Web Datamining

Information Retrieval and Machine Learning for the Web

Guibert J. TCHINDE
Data Scientist | Solocal group - 2014

PagesJaunes

Intro

- Le séminaire de webmining
- Introduction au webmining
- Programme du cours
- Le logiciel R pour le webmining
- les choses essentielles qu'il faudra retenir
- Les liens pertinents

PagesJaunes

On dira Webmining ou web-datamining

PagesJaunes

Guibert J. TCHINDE
gtchinde@pagesjaunes.fr
@GjTch
github.com/gtanalytics
sciencendata.wordpress.com

> Master de mathématiques appliquées
> DEA d'Epistémologie et Philosophie

—

> Géostatistiques
> Marketing Quantitatif
> Machine Learning
> Big Data

—

> Technologies autour de la transformation digitale
> Le paradigme NoSQL
> La stratégie autour des données

PagesJaunes

- Le séminaire de webmining
- Introduction au webmining
- Programme du cours
- Le logiciel R pour le webmining
- les choses essentielles qu'il faudra retenir
- Les liens pertinents

Séminaire de Webmining

- 24 heures
  - 18 heures de pratiques et 6 heures de théorie
  - De Novembre à Décembre (tous les lundis)

- Outils & Techniques
  - On utilisera essentiellement le logiciel R
  - Techniques de datamining appliquées au web

- Les cours, Les TP
  - github : www.github.com/gtanalytics

- Evaluation : Il sera noté en priorité votre implication
  - QCM : 20% de la note finale
  - 20% x 4 Travaux Pratiques (fait en séance !)

PagesJaunes

- Le séminaire de webmining
- Introduction au webmining
- Programme du cours
- Le logiciel R pour le webmining

- les choses essentielles qu'il faudra retenir
- Les liens pertinents

Introduction au Webmining
Définitions

- Web Data-mining

  - Applications des techniques de datamining aux données provenant du WEB
  - Subdivisé en général en trois sous domaines
    - Web Content mining : Comment décrire les contenus des pages web ?
    - Web Usage Mining :Comment se comportent les utilisateurs du we
    - Web Structure Mining : Comment est structuré le web ?

- Applications

  - E-commerce : Transactions commerciales par le biais d'interactions digitales
  - Search : PagesJaunes, Google
  - Détection de fraude
  - etc…

PagesJaunes

Introduction au Webmining
Histoire du web

- Inventeur du web
  - Tim Berners-Lee : http://fr.wikipedia.org/wiki/Tim_Berners-Lee
- Quelques dates :
  - 1989 : Invention du WEB
  - 1990 : Ecriture du premier serveur
  - 1993 : Mosaic Browser (un des premiers moteurs web)
  - 1998 : HTML4
- Web 2.0 : Interactivité (social)
- Web 3.0 : Sémantique

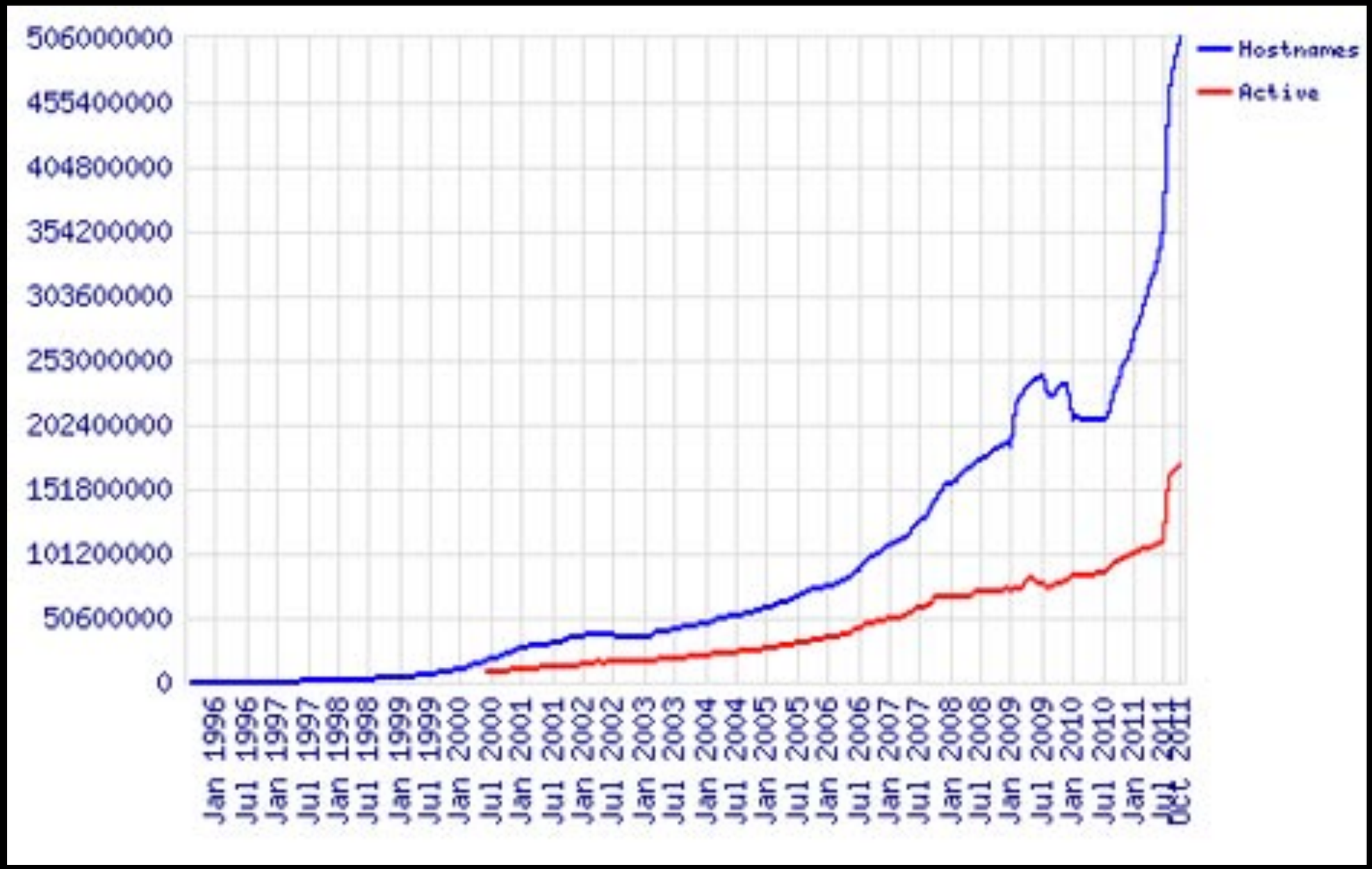
http://www.evolutionoftheweb.com/?hl=fr

- le Web aujourd'hui
  - Infini ?
  - Le graphe est sans doute la meilleure représentation du web
  - Les grands acteurs

    Google, Facebook, AT&T, Yahoo, Microsoft, Apple, Amazon

PagesJaunes

Plus de 100 millions de sites web dans le monde en 2013



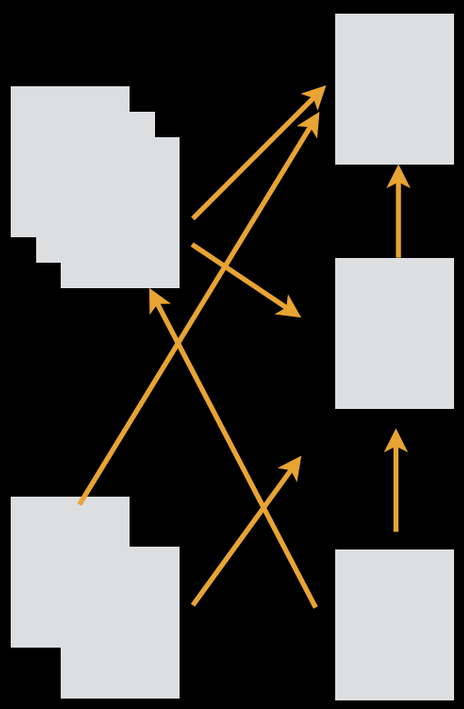
Introduction au Webmining
Information Retrieval

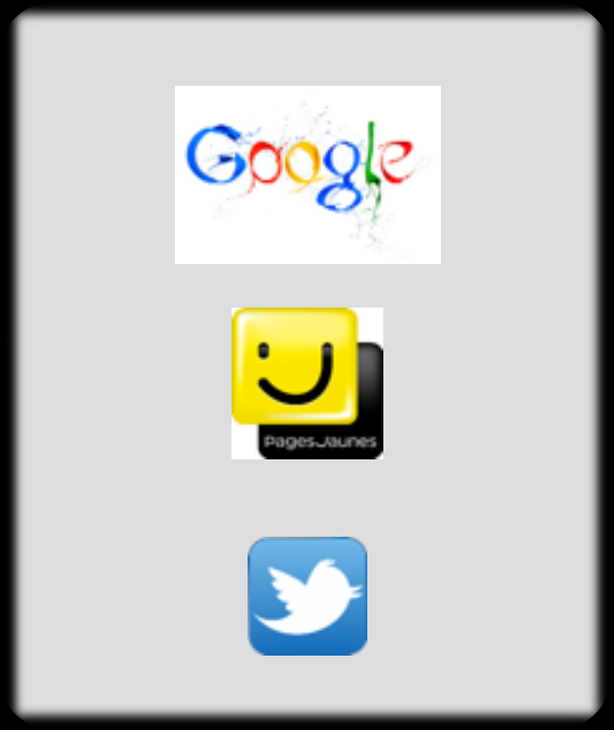
- Recherche d'informations

WEB

Aggregateurs de contenus

User





PagesJaunes

Introduction au Webmining
Les mots du web mining

- Crawling
- Information Retrieval | Web Search
- Web graph analysis
- Structured data extraction
- Classification and vertical search
- Collaborative filtering
- Web advertising and optimisation
- Mining web logs
- Opinion Mining

Prenez 15 minutes pour vous familiariser avec toutes ces notions en recherchant sur le web et ensuite, partageons ensemble vos connaissances

PS : Trois de ces définitions feront parties des QCM

PAUSE

Introduction au Webmining
Big Picture du web search

Le search est probablement la plus grande application sur le Web. Il a sa racine dans la recherche d'information qui est un domaine d'étude qui aide l'utilisateur à trouver l'information nécessaire à partir d'une grande collection de documents de texte.

Compte tenu de la requête (par exemple, un ensemble de mots clés), qui exprime le besoin d'information de l'utilisateur, un système infrarouge trouve un ensemble de documents qui sont pertinents à la requête de sa collection sous-jacente : Moteur de recherche

Caractéristiques particulières des données Web
Tout d'abord, les pages Web ne sont pas les mêmes que les documents en texte brut, car ils sont semi-structurées et contiennent des liens. Ainsi, de nouvelles méthodes ont été conçus pour produire de meilleurs systèmes de recherches.

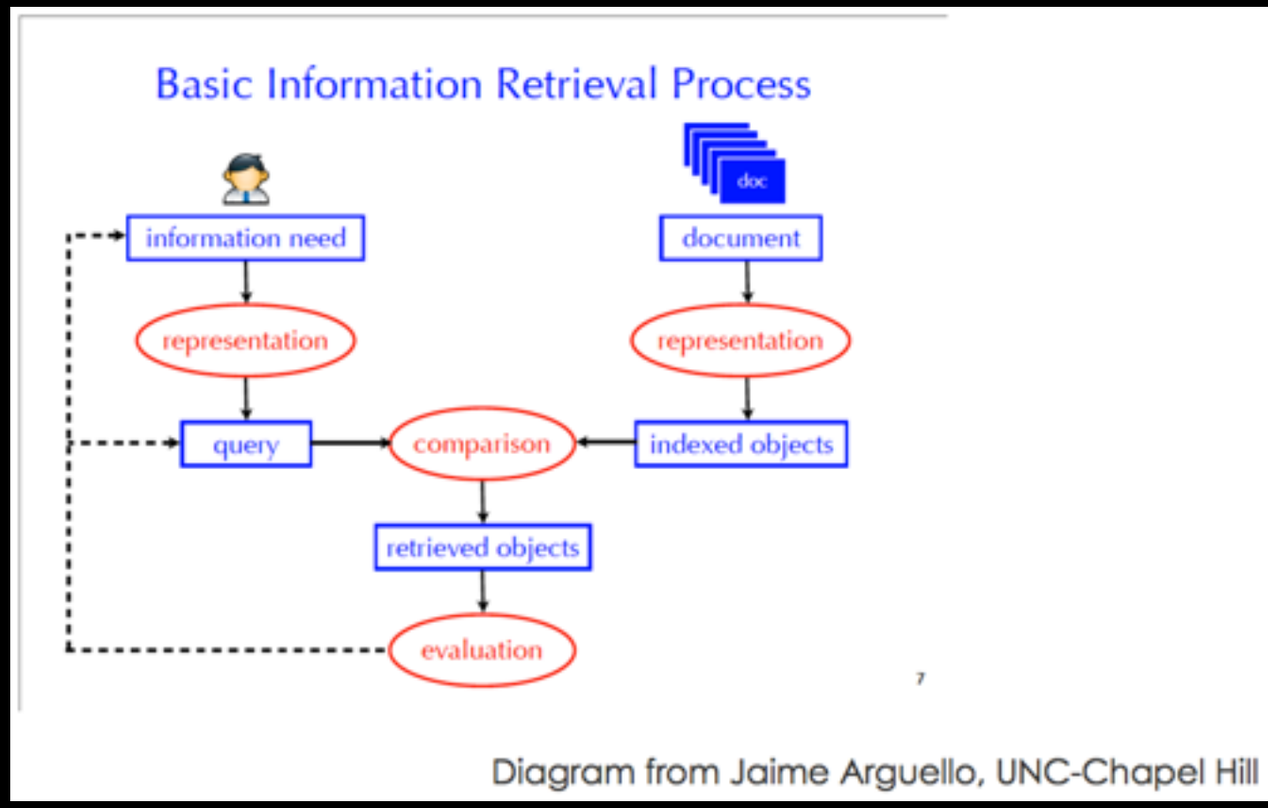
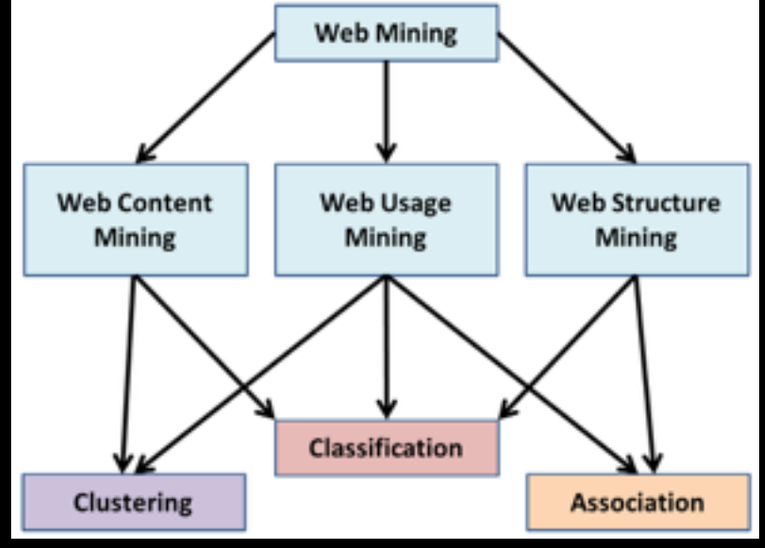


Diagram from Jaime Arguello, UNC-Chapel Hill

Introduction au Webmining
Big Picture du web mining - 1

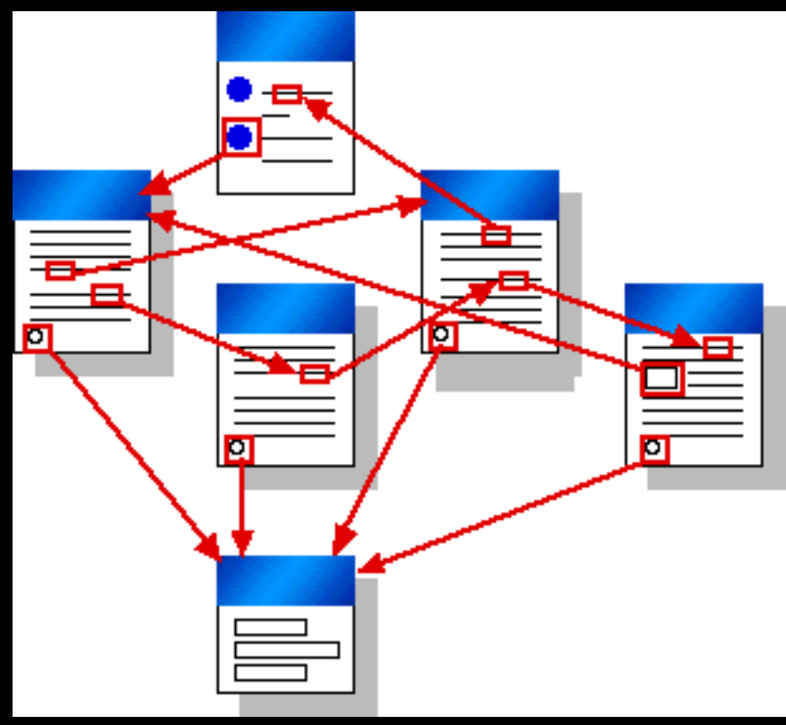


Web content mining

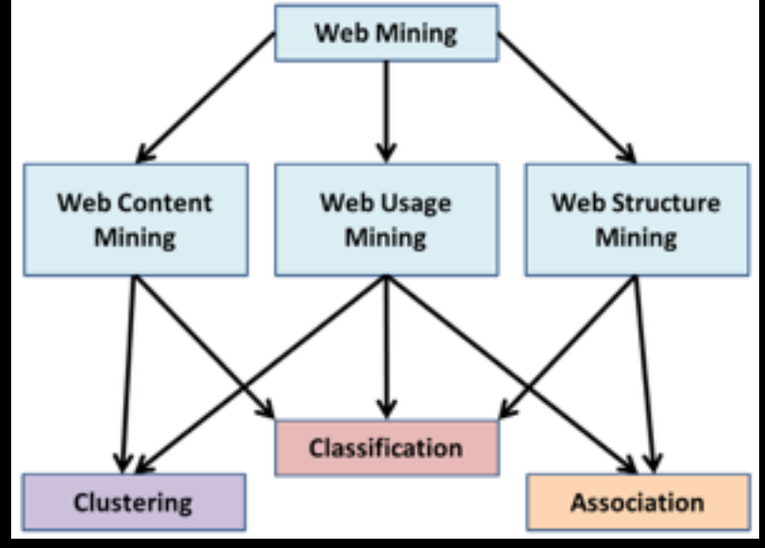
- Definition :
  - C'est le processus d'extraction d'information utile à partir du contenu des documents web. Le contenu d'un document correspond aux informations dont la page web est conçue pour les transférer aux utilisateurs. Ça peut être du texte, des images, des vidéos ou des enregistrements structurés comme les listes et les tables.

- Principales techniques :
  - Les principales issues adressées dans le texte mining sont : la découverte de catégorie, l'extraction des modèles d'association, le clustering (regroupement) des documents web et la classification des pages web.

- Applications
  - Identification des documents Web
  - Catégorisation
  - Trouver des pages webs qui traitent de la même question sur d'autres serveurs
  - Mesurer la pertinence
  - Filtrer et Inférer



Introduction au Webmining
Big Picture du web mining - 3



Web usage mining

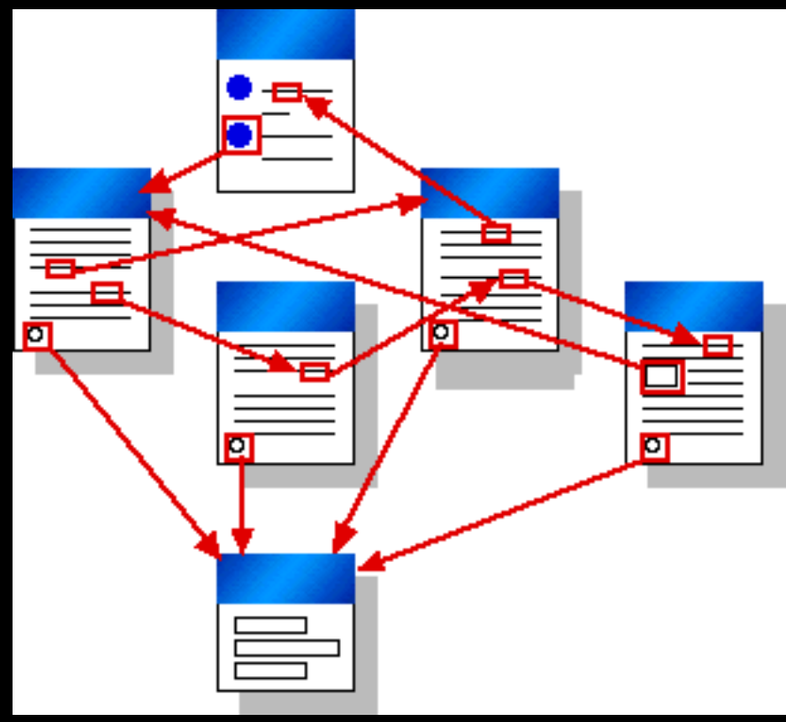
- Definition :

Le Web Structure Mining est le processus de découvrir la structure d'information à partir du web. Ça peut être divisé en deux types, en se basant sur le type de la structure d'information utilisée :

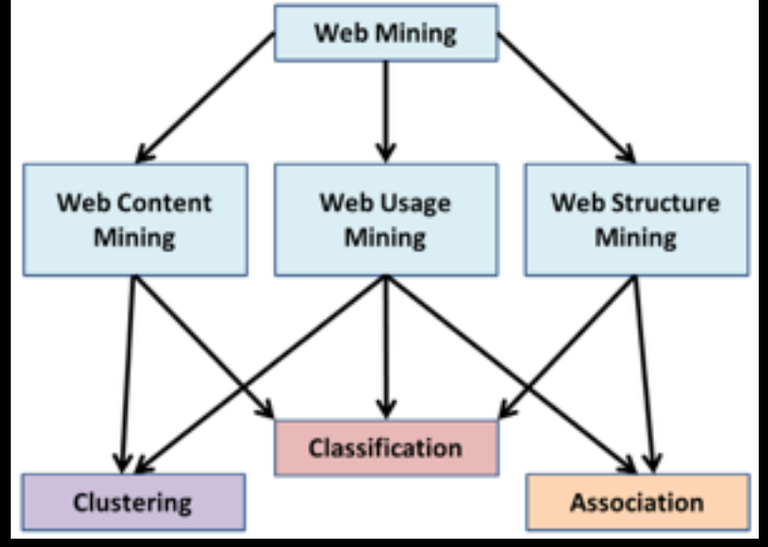
- Lien Hypertexte : Unité structurelle qui connecte un emplacement à un autre dans une page web
- Structure de document : Le contenu d'un page peut être organisé dans le format d'un arbre structuré (tags HTML ou XML)

- Principales techniques :
  - Les règles d'association : Si (A et B) => C

- Applications
  - Qualité d'une page web(SEO)
  - Pages reliées
  - Structure d'un site web



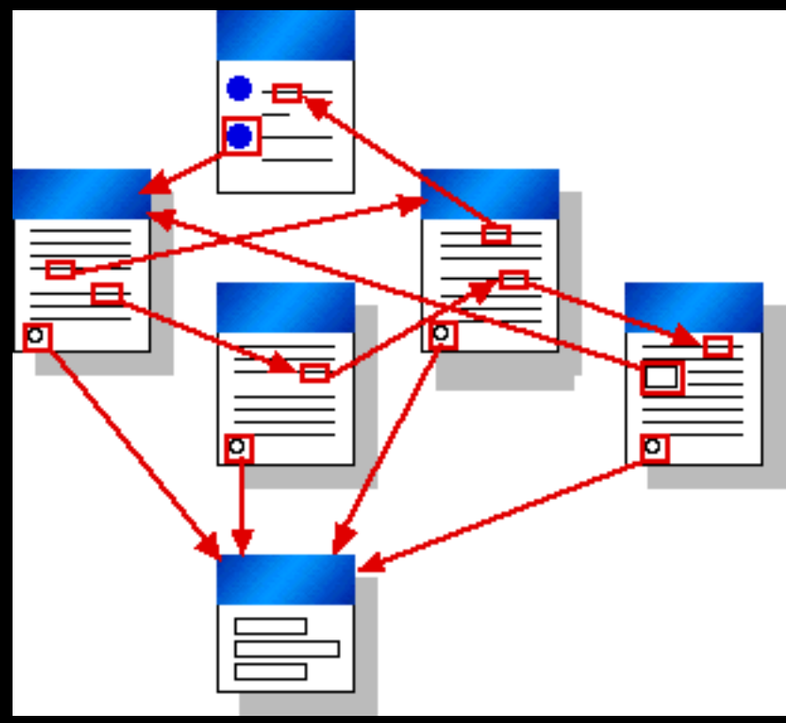
Introduction au Webmining
Big Picture du web mining - 2



Web structure mining

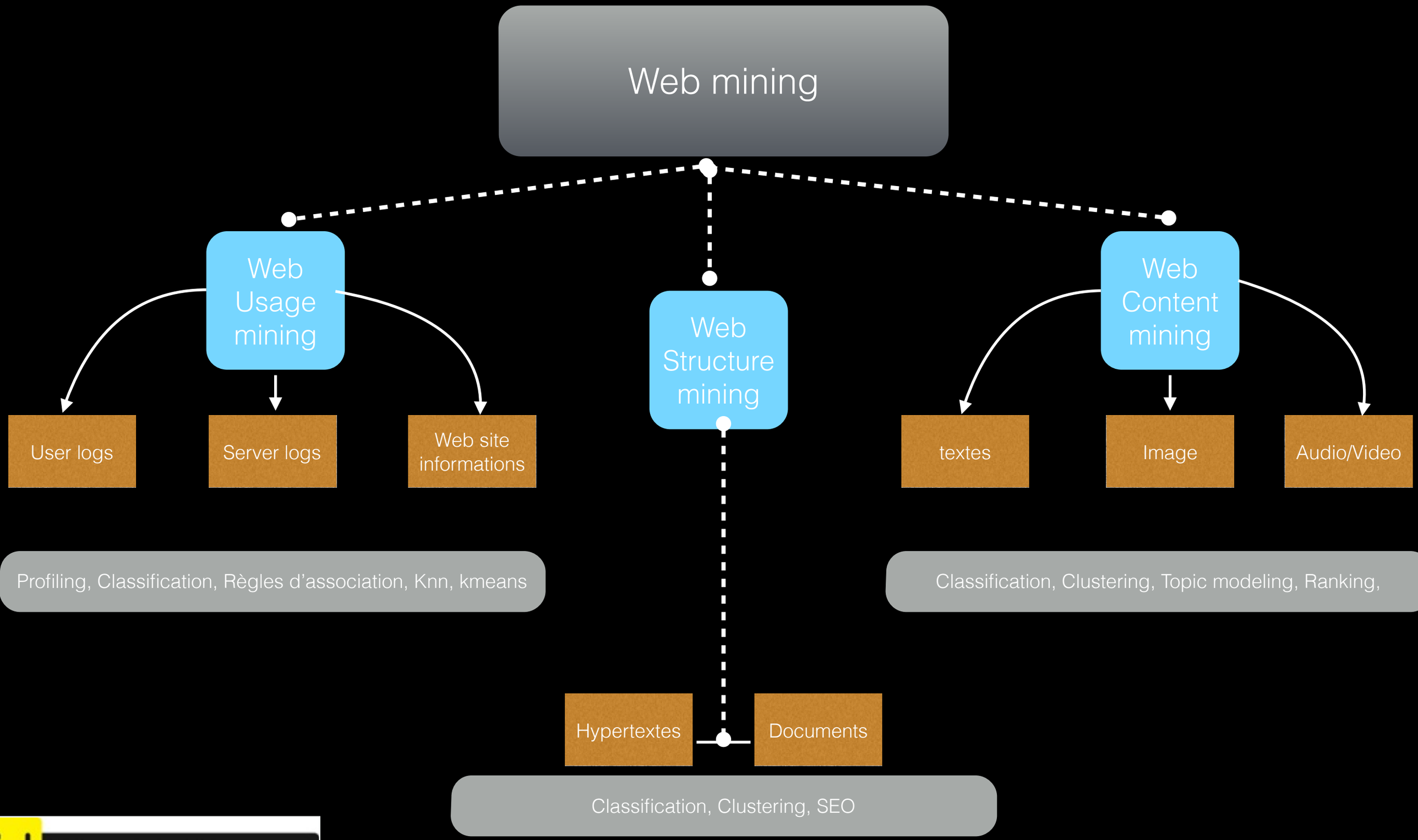
- Definition :
  - C'est l'application des techniques du data Mining pour découvrir des modèles d'usage intéressants à partir des données du web, et cela, pour mieux comprendre et bien servir les besoins des applications web. Les données d'utilisation capturent l'identité et l'origine des utilisateurs web tout au long de leurs navigations dans un site web.
  - User Logs : Infos de session, Pages visitées, @IP
  - ~~Données des serveurs d'application~~

- Principales techniques :
  - Les règles d'association, les motifs séquentiels, Classification



- Applications
  - Extraire et organiser l'information
  - Stockage (Matrices, NoSQL, BDD structurées, etc…)
  - Data mining : Fouille et analyse de données
  - Analyse et interprétation

Introduction au Webmining
Big Picture : Wrap up



- Le séminaire de webmining
- Introduction au webmining
- Programme du cours
- Le logiciel R pour le webmining

- les choses essentielles qu'il faudra retenir
- Les liens pertinents

Programme du cours

- Programme
  - https://github.com/gtanalytics/ensai/tree/master/webdatamining
- Organisation
  - Chaque vendredi : Un lien Github avec le cours du lundi
  - Chaque lundi : Un lien avec les corrections des TP
- Formation des binômes pour les travaux pratiques

PagesJaunes

- Le séminaire de webmining
- Introduction au webmining
- Programme du cours
- Le logiciel R pour le webmining

- les choses essentielles qu'il faudra retenir
- Les liens pertinents

Le logiciel R pour le webmining

- R
  - Logiciel Open Source de statistiques le plus utilisé dans le monde
  - Développé par une communauté : http://cran.r-project.org/
  - De plus en plus utilisé dans les entreprises
  - Plus de 6000 librairies complémentaires
- Tutoriels autour de R
  - Voir https://github.com/gtanalytics/ensai/tree/master/R-Init
- R pour le webmining
  - text processing library : base, XML, stringr, stringdist,
  - Web requests: RCurl, httr,rjson, RJSONIO,
  - Documentation
    - http://cran.r-project.org/web/packages/tm.plugin.webmining/vignettes/ShortIntro.pdf
    - http://cran.r-project.org/doc/contrib/Zhao_R_and_data_mining.pdf

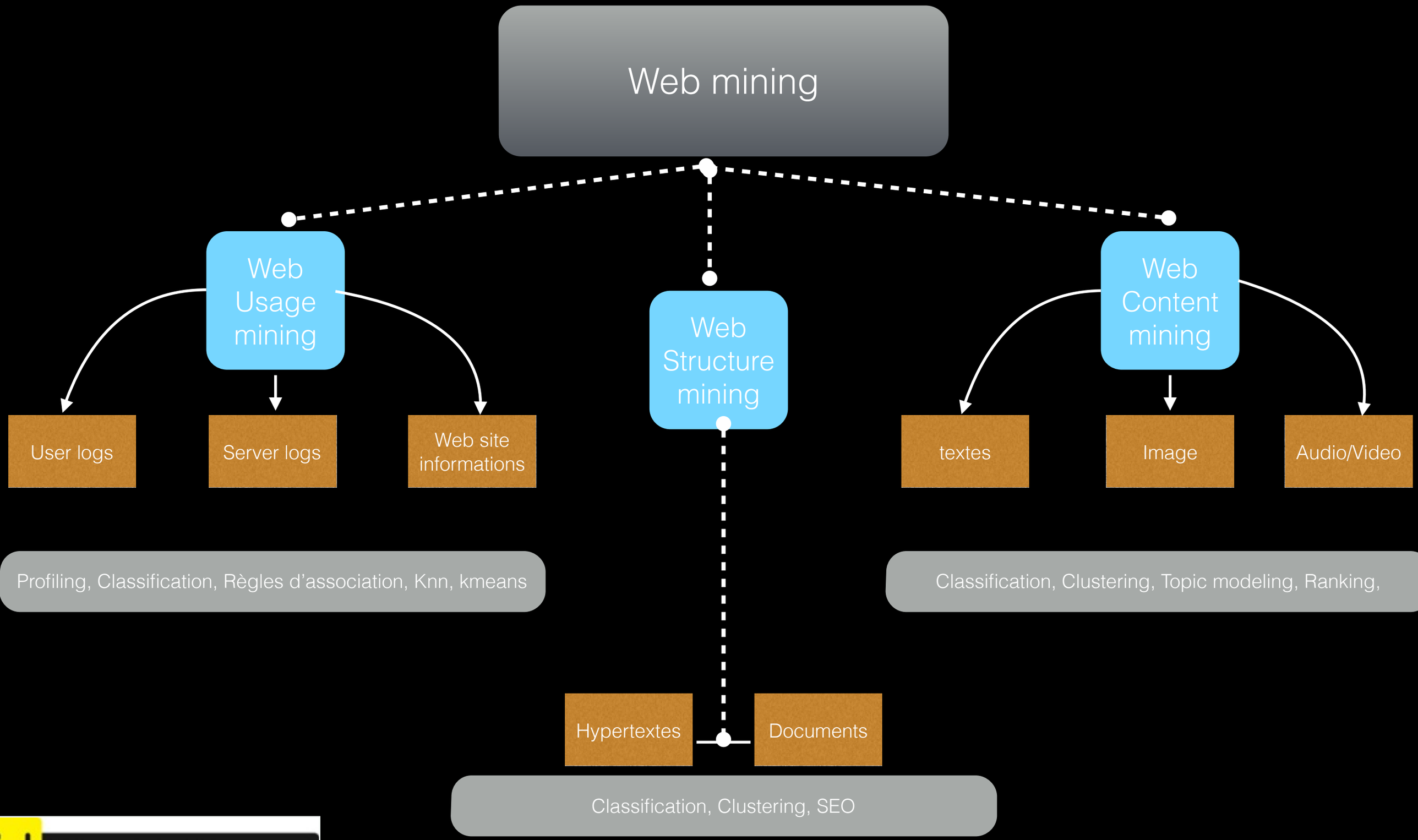
Prenez 30 minutes pour :

- Installer R
- Installer toutes les librairies utiles au web-mining
- Lire les vignettes des principales librairies pour le webmining

- Le séminaire de webmining
- Introduction au webmining
- Programme du cours
- Le logiciel R pour le webmining

- les choses essentielles qu'il faudra retenir
- Les liens pertinents

A retenir
Le webmining c'est…



- Le séminaire de webmining
- Introduction au webmining
- Programme du cours
- Le logiciel R pour le webmining
- les choses essentielles qu'il faudra retenir
- Les liens pertinents

Documentation

- Ouvrages de référence
  - Web DataMining, Exploring Hyperlinks, Contents, and Usage Data, Bing Liu, Springer (Chapitre 6 à 13)
  - Information Retrieval, [http://nlp.stanford.edu/IR-book/pdf/irbookonlinereading.pdf (chapitres 1-3)]

- Documents de référence
  - Infrastructure of Texmining with R, http://www.jstatsoft.org/v25/i05/paper
  - PageRank, [http://ilpubs.stanford.edu:8090/422/1/1999-66.pdf]
  - Mining the social web, [https://github.com/ptwobrussell/Mining-the-Social-Web]

- Google

- Stackoverflow

PagesJaunes

Intro

- Le séminaire de webmining
- Introduction au webmining
- Programme du cours
- Le logiciel R pour le webmining
- les choses essentielles qu'il faudra retenir
- Les liens pertinents

Web Datamining

Information Retrieval and Machine Learning for the Web

Guibert J. TCHINDE
Data Scientist | Solocal group - 2014

PagesJaunes