Introduction:

fouille de données et fouille du web, communauté web et réseaux sociaux.

Rappels

Internet:

TCP/IP

HTTP

URL

Hyperlinks

HTML

Qu'est-ce que le World Wide Web ?

Le web est une application de l’internet qui permet de consulter, avec un navigateur, des pages accessibles sur des sites.(partage de documents) puis est devenu une plateforme sur laquelle sont développées des nouvelles technologies. Les bases de ces technologies sont le protocole HTTP et le format de document HTML et URL.

Le Web est l’une des applications d’Internet, comme d’autres applications comme le courrier électronique, la visioconférence et le partage de fichiers en pair à pair

Modèle client-serveur.

Caractéristiques des données Web

- **La taille énorme:** Le volume du Web croit de manière exponentielle
- **L'hétérogénéité**: des données textuelles, d'images, de fichiers audio et vidéo, ainsi que des programmes dans la même page
- **La distribution**: Les données se trouvent éparpillées géographiquement sur des ordinateurs et des plateformes
- **La non structuration**: il intègre conjointement les données structurées telles que les bases de données, les données semi-structurées, dites a auto description, tels que les documents HTML ou XML, et les données non structurées comme les documents textuels libres.
- **La dynamicité**: des données et des liens différents et nouveaux sont perpétuellement ajoutés, mise à jour, ou supprimés

Qu'est-ce que le Data Mining?

Le Data mining connue aussi sous les expressions de fouille de données, exploration de données ou encore extraction de connaissances à partir de données, a pour objet de développer des méthodes et des outils permettant d'automatiser le processus de l’extraction des connaissances et la découverte de modèles dans de grandes quantités de données.

La source de données peut être databases, texts, images, the Web, etc.

Le DM est un processus non trivial d’extraction, à partir de gros ensembles de données de l’information valide, compréhensible, préalablement inconnue et potentiellement utile pour l’utilisateur. (Fayyad et al., 1996)

Data mining est un domaine multidisciplinaire impliquant l'apprentissage automatique, les statistiques, les bases de données, l'intelligence artificielle, la recherche d'informations et la visualisation.

Tâches du datamining 1/2

Les Tâches les plus utilisées sont:

Supervised learning: L'objectif principal est de prédire ou de classifier des données en fonction d'un ensemble d'exemples d'entraînement étiquetés.

La classification: Il s'agit de la catégorisation des données en classes ou catégories distinctes. Par exemple, vous pouvez classifier des courriels comme spam ou non spam en fonction de diverses caractéristiques.

L’estimation (Regression): consiste à prédire une valeur numérique en fonction des données disponibles. Par exemple, prédire le prix d'une maison en fonction de ses caractéristiques.

Association : consiste à découvrir des relations fréquentes entre les éléments d'un ensemble de données. Par exemple, identifier les produits fréquemment achetés ensemble dans un magasin.

Tâches du datamining 2/2

Unsupervised learning: vise à découvrir des structures ou des modèles cachés dans les données sans l'aide d'étiquettes ou de réponses connues.

La Segment (Clustering): Elle permet le partitionnement ou le regroupement des données similaires en groupes, ce qui peut révéler des structures cachées ou des similitudes entre les données.

Analyse de séries temporelles : consiste à identifier des tendances, des modèles saisonniers et des variations dans les données chronologiques, ce qui est crucial dans des domaines tels que la prévision financière.

Détection d'anomalies : la recherche d'éléments de données qui diffèrent considérablement du reste des données. Cela peut être utile pour détecter des fraudes ou des erreurs.

Les étapes d’application de data minig

1. Compréhension du domaine d’application (cerner les Objectifs)
2. Préparation des données
    1. Recueil de données
    2. Nettoyage
    3. Intégration
    4. Sélection
    5. Transformation
3. Fouille de données
    1. Définition des taches
    2. Choix des algorithmes
    3. Fouille
4. Analyse des résultats
    1. Présentation et interprétation des Formes extraites
    2. Évaluation et validation
5. Exploitation des résultats

Web mining

La fouille du Web est le processus d'extraction de connaissances à partir du Web. Cad l'application de techniques de la fouille de données pour découvrir des modèles à partir du Web.

il peut être utilisées à diverses fins, notamment :

- Amélioration de la prise de décision
- Recherche d'informations
- Personnalisation
- Veille stratégique :
- Détection de fraudes et de cybermenaces

Les catégories du Web mining

Selon les objectifs d'analyse, le Web Mining pourrait être classé en trois catégories.

- la fouille de l'usage du web:

  Révéler les modèles d'accès sous-jacents à partir des données de transaction Web ou de session utilisateur enregistrées dans les fichiers journaux Web (log)

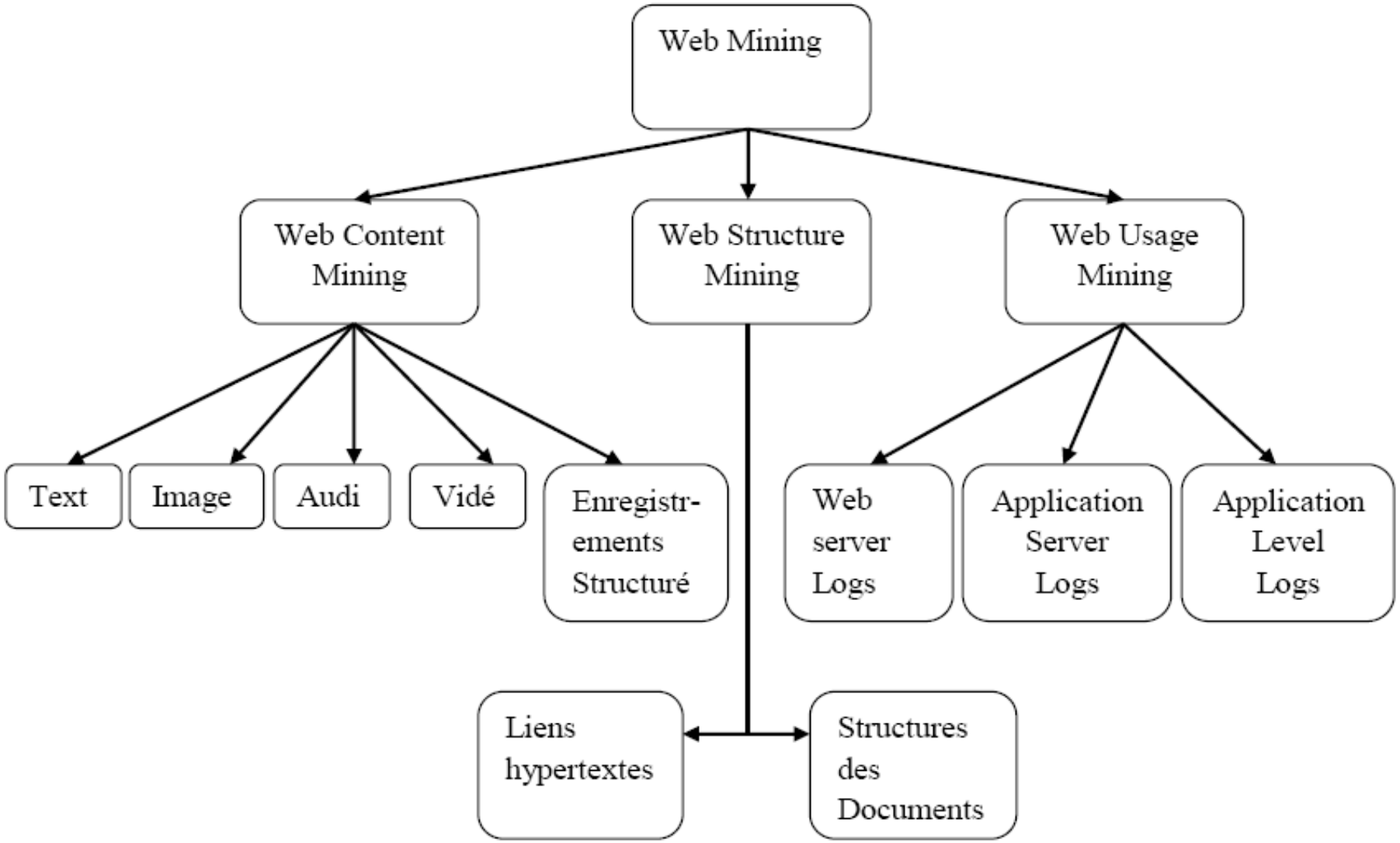
- la fouille du contenu du web

  Découvrir des informations précieuses à partir de contenus Web (c'est-à-dire des documents Web). également appelé text mining

- la fouille de la structure du web.

  Consiste à modéliser des sites Web en termes de structures de liaison. Les informations de liaison mutuelle obtenues pourraient être utilisées pour construire des communautés de pages Web ou trouver des pages pertinentes en se basant sur la similitude ou de la connection entre deux pages Web.

La taxonomie du Web mining



Communauté web

- la communauté Web est définie comme une agrégation d’objets Web en termes de pages Web ou d’utilisateurs, dans laquelle chaque objet est lié à l’autre sous un certain espace de distance.
- Contrairement à la gestion de base de données conventionnelle dans laquelle les modèles de données et les schémas sont définis, une communauté Web est une autre approche efficace pour réorganiser les objets Web, prendre en charge la récupération d’informations et implémenter diverses applications.
- L'analyse des communautés web peut être utile pour diverses applications, telles que la recherche d'informations, l'optimisation des moteurs de recherche, la détection de spam, la catégorisation de sites web, l'analyse de la structure du web, et bien d'autres.

Exemple : un groupe de pages Web, au sein duquel tous les membres partagent une topologie de lien hypertexte similaire vers une page Web spécifique.

Réseaux sociaux

- Récemment, le développement du Web 2.0, des services et des applications Web de plus en plus avancés émergent pour que les utilisateurs Web puissent facilement générer et distribuer des contenus Web et partager facilement des informations dans un environnement collaboratif.
- La composante principale du Web 2.0 est constituée de communautés Web et de services hébergés, tels que les sites de réseaux sociaux, les wikis et les folksonomies, qui se caractérisent par les caractéristiques de communication ouverte, de décentralisation de l'autorité et de liberté de partage et d'autogestion. Ces fonctionnalités Web nouvellement améliorées permettent aux utilisateurs Web de partager et de localiser facilement les contenus Web nécessaires, de collaborer et d'interagir socialement les uns avec les autres, et de réaliser librement l'utilisation et la gestion des connaissances sur le Web