Fouille d’usage web:

modélisation d’intérêts de l’internaute par clustering, analyse de sémantique latente, découverte des patterns d’accès de l’utilisateur, exploitation des fichiers logs (weblogs).

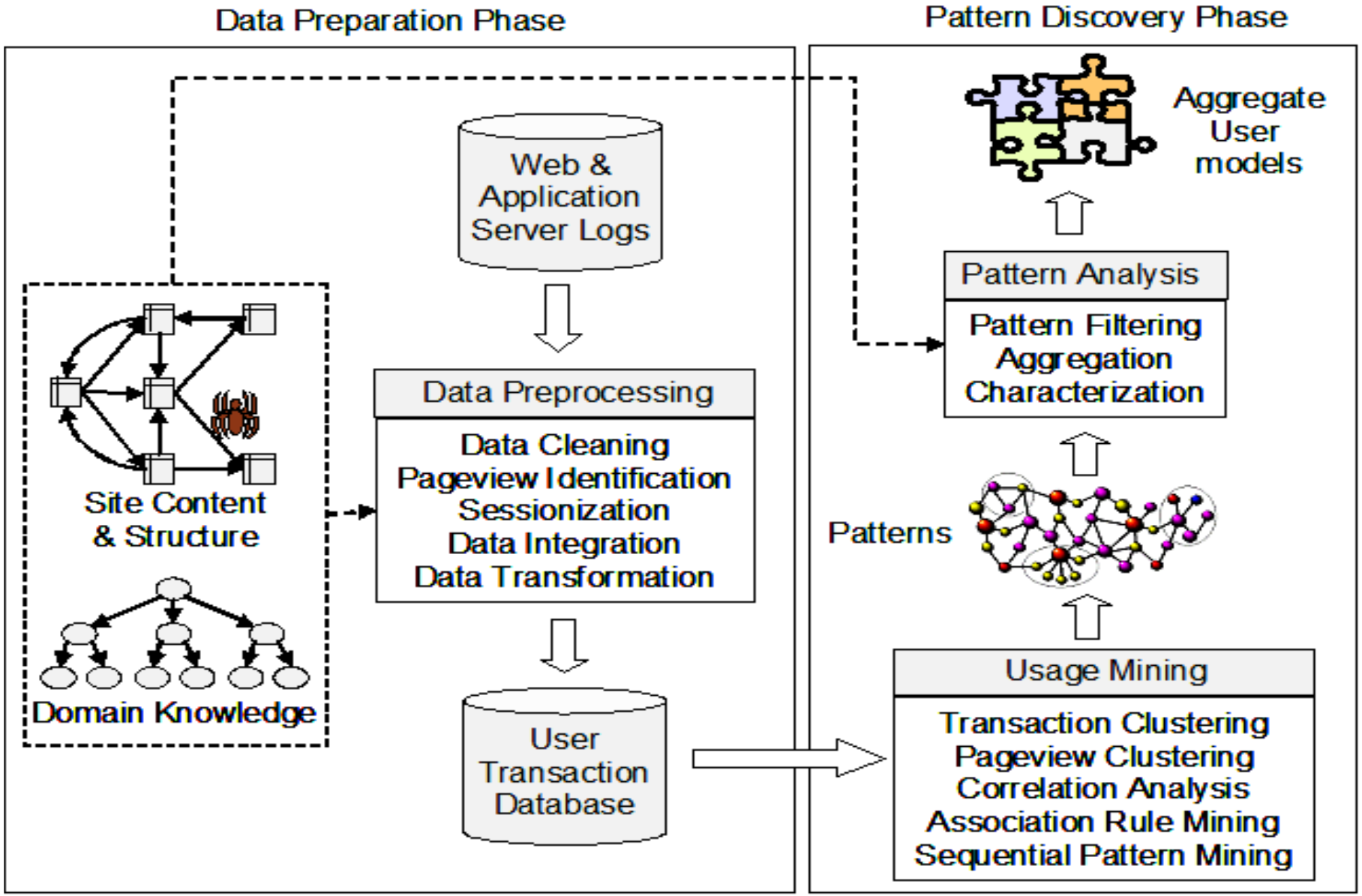
Le web usage mining (WUM)

Exploration de l'utilisation du Web : est une branche du Web Mining qui se centre sur la découverte automatique des modèles dans les flux de clics et des données associées, collectées ou générées à la suite d'interactions de l'utilisateur avec un ou plusieurs sites Web.

Objectif : analyser les modèles de comportement et les profils des utilisateurs interagissant avec un site Web, afin de comprendre et de mieux répondre aux besoins des applications Web.

- Il est utilisé de plus en plus par de nombreuses entreprises et par les propriétaires de sites, afin de mesurer leurs fréquentations, suivre (tracker) leurs utilisateurs, anticiper leurs besoins, et offrir des contenus adaptés et personnalisés.
- Le web usage mining est utilisée par les sites de commerce électronique pour organiser leurs sites et augmenter leurs profits.
- Recommandation
- permet aux développeurs d'un site Web de repenser leur design
- Il est utilisé par les moteurs de recherche pour améliorer la qualité de la recherche et pour évaluer les résultats de la recherche.

Processus du web usage mining



Processus du web usage mining

Ce processus se compose de quatre phases :

1. **Input stage**: les fichiers journaux Web bruts sont récupérés ainsi que des informations d'enregistrement (le cas échéant) et des informations concernant la topologie du site.
2. **Preprocessing stage**: Les tâches de prétraitement les plus courantes sont (1) le nettoyage et le filtrage des données, (2) le de-spidering, (3) l'identification de l'utilisateur, (4) l'identification de la session et (5) l'achèvement du chemin.
3. **Pattern discovery stage**: Ces méthodes comprennent (1) l'analyse statistique standard, (2) les algorithmes de clustering, (3) les règles d'association, (4) les algorithmes de classification et (5) les patterns séquentiel.
4. **Pattern analysis stage**: Les analystes humains examinent les résultats de l'étape de découverte de motifs et en extraient les motifs les plus intéressants, utiles et exploitables.

Recueil de données

Les données les plus communément exploitées sont les fichiers log, les données issues des procédures d'inscription, et les données sur la structure et le contenu des sites .

Le WUM est basé sur les données de trafic collectées à partir de trois sources :

- **les serveurs Web**: les traces des requêtes effectuées par les utilisateurs sont automatiquement enregistrées dans des fichiers journaux log.
- **les serveurs proxy**: Un proxy est une machine intermédiaire placée entre le client et le serveur, qui permet d'acheminer indirectement les requêtes de l'un vers l'autre. Ces machines servent à plusieurs fonctions (offrir des connexions Web à un groupe d'utilisateurs, cache, pare-feu …).

les clients Web: plugins …

Analyse du Flux de Clics « **Clickstream** »

- Un **clickstream** est la séquence agrégée de visites de pages exécutées par un utilisateur particulier sur un site Web.
- Toutes les **hits** (l’ensemble des téléchargements) doivent être agrégées en pages view au stade du prétraitement. Ensuite, une série de pages consultées peuvent être regroupées en une session.
- Les données du parcours de navigation nécessitent un prétraitement avant de pouvoir analyser le comportement de l'utilisateur.

Questions

Une fois que les données du parcours de navigation ont été prétraitées, les analystes peuvent commencer à s'attaquer à des questions telles que les suivantes :

- Quelle page web est le point d'entrée le plus courant pour les utilisateurs ?
- Qu'est-ce qui les fait venir ? (référencement)
- Dans quel ordre les pages ont-elles été consultées ? (web structure)
- Comment utilisent-ils vos services ?
- Combien de pages web ont été consultées lors d'une visite typique ?
- Combien de temps le visiteur type reste-t-il sur notre site Web ?
- Comment améliorer leur expérience ?
- Quelle page web est le point de départ le plus courant pour les utilisateurs ?

répondre à ces questions il faut utiliser Les outils de web métriques

Web server logs

1	2006-02-01 00:08:43 1.2.3.4 - GET /classes/cs589/papers.html - 200 9221 HTTP/1.1 maya.cs.depaul.edu Mozilla/4.0+(compatible;+MSIE+6.0;+Windows+NT+5.1;+SV1;+.NET+CLR+2.0.50727) http://dataminingresources.blogspot.com/
2	2006-02-01 00:08:46 1.2.3.4 - GET /classes/cs589/papers/cms-tai.pdf - 200 4096 HTTP/1.1 maya.cs.depaul.edu Mozilla/4.0+(compatible;+MSIE+6.0;+Windows+NT+5.1;+SV1;+.NET+CLR+2.0.50727) http://maya.cs.depaul.edu/~classes/cs589/papers.html
3	2006-02-01 08:01:28 2.3.4.5 - GET /classes/ds575/papers/hyperlink.pdf - 200 318814 HTTP/1.1 maya.cs.depaul.edu Mozilla/4.0+(compatible;+MSIE+6.0;+Windows+NT+5.1) http://www.google.com/search?hl=en&lr=&q=hyperlink+analysis+for+the+web+survey
4	2006-02-02 19:34:45 3.4.5.6 - GET /classes/cs480/announce.html - 200 3794 HTTP/1.1 maya.cs.depaul.edu Mozilla/4.0+(compatible;+MSIE+6.0;+Windows+NT+5.1;+SV1) http://maya.cs.depaul.edu/~classes/cs480/

Pré traitement des données log :

Après ce pré traitement, on peut faire des statistiques simples , les pages les plus visitées, l'ordre de visite des pages, la durée de visite pour chaque page ….ou appliquer les techniques data mining (clustering et règles d'association pour découvrir des patterns d'utilisation (profil) ou détecter des groupes d'utilisateur….

Nettoyage des données:

- Extraction de variables
- Supprimer les références et les champs non pertinents dans les journaux du serveur.
- Supprimer l'ensemble des données les demandes automatiques généré par la page web
- Supprimer les références dues à la navigation du spider crawler.
- Supprimer les références erronées.
- Ajouter les références manquantes en raison de la mise en cache (effectué après la sessionisation).

Identification de l’utilisateur

- Adresse IP + agent
- Inscription Login
- Cookie
- Plug-in

La procédure d’identification d’ utilisateur :

1. Trier le fichier web log par adresse IP puis par temps (time stamp)
2. Pour chaque adresse IP différente , identifier chaque agent comme appartenant à un user différent
3. Pour chaque utilisateur identifié dans l’étape 2 , appliquer l’information du champ référer et la structure du site pour déterminer si le comportement est le résultat d’un seul user ou plusieurs.
4. Pour identifier chaque user, combiner l’information du user extraite à partir des étapes 1 à 3 avec le fichier cookies et les infos username .

Identifier les sessions (sessionisation) :

- les activités effectuées par un utilisateur à partir du moment où il entre sur le site jusqu'au moment où il le quitte.

Procédure d'identification de session

1. Pour chaque user distinct identifié , assigner un ID unique.
2. Définir le seuil timeout t
3. Pour chaque user,

a. trouver le temps séparant entre deux entrées consécutives dans le fichier web log

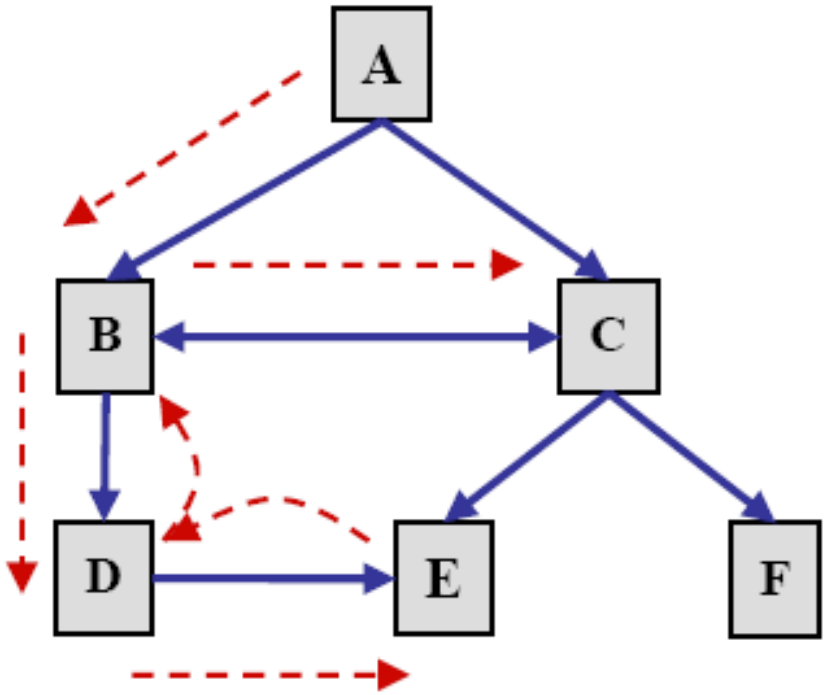
b. Si cette différence de temps dépasse le time out t , assigner un nouvel ID à l''entrée dernière.

4. Trier les entries par session ID

Difficile d'obtenir des données d'utilisation fiables en raison des serveurs proxy et des anonyma, des adresses IP dynamiques, des références manquantes dues à la mise en cache et de l'incapacité des serveurs à distinguer les différentes visites.

Achèvement du chemin

- La mise en cache côté client ou proxy peut souvent entraîner l'absence de références d'accès aux pages ou aux objets qui ont été mis en cache.



User's actual navigation path:

A → B → D → E → D → B → C

What the server log shows:

URL	Referrer
A	--
B	A
D	B
E	D
C	B

Fig. 12.7. Missing references due to caching.

modélisation d’intérêts de l’internaute par clustering

Le clustering de sites Web est l'une des techniques les plus utilisées dans le contexte de l'exploration du Web, qui consiste à regrouper des objets Web similaires, tels que des pages Web ou des sessions d'utilisateurs, en un certain nombre de groupes d'objets en mesurant leur distance vectorielle mutuelle.

L'objectif de la recherche d'intérêts similaires parmi les utilisateurs du Web est de découvrir des connaissances à partir du profil de l'utilisateur. Si un site Web est bien conçu, il y aura une forte corrélation entre la similarité des chemins de navigation et la similarité des intérêts des utilisateurs. Par conséquent, le regroupement des premiers pourrait être utilisé pour regrouper les seconds.

La fonction de similarité peut être basée sur la visite de pages identiques ou similaires, sur la fréquence d'accès à une page, ou même sur l'ordre de visite des liens.