'information est utile. Le reste est considéré comme du bruit.

➢Pas de contrôle de la qualité de l'information

Sur le web, n'importe qui peut écrire ce qu'il veut , de ce fait, une grande quantité d'informations sur le Web est de mauvaise qualité, erronée, voir trompeuse.

Particularités et défis du web-3-

➢ L'interconnexion

Les données sur le Web sont interconnecté via des hyperliens qui peuvent être internes (existent entre les pages Web d'un site) ou externes (connecte les pages de différents sites).

➢ Le web est une société virtuelle

Plus le grand volume de données et d'information, Le web constitue une plateforme d'interactions entre les personnes et les organisations(les forums, les blogs, les sites de réseaux sociaux)

Ces particularités présentent à la fois des défis et des opportunités pour la découverte d'informations et de connaissances sur le Web

Le web est une source très riche pour la fouille de donnée

Le web mining

- Le Web Mining est un domaine de l'informatique qui utilise des techniques de data mining pour extraire des informations et des connaissances utiles à partir des données disponibles sur le Web.

- Il s'agit d'un processus d'exploration de données à grande échelle qui combine des méthodes d'intelligence artificielle, de traitement du langage naturel (NLP), de machine learning et d'analyse statistique pour analyser et découvrir des motifs cachés dans les données web

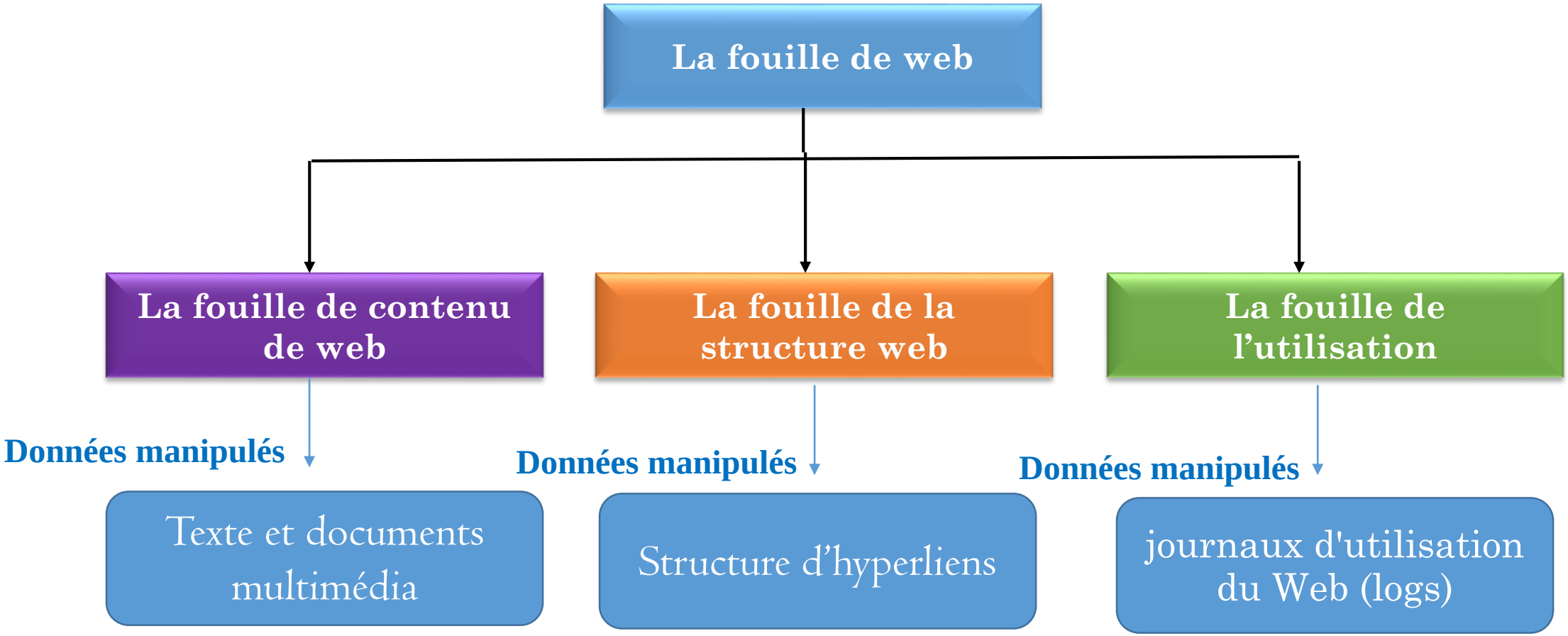
Le web mining

❑ Les données de Web peuvent être :

- ✓ Documents (pages web)
- ✓ Des liens entre ces documents
- ✓ Des informations associées à l'utilisation de ces documents

❑ le Web Mining vise à découvrir des informations et des connaissances utiles à partir du contenu des pages, de **la structure des hyperliens** et des données **d'utilisation des sites web**.

Catégories du web mining



La fouille de contenu de web/ web content mining

- le WCM s'intéresse à Analyse du contenu des pages web afin d'en extraire des informations (ex: l'extraction d'opinions à partir de forums ou de réseaux sociaux)

- Les données manipulées peuvent inclure :
  - ✓**Texte**
  - ✓**Multimédia** (images, vidéos ou audios)
  - ✓**HTML et XML** (données structurées.)

Techniques de web content mining

Le web content mining utilise les techniques de:

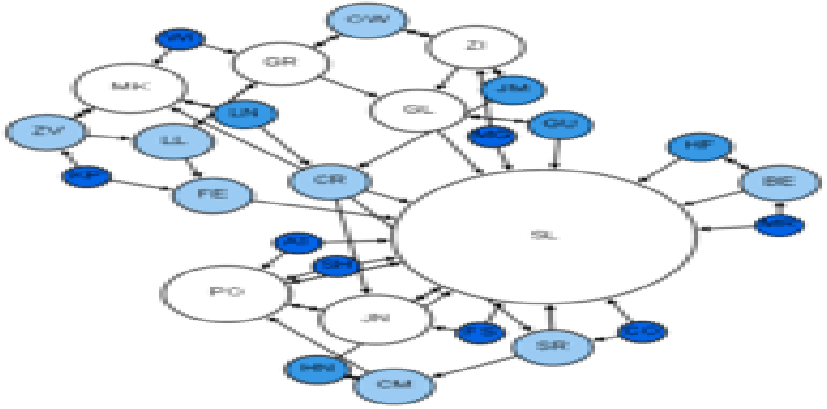
- Text Mining : Analyse des textes pour extraire des informations telles que des mots-clés, des résumés, des sentiments, etc.
- Extraction d'informations : Identification de données spécifiques telles que des noms, des dates, des lieux, etc.
- Classification et clustering : Regroupement des documents similaires ou classification des documents dans des catégories prédéfinies
- ……..

La fouille de la structure web (WSM)

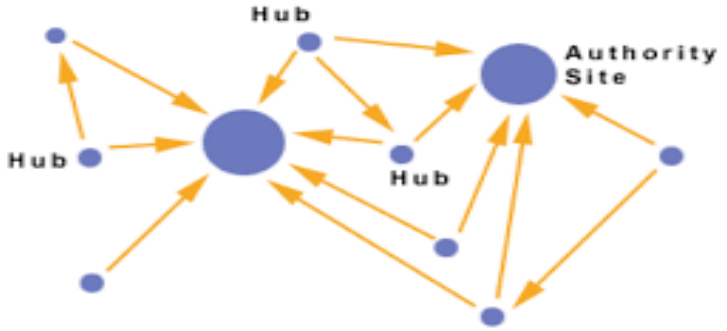
- La fouille de la structure Web appelé également link analysis peut être défini comme le processus de découverte d'information à partir de la structure des liens entre les entités Web.
- La fouille de la structure web s'intéresse à l'analyse des hyperliens qui existent entre les entités du Web.
- Ces entités peuvent être des utilisateurs de réseaux sociaux liés par des liens d'amitié ou des pages internet liées entre elles par des hyperliens.

La fouille de la structure web

Techniques



- La théorie des graphes pour analyser les structures de liens.
- **PageRank** : Algorithme utilisé par Google pour mesurer l'importance des pages web.
- **HITS (Hyperlink-Induced Topic Search)** : Algorithme qui classe les pages web en tant qu'autorités et hubs.
- ….



Le Web usage mining

- Le Web Usage Mining se focalise sur l'extraction de connaissance des données de journaux Web (web log data), afin de connaître le comportement des utilisateurs d'après leurs activités effectuées sur le Web.
- Les fichiers log enregistrent les traces de navigation (click stream) , et les actions effectuées par les utilisateurs lors de leurs visites des sites Web

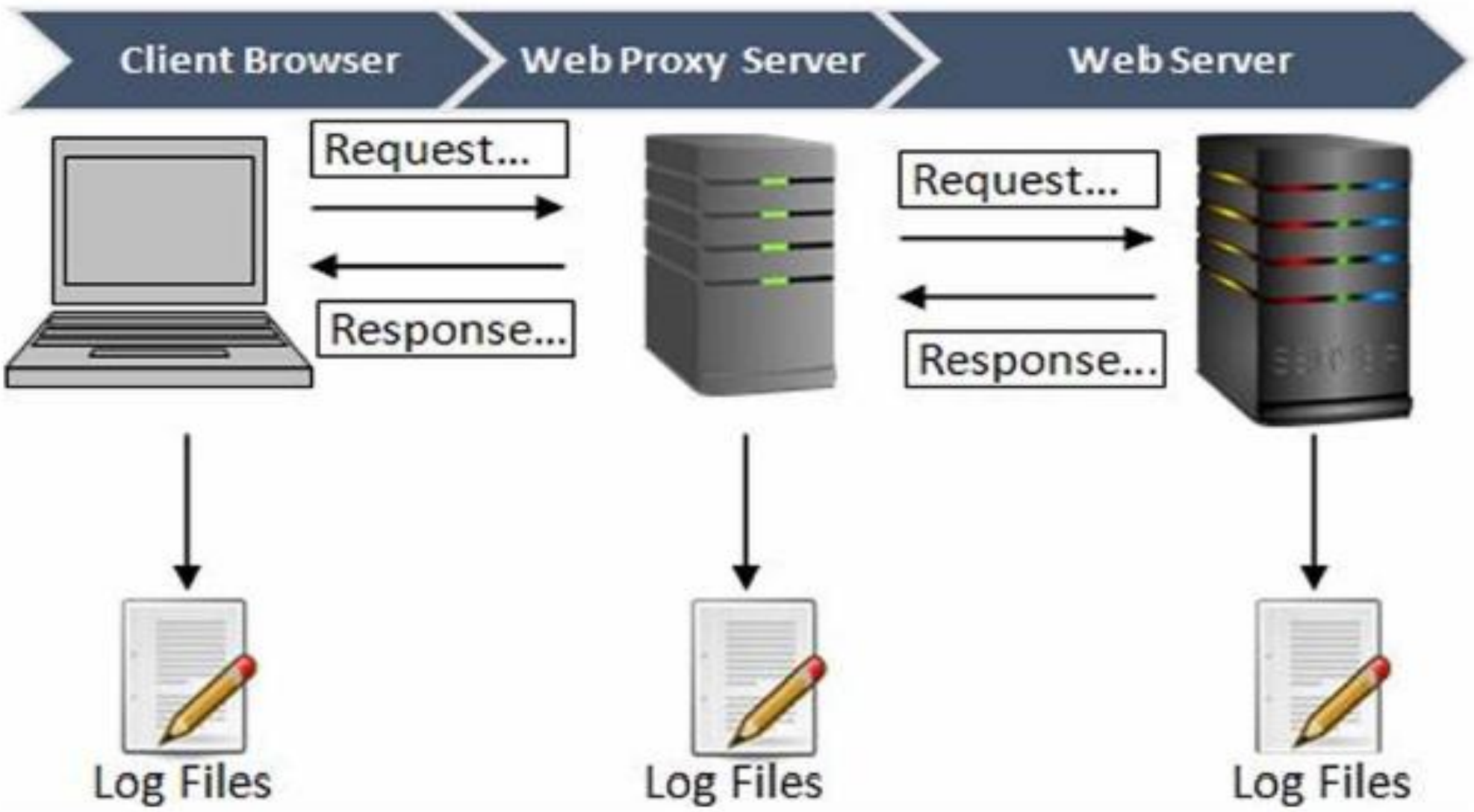
Source de données dans le WUM

Le WUM a besoin des données du journaux Web (log data) afin de les traiter et extraire des connaissances

Un journal Web (web log) est un fichier dans lequel le serveur Web écrit les informations à chaque fois qu'un utilisateur demande un site Web à partir de ce serveur particulier.

- ❑ les journaux web (log data) sont extraites à partir de:
  - ➢ serveurs Web: un programme applicatif acceptant des connexions d'accès a ses ressources dans le but de traiter des requêtes en délivrant une réponse.
  - ➢ serveurs proxy: serveur intermédiaire qui existe entre le client et le serveur Web. Il permet d'acheminer indirectement les requêtes de l'un vers l'autre
  - ➢ Navigateur client: un programme applicatif dont la fonction principale est d'émettre des requêtes,

Source de données dans le WUM



Processus général du Web mining

1. **Collecte des données** : Extraire ou collecter les données disponibles sur le web. Il existe plusieurs méthodes (Web Crawling , Web Scraping, APIs)
2. **Prétraitement des données** : Nettoyage des données, gestion des données manquantes, suppression des doublons, Transformation des données (non structurées) en un format structuré (un tableau ou une base de données) pour faciliter l'analyse.
3. **Modélisation** : Application d’algorithmes pour extraire des connaissances utiles (classification, clustering, prédiction).
4. **Évaluation** : Validation des résultats obtenus par des techniques telles que la validation croisée ou les tests statistiques.
5. **Visualisation** : Présentation des résultats sous forme de graphiques ou de tableaux pour faciliter l'interprétation.

Applications du Web Mining

- **Optimiser les moteurs de recherche** : Google et d'autres moteurs utilisent des algorithmes de Web Mining pour améliorer les résultats de recherche.
- **Recherche d’informations spécialisées (Domain-Specific Information Retrieval)** : construire des moteurs de recherche spécialisés dans des domaines précis (comme Google Scholar)
- **Analyse des sentiments et des opinions (Sentiment Analysis)**
- **Recommandation de produits** : Les sites de e-commerce utilisent le Web Mining pour recommander des produits basés sur l'historique de navigation.
- **Publicité ciblée** : Les entreprises utilisent ces techniques pour analyser les comportements des utilisateurs et proposer des publicités adaptées.
- **Détection des fraudes et cybersécurité**
- **Recherche académique et analyse des réseaux sociaux**
- **Éducation et e-learning**
- **……..**

Enjeux et Défis du Web Mining

- **Hétérogénéité des données** : Les données web existent sous diverses formes (texte, images, vidéos), ce qui rend leur extraction et leur analyse complexes.
- **Volume de données** : Le Web génère d'énormes volumes de données quotidiennement. Le traitement de ces données nécessite des algorithmes optimisés et de grandes capacités de calcul.
- **Qualité des données** : Les données web peuvent être bruyantes, incomplètes ou redondantes, nécessitant des techniques de prétraitement.
- **Confidentialité** : La collecte et l'analyse des données d'utilisateur soulèvent des questions de respect de la vie privée et de conformité avec les lois de protection des données (RGPD, par exemple).

Thank you