Fouille des données de la toile

Web Data Mining

Maria Malek

maria.malek@eisti.fr

Ecole Internationale des Sciences de Traitement de l'Information (EISTI)

Avant-Propos

- La toile : **World Wide Web**

Avant-Propos

- La toile : **World Wide Web**
- "Wide-area Hypermedia Information retrieval initiative aiming to give universal access to a large universe of documents".

Avant-Propos

- La toile : **World Wide Web**
- "Wide-area Hypermedia Information retrieval initiative aiming to give universal access to a large universe of documents".
- *Réseau* : internet, *Modèle* : client serveur, *Navigation* : navigateur.

Avant-Propos

- La toile : **World Wide Web**
- "Wide-area Hypermedia Information retrieval initiative aiming to give universal access to a large universe of documents".
- *Réseau* : internet, *Modèle* : client serveur, *Navigation* : navigateur.
- Structurer les documents : Hypertexte, hyperliens, Hypermédias, etc.

Explorer les données de la toile

- Données - Informations : quantité énorme, diversité, couverture.

Explorer les données de la toile

- Données - Informations : quantité énorme, diversité, couverture.
- Différents types et natures de données.

Explorer les données de la toile

- Données - Informations : quantité énorme, diversité, couverture.
- Différents types et natures de données.
- Information présentée d’une façon hétérogène.

Explorer les données de la toile

- Données - Informations : quantité énorme, diversité, couverture.
- Différents types et natures de données.
- Information présentée d’une façon hétérogène.
- Liens et Hyperliens, **autorité** de certaines pages.

Explorer les données de la toile

- Données - Informations : quantité énorme, diversité, couverture.
- Différents types et natures de données.
- Information présentée d'une façon hétérogène.
- Liens et Hyperliens, **autorité** de certaines pages.
- **Informations bruitées** : pour une application donnée, une partie de la page est considérée.

La toile d’aujourd’hui

- La toile "propose" des services : commande de produits, paiement de factures.

La toile d’aujourd’hui

- La toile "propose" des services : commande de produits, paiement de factures.
- La toile est dynamique, les informations changent d’une façon continue.

La toile d’aujourd’hui

- La toile "propose" des services : commande de produits, paiement de factures.
- La toile est dynamique, les informations changent d’une façon continue.
- La toile est une société virtuelle : interactions entre les internautes, etc.

Exploration & fouille de données

- ”Data Mining”

Exploration & fouille de données

- **”Data Mining”**
- Découverte de connaissances à partir de données
  - Un processus itératif par lequel on extrait des connaissances valides, nouvelles, potentiellement utiles et compréhensibles [Fayyad et al., 1995].

Exploration & fouille de données

- **”Data Mining”**
- Découverte de connaissances à partir de données
  - Un processus itératif par lequel on extrait des connaissances valides, nouvelles, potentiellement utiles et compréhensibles [Fayyad et al., 1995].
- "Comment faire parler les données ? "

Processus de la fouille de données - 1

- Analyse du problème d’application.

Processus de la fouille de données - 1

- Analyse du problème d’application.
- Sélection et exploration des données
  - évaluer la qualité des données,
  - visualiser, analyser les distributions et les regroupements,
  - détecter les insuffisances, pathologies des données.

Processus de la fouille de données - 1

- Analyse du problème d’application.
- Sélection et exploration des données
  - évaluer la qualité des données,
  - visualiser, analyser les distributions et les regroupements,
  - détecter les insuffisances, pathologies des données.
- Pré-traitement des données

Processus de la fouille de données - 1

- Analyse du problème d'application.
- Sélection et exploration des données
  - évaluer la qualité des données,
  - visualiser, analyser les distributions et les regroupements,
  - détecter les insuffisances, pathologies des données.
- Pré-traitement des données

  nettoyage bruit, valeurs manquantes,

Processus de la fouille de données - 1

- Analyse du problème d'application.
- Sélection et exploration des données
  - évaluer la qualité des données,
  - visualiser, analyser les distributions et les regroupements,
  - détecter les insuffisances, pathologies des données.
- Pré-traitement des données

nettoyage bruit, valeurs manquantes,

réduction sélection des instances, extraction, combinaison des variables,

Processus de la fouille de données - 1

- Analyse du problème d'application.
- Sélection et exploration des données
  - évaluer la qualité des données,
  - visualiser, analyser les distributions et les regroupements,
  - détecter les insuffisances, pathologies des données.
- Pré-traitement des données

nettoyage bruit, valeurs manquantes,

réduction sélection des instances, extraction, combinaison des variables,

transformation discrétisation des variables continues, ajout de nouvelles variables (induction constructive).

Processus de la fouille de données - 2

- L’apprentissage «data mining» : une méthode d’extraction de connaissances.

Processus de la fouille de données - 2

- L’apprentissage «data mining» : une méthode d’extraction de connaissances.
- Evaluation et interprétation des résultats : critères différents suivant la tâche.

Résultat du processus

- Connaissances sont extraites sous forme d’un :

Résultat du processus

- Connaissances sont extraites sous forme d'un :

  modèle : un résumé global de l'ensemble de données applicable sur n'importe quelle instance appartenant à l'espace des données ;

Résultat du processus

- Connaissances sont extraites sous forme d'un :

modèle : un résumé global de l'ensemble de données applicable sur n'importe quelle instance appartenant à l'espace des données ;

motif (pattern) : résumé local d'une région de l'espace des données ; exemple : une règle d'association ou de classement.

Applications de données

- Domaines supervisés :

Applications de données

- Domaines supervisés :
- *Chaque instance = p variables prédictives + 1 variable cible (à prédire)*

Applications de données

- Domaines supervisés :
- *Chaque instance = p variables prédictives + 1 variable cible (à prédire)*

  Classification : variable cible discrète *Exemple : diagnostiquer une maladie.*

Applications de données

- Domaines supervisés :
- *Chaque instance = p variables prédictives + 1 variable cible (à prédire)*

  Classification : variable cible discrète *Exemple : diagnostiquer une maladie.*

  Régression : variable cible continue *Exemple : estimer la valeur d'un bien.*

Applications de données

- Domaines supervisés :
- *Chaque instance = p variables prédictives + 1 variable cible (à prédire)*

  Classification : variable cible discrète *Exemple : diagnostiquer une maladie.*

  Régression : variable cible continue *Exemple : estimer la valeur d'un bien.*
- Domaines non supervisés :

Applications de données

- Domaines supervisés :
- *Chaque instance = p variables prédictives + 1 variable cible (à prédire)*

  Classification : variable cible discrète *Exemple : diagnostiquer une maladie.*

  Régression : variable cible continue *Exemple : estimer la valeur d'un bien.*
- Domaines non supervisés :

  Regroupement (clustering) *Exemple : détecter le profil utilisateur.*

Applications de données

- Domaines supervisés :
- *Chaque instance = p variables prédictives + 1 variable cible (à prédire)*

  Classification : variable cible discrète *Exemple : diagnostiquer une maladie.*

  Régression : variable cible continue *Exemple : estimer la valeur d'un bien.*
- Domaines non supervisés :

  Regroupement (clustering) *Exemple : détecter le profil utilisateur.*

  Association *Exemple analyser les logs utilisateurs d'un serveur web.*

Fouille des Données de la Toile

- **Contenu** *(Web Content Mining)* Analyse des contenus des pages web :
  - Classer et/ou segmenter les pages selon le thème.
  - Chercher des descriptions de produits, etc.

Fouille des Données de la Toile

- **Contenu** *(Web Content Mining)* Analyse des contenus des pages web :
  - Classer et/ou segmenter les pages selon le thème.
  - Chercher des descriptions de produits, etc.
- **Comportement** *(Web Usage Mining)* Analyse les traces de navigations des internautes (logs)
  - Algorithmes d'analyse et de traitement de séquences
  - Besoin d'une phase de pré-traitement

Fouille des Données de la Toile

- **Contenu** *(Web Content Mining)* Analyse des contenus des pages web :
  - Classer et/ou segmenter les pages selon le thème.
  - Chercher des descriptions de produits, etc.
- **Comportement** *(Web Usage Mining)* Analyse les traces de navigations des internautes (logs)
  - Algorithmes d'analyse et de traitement de séquences
  - Besoin d'une phase de pré-traitement
- **Structure** *(Web Structure Mining)* Découverte des connaissances à partir des hyperliens.

Analyse des données d'usage - 1

- I. Collection de données & pré-traitement
  - transformer les *clicks* après nettoyage en *sessions* utilisateurs.
  - intégration d'autres type de connaissances : ontologies, catalogues de produits, etc.

Analyse des données d'usage - 1

- I. Collection de données & pré-traitement
  - transformer les *clicks* après nettoyage en *sessions* utilisateurs.
  - intégration d'autres type de connaissances : ontologies, catalogues de produits, etc.
- II. Découverte des schemas résultants
  - Découverte des profils utilisateurs.
  - Statistique sur les ressources, *les sessions* et *les utilisateurs*.

Analyse des données d'usage - 1

- I. Collection de données & pré-traitement
  - transformer les *clicks* après nettoyage en *sessions* utilisateurs.
  - intégration d'autres type de connaissances : ontologies, catalogues de produits, etc.
- II. Découverte des schemas résultants
  - Découverte des profils utilisateurs.
  - Statistique sur les ressources, *les sessions* et *les utilisateurs*.
- III. Analyse des schemas résultants.

Analyse des données d’usage - 2

- I. Collection de données & pré-traitement.

Analyse des données d’usage - 2

- I. Collection de données & pré-traitement.
- II. Découverte des schemas résultants.

Analyse des données d’usage - 2

- I. Collection de données & pré-traitement.
- II. Découverte des schemas résultants.
- III. Analyse des schemas résultants
  - Systèmes de recommandation.
  - Outils de visualisation.
  - Outils d’analyse de reporting.

I. Collection de données et pré-traitement

- Préparation de données
  - Traitement des données bruts (les clicks).
  - Intégration des données à partir de plusieurs ressources.
  - Transformer les données pour pouvoir leurs appliquer des algorithmes de fouille de données.

I. Collection de données et pré-traitement

- Préparation de données
  - Traitement des données bruts (les clicks).
  - Intégration des données à partir de plusieurs ressources.
  - Transformer les données pour pouvoir leurs appliquer des algorithmes de fouille de données.
- Besoins de techniques pour :
  - Nettoyage et fusions de données.
  - Identification des *vues* (des pages).
  - Identification des *utilisateurs* et des *sessions*.

Sources et types de données

- Logs relevés sur le serveur ...
  - **Click** : requête http qui génère une entrée d'une donnée logs.
  - Une donnée *log* contient les champs :
    - L'heure et la data de la requête, l'adresse IP client
    - La ressource demandée, l'agent utilisateur, la source référant, etc.
    - Éventuellement : sur la machine client : cookies.
- D'autre sources de données : base de données, méta-données, templates d'application, ontologies, etc.

Sources et types de données

- Logs relevés sur le serveur ...
  - **Click** : requête http qui génère une entrée d'une donnée logs.
  - Une donnée *log* contient les champs :
    - L'heure et la data de la requête, l'adresse IP client
    - La ressource demandée, l'agent utilisateur, la source référant, etc.
    - Éventuellement : sur la machine client : cookies.
- D'autre sources de données : base de données, méta-données, templates d'application, ontologies, etc.

Exemple d'une donnée log

- Une donnée *log* contient les champs :
  - L'heure et la data de la requête, l'adresse IP client
  - La ressource demandée, l'agent utilisateur, la source référant, etc.

Exemple d'une donnée log

- Une donnée *log* contient les champs :
  - L'heure et la data de la requête, l'adresse IP client
  - La ressource demandée, l'agent utilisateur, la source référant, etc.
- Exemple :
  - 2006-02-01 00:08:43 1.2.3.4 -
  - Get /classes/cs589/papers.html 200 9221 HTTP://1.1 maya.cs.depau1.edu
  - Mozilla/4.0+ (compatible ; +Windos+NT+5;+2.O .50727)
  - http://dataminingresource.blogspot.com/

Elements clés pour le pré-traitement

- Le nettoyage et la fusion des données.

Elements clés pour le pré-traitement

- Le nettoyage et la fusion des données.
- L'identification des *pages (vues)*.

Elements clés pour le pré-traitement

- Le nettoyage et la fusion des données.
- L'identification des *pages (vues)*.
- L'identification des *utilisateurs*.

Elements clés pour le pré-traitement

- Le nettoyage et la fusion des données.
- L'identification des *pages (vues)*.
- L'identification des *utilisateurs*.
- L'identification des *sessions*.

Elements clés pour le pré-traitement

- Le nettoyage et la fusion des données.
- L'identification des *pages (vues)*.
- L'identification des *utilisateurs*.
- L'identification des *sessions*.
- Comment compléter les chemins.

Elements clés pour le pré-traitement

- Le nettoyage et la fusion des données.
- L'identification des *pages (vues)*.
- L'identification des *utilisateurs*.
- L'identification des *sessions*.
- Comment compléter les chemins.
- L'intégration des données.

Fusion & Nettoyage

- Fusion à partir de plusieurs serveurs.

Fusion & Nettoyage

- Fusion à partir de plusieurs serveurs.
- Nettoyage selon le site.
  - Effacer les références aux certains objets imbriqués (styles, graphiques ou son).
  - Effacer le contenu de certain champ : le nombre de bytes, la version http, etc.
  - Effacer les références provenant des robots de navigation (crawlers).

Identification des pages

- Dépendance de la structure interne du site et du contenu.

Identification des pages

- Dépendance de la structure interne du site et du contenu.
- Une page (une vue) correspond à une collection d’objets ou de ressources suite à un événement effectué par l’utilisateur (click, etc.).

Identification des pages

- Dépendance de la structure interne du site et du contenu.
- Une page (une vue) correspond à une collection d'objets ou de ressources suite à un événement effectué par l'utilisateur (click, etc.).
- Site statique : une page correspond un fichier HTML.

Identification des pages

- Dépendance de la structure interne du site et du contenu.
- Une page (une vue) correspond à une collection d'objets ou de ressources suite à un événement effectué par l'utilisateur (click, etc.).
- Site statique : une page correspond un fichier HTML.
- Site dynamique : une page est une combinaison de templates statiques et de contenu généré par des applications liées au serveur.

Identification des pages

- Dépendance de la structure interne du site et du contenu.
- Une page (une vue) correspond à une collection d'objets ou de ressources suite à un événement effectué par l'utilisateur (click, etc.).
- Site statique : une page correspond un fichier HTML.
- Site dynamique : une page est une combinaison de templates statiques et de contenu généré par des applications liées au serveur.
- Pour identifier les pages : bien spécifier l'ensemble des événements liés aux actions utilisateurs.

Identification des pages

- Dépendance de la structure interne du site et du contenu.
- Une page (une vue) correspond à une collection d'objets ou de ressources suite à un événement effectué par l'utilisateur (click, etc.).
- Site statique : une page correspond un fichier HTML.
- Site dynamique : une page est une combinaison de templates statiques et de contenu généré par des applications liées au serveur.
- Pour identifier les pages : bien spécifier l'ensemble des événements liés aux actions utilisateurs.

Identification des utilisateurs

- Mécanisme d’authentification

Identification des utilisateurs

- Mécanisme d’authentification
- Cookies côté client.

Identification des utilisateurs

- Mécanisme d’authentification
- Cookies côté client.
- Adresse IP + Agent.

Identification d'utilisateurs - 1

Time	IP	URL	Ref	Agent
0:01	1.2.3.4	A	-	IE5;Win2K
0:09	1.2.3.4	B	A	IE5;Win2K
0:10	2.3.4.5	C	-	IE6;WinXP;SP1
0:12	2.3.4.5	B	C	IE6;WinXP;SP1
0:15	2.3.4.5	E	C	IE6;WinXP;SP1
0:19	1.2.3.4	C	A	IE5;Win2K
0:22	2.3.4.5	D	B	IE6;WinXP;SP1
0:22	1.2.3.4	A	-	IE6;WinXP;SP2
0:25	1.2.3.4	E	C	IE5;Win2K
0:25	1.2.3.4	C	A	IE6;WinXP;SP2
0:33	1.2.3.4	B	C	IE6;WinXP;SP2
0:58	1.2.3.4	D	B	IE6;WinXP;SP2

Identification d'utilisateurs - 2

- user1

Time	IP	URL	Ref
0:01	1.2.3.4	A	-
0:09	1.2.3.4	B	A
0:19	1.2.3.4	C	A
0:25	1.2.3.4	E	C

- user2

Time	IP	URL	Ref
0:10	2.3.4.5	C	-
0:12	2.3.4.5	B	C
0:15	2.3.4.5	E	C
0:22	2.3.4.5	D	B

Identification d'utilisateurs - 3

- user3

Time	IP	URL	Ref
0:22	1.2.3.4	A	-
0:25	1.2.3.4	C	A
0:33	1.2.3.4	B	C
0:58	1.2.3.4	D	B

Identification de sessions

- Une session : une visite au site effectué par un utilisateur.

Identification de sessions

- Une session : une visite au site effectué par un utilisateur.
- Une séquence de navigations effectuée par un utilisateur.

Identification de sessions

- Une session : une visite au site effectué par un utilisateur.
- Une séquence de navigations effectuée par un utilisateur.
- Une heuristique fondée sur le temps ou la structure.

Identification de sessions

- Une session : une visite au site effectué par un utilisateur.
- Une séquence de navigations effectuée par un utilisateur.
- Une heuristique fondée sur le temps ou la structure.
- Le temps permet de découvrir deux sessions consécutives (par utilisateur).

Session & heuristique

- Exemples d’heuristique :
  - h1 : La durée globale d’une session ne doit pas dépasser un certain seuil.
  - h2 : La durée d’une page ne doit pas dépasser un certain seuil.
  - href : Q est ajoutée à une session si elle y est référée.

Identification de sessions - 1

Time	IP	URL	Ref
0:01	1.2.3.4	A	-
0:09	1.2.3.4	B	A
0:19	1.2.3.4	C	A
0:25	1.2.3.4	E	C
1:15	1.2.3.4	A	-
1:26	1.2.3.4	F	C
1:30	1.2.3.4	B	A
1:36	1.2.3.4	D	B

Identification de sessions par h1

Time	IP	URL	Ref
0:01	1.2.3.4	A	-
0:09	1.2.3.4	B	A
0:19	1.2.3.4	C	A
0:25	1.2.3.4	E	C
-	-	-	-
1:15	1.2.3.4	A	-
1:26	1.2.3.4	F	C
1:30	1.2.3.4	B	A
1:36	1.2.3.4	D	B

Identification de sessions par href

Time	IP	URL	Ref
0:01	1.2.3.4	A	-
0:09	1.2.3.4	B	A
0:19	1.2.3.4	C	A
0:25	1.2.3.4	E	C
1:26	1.2.3.4	F	C
-	-	-	-
1:15	1.2.3.4	A	-
1:30	1.2.3.4	B	A
1:36	1.2.3.4	D	B

Comment compléter le chemin - 1

- Chemin de navigation incomplet (les caches utilisateurs).

Comment compléter le chemin - 1

- Chemin de navigation incomplet (les caches utilisateurs).
- Utiliser des heuristiques à partir de la structure du site, exemple :
  - Soit le log :

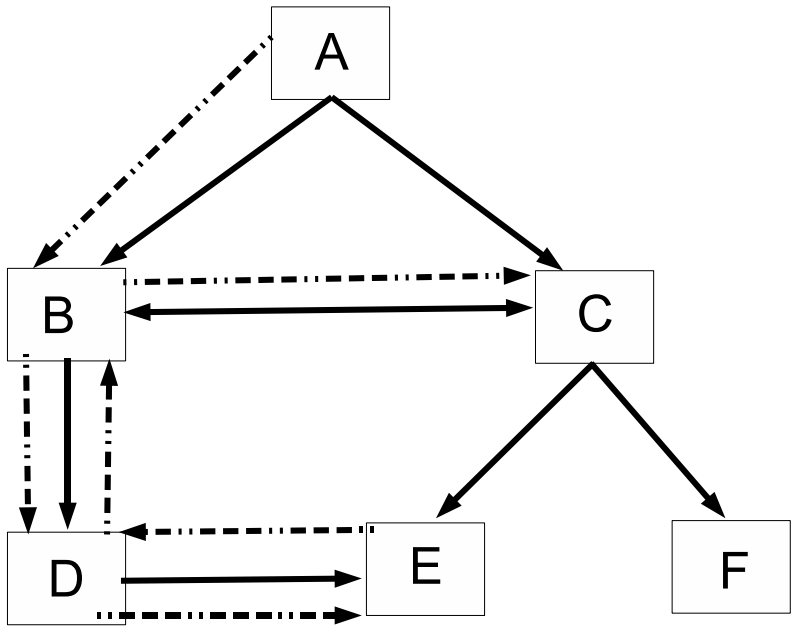
URL	Référant
A	
B	A
D	B
E	D
C	B

Comment compléter le chemin - 2

- Chemin détecté : A,B,D,E,C

Comment compléter le chemin - 2

- Chemin détecté : A,B,D,E,C



Intégration de données

- Intégration des données d’usage avec les autres données.

Intégration de données

- Intégration des données d’usage avec les autres données.
- Exemple E-Commerce : afin de déterminer le modèle utilisateur, etc.
  - Données liées au produits (extraites de la base de données).
  - Les achats, les intérêts des clients (selon les clicks).

Analyse des données d’usage - 1

- I. Collection de données & pré-traitement
  - transformer les clicks après nettoyage en sessions utilisateurs.
  - intégration d’autres type de connaissances : ontologies, catalogues de produits, etc.

Analyse des données d'usage - 1

- I. Collection de données & pré-traitement
  - transformer les clicks après nettoyage en sessions utilisateurs.
  - intégration d'autres type de connaissances : ontologies, catalogues de produits, etc.
- II. Découverte des schemas résultants
  - Découverte des profils utilisateurs.
  - Statistique sur les ressources, *les sessions* et *les utilisateurs*.

Analyse des données d'usage - 1

- I. Collection de données & pré-traitement
  - transformer les clicks après nettoyage en sessions utilisateurs.
  - intégration d'autres type de connaissances : ontologies, catalogues de produits, etc.
- II. Découverte des schemas résultants
  - Découverte des profils utilisateurs.
  - Statistique sur les ressources, *les sessions* et *les utilisateurs*.
- III. Analyse des schemas résultants.

Modélisation des données d'usage

- Un ensemble de n pages : $P = \{p_1, p_2, .., p_n\}$
- Un ensemble de m transactions par utilisateur $T = \{t_1, t_2, .., t_n\}$ Une transaction est définie par :

$$t =< (p_1^t, w(p_1^t)), (p_2^t, w(p_2^t)), .., , (p_l^t, w(p_l^t)) >$$

où $w(p_i^t)$ étant un poids associé à la page en question

1. le poids peut être binaire,
2. la durée de la vue,
3. le poids : ciblé pour une application (exemple filtrage collaborative).

La matrice utilisateur-page

- Une ligne par utilisateur

	A	B	C	D	E	F
user1	15	5	0	0	0	185
user2	0	O	32	4	0	185
user3	12	O	0	56	236	0
user4	9	47	0	0	0	134

Exemple : Découverte de profil utilisateur - 1

- Matrice utilisateurs-pages.

Exemple : Découverte de profil utilisateur - 1

- Matrice utilisateurs-pages.
- Matrice termes-pages (données intégrées).

	A	B	C	D	E	F
terme1	1	0	0	0	0	1
terme2	0	0	1	1	0	0
terme3	0	1	0	1	0	1
terme4	1	0	0	1	1	0

Exemple : Découverte de profil utilisateur - 1

- Matrice utilisateurs-pages.
- Matrice termes-pages (données intégrées).
  -

	A	B	C	D	E	F
terme1	1	0	0	0	0	1
terme2	0	0	1	1	0	0
terme3	0	1	0	1	0	1
terme4	1	0	0	1	1	0

- Découverte de la matrice profil (en fonction du contenu).

Exemple : Découverte de profil utilisateur - 2

- Matrice utilisateurs-pages.

Exemple : Découverte de profil utilisateur - 2

- Matrice utilisateurs-pages.
- Matrice termes-pages (données intégrées).

Exemple : Découverte de profil utilisateur - 2

- Matrice utilisateurs-pages.
- Matrice termes-pages (données intégrées).
- Découverte de la matrice profil (en fonction du contenu)

	terme1	terme2	terme3	terme4
user1	2	0	2	1
user2	1	1	1	0
user3	1	1	1	3
user4	2	1	2	1

III. Analyse des schémas résultants

- Analyse des sessions utilisateurs.
  - Elements : temps d’affichage, les pages les plus visitées.
  - Comportement utilisateur (E-Commerce).

III. Analyse des schémas résultants

- Analyse des sessions utilisateurs.
  - Elements : temps d'affichage, les pages les plus visitées.
  - Comportement utilisateur (E-Commerce).
- Analyse des *clusters* & Segmentation des visiteurs
  - Cluster : groupe d'individus similaires.
  - Similarité sur le comportement.
  - Application : E-Commerce, Personnalisation & Communautés.
  - Algorithme : les centres mobiles (K-Means).

Exemple de Clusters

	A	B	C	D	E	F
user1	0	0	1	1	0	0
user4	0	0	1	1	0	0
user7	0	0	1	1	0	0
user0	1	1	0	0	0	1
user3	1	1	0	0	0	1
user6	1	1	0	0	0	1
user9	0	1	1	0	0	1
user2	1	0	0	1	1	0
user5	1	0	0	1	1	0
user8	1	0	1	1	1	0

Profile d'un cluster

	A	B	C	D	E	F
user0	1	1	0	0	0	1
user3	1	1	0	0	0	1
user6	1	1	0	0	0	1
user9	0	1	1	0	0	1

Poids	Page
1.00	B
1.00	F
0.75	A
0.25	C

Analyse des associations - 1

- Trouver les règles d'associations
  - Exemple : Utilisation de l'algorithme Apriori[Agrawal 93] pour trouver les associations entre les pages ou les thèmes (visité(e)s).
  - Exemple d'une règle *specialoffers/*, */products/software/* $\longrightarrow$ *shopping-cart*

Analyse des associations - 1

- Trouver les règles d'associations
  - Exemple : Utilisation de l'algorithme Apriori[Agrawal 93] pour trouver les associations entre les pages ou les thèmes (visité(e)s).
  - Exemple d'une règle *specialoffers/, /products/software/* $\rightarrow$ *shopping-cart*
- Optimiser la structure du site.
  - Exemple : Si $A \rightarrow B$ est une règle découverte avec une confiance élevée.
  - Ajout d'un lien de la page A vers la page B.

Les règles d’association - Terminologie

- Domaine décrit par une liste d’atomes appelée *items*.

Les règles d’association - Terminologie

- Domaine décrit par une liste d’atomes appelée *items*.
- Application : panier de ménagère dans un supermarché : [vin, fromage, chocolat]

Les règles d'association - Terminologie

- Domaine décrit par une liste d'atomes appelée *items*.
- Application : panier de ménagère dans un supermarché : [vin, fromage, chocolat]
- Un ensemble d'*items* est une suite d'*items* exprimée dans un ordre donné.

Les règles d'association - Terminologie

- Domaine décrit par une liste d'atomes appelée *items*.
- Application : panier de ménagère dans un supermarché : [vin, fromage, chocolat]
- Un ensemble d'*items* est une suite d'*items* exprimée dans un ordre donné.
- Une transaction est un ensemble d'items, exemples :

Les règles d'association - Terminologie

- Domaine décrit par une liste d'atomes appelée *items*.
- Application : panier de ménagère dans un supermarché : [vin, fromage, chocolat]
- Un ensemble d'*items* est une suite d'*items* exprimée dans un ordre donné.
- Une transaction est un ensemble d'items, exemples :

  T1 [vin, fromage, viande]

Les règles d'association - Terminologie

- Domaine décrit par une liste d'atomes appelée *items*.
- Application : panier de ménagère dans un supermarché : [vin, fromage, chocolat]
- Un ensemble d'*items* est une suite d'*items* exprimée dans un ordre donné.
- Une transaction est un ensemble d'items, exemples :

  T1 [vin, fromage, viande]

  T2 [vin, fromage, chocolat]

Les règles d'association - Terminologie

- Domaine décrit par une liste d'atomes appelée *items*.
- Application : panier de ménagère dans un supermarché : [vin, fromage, chocolat]
- Un ensemble d'*items* est une suite d'*items* exprimée dans un ordre donné.
- Une transaction est un ensemble d'items, exemples :

  T1 [vin, fromage, viande]

  T2 [vin, fromage, chocolat]
- Un ensemble D de transactions correspond à un ensemble d'apprentissage.

Les règles d’association - Objectif

- Objectif : chercher les associations à partir de D

Les règles d'association - Objectif

- Objectif : chercher les associations à partir de D

  T1 $vin \rightarrow fromage$

Les règles d'association - Objectif

- Objectif : chercher les associations à partir de D

T1 $vin \rightarrow fromage$

T2 $vinfromage \rightarrow jambon$

Les règles d’association - Définition

- Ensemble d’items fréquents : motif fréquent dans la base de transactions.

Les règles d’association - Définition

- Ensemble d’items fréquents : motif fréquent dans la base de transactions.

  minSupp un paramètre

Les règles d’association - Définition

- Ensemble d’items fréquents : motif fréquent dans la base de transactions.

  minSupp un paramètre

- Trouver tous les ensembles d’items fréquents de longueurs différentes, exemple :

Les règles d’association - Définition

- Ensemble d’items fréquents : motif fréquent dans la base de transactions.

  minSupp un paramètre

- Trouver tous les ensembles d’items fréquents de longueurs différentes, exemple :

  1. Si ABCD est un ensemble d’items fréquent

Les règles d'association - Définition

- Ensemble d'items fréquents : motif fréquent dans la base de transactions.

  minSupp un paramètre

- Trouver tous les ensembles d'items fréquents de longueurs différentes, exemple :
  1. Si ABCD est un ensemble d'items fréquent
  2. Construire la règle $AB \Rightarrow CD$ ssi $support(ABCD)/support(AB) \geq minConf$

Les règles d'association - Définition

- Ensemble d'items fréquents : motif fréquent dans la base de transactions.

  minSupp un paramètre

- Trouver tous les ensembles d'items fréquents de longueurs différentes, exemple :
  1. Si ABCD est un ensemble d'items fréquent
  2. Construire la règle $AB \Rightarrow CD$ ssi $support(ABCD)/support(AB) \geq minConf$
  3. minConf est un paramètre

Les règles d’association - Algorithmes

- Terminologie

Les règles d'association - Algorithmes

- Terminologie

  L_k est l'ensemble constitué des sous-ensembles d'items fréquents de longueur k.

Les règles d'association - Algorithmes

- Terminologie

  L_k est l'ensemble constitué des sous-ensembles d'items fréquents de longueur k.

  C_k est un ensemble constitué des sous-ensembles d'items candidats de longueur k, notons bien que $L_k \subset C_k$

Les règles d'association - Algorithmes

- Terminologie

  L_k est l'ensemble constitué des sous-ensembles d'items fréquents de longueur k.

  C_k est un ensemble constitué des sous-ensembles d'items candidats de longueur k, notons bien que $L_k \subset C_k$

- **Propriété** Soit X_k un sous-ensemble d'items fréquent, tous les sous-ensembles d'items contenus dans X_k et qui soient de longueurs inférieurs à k sont fréquents.

Les règles d'association - Algorithmes

- Terminologie

  L_k est l'ensemble constitué des sous-ensembles d'items fréquents de longueur k.

  C_k est un ensemble constitué des sous-ensembles d'items candidats de longueur k, notons bien que $L_k \subset C_k$

- **Propriété** Soit X_k un sous-ensemble d'items fréquent, tous les sous-ensembles d'items contenus dans X_k et qui soient de longueurs inférieurs à k sont fréquents.

  1. Si ABCD est un ensemble d'items fréquent

Les règles d'association - Algorithmes

- Terminologie

  L_k est l'ensemble constitué des sous-ensembles d'items fréquents de longueur k.

  C_k est un ensemble constitué des sous-ensembles d'items candidats de longueur k, notons bien que $L_k \subset C_k$

- **Propriété** Soit X_k un sous-ensemble d'items fréquent, tous les sous-ensembles d'items contenus dans X_k et qui soient de longueurs inférieurs à k sont fréquents.

  1. Si ABCD est un ensemble d'items fréquent
  2. ABC,ABD, BCD, AB,AC,BC,BD,CD,A,B,C,D les sont aussi.

Génération de règles - 1

- Approche descendante de génération fondée sur deux propriétés :

Génération de règles - 1

- Approche descendante de génération fondée sur deux propriétés :
  1. **Redondance simple** : Nous testons les règles ayant le nombre de conditions minimal pour un sous-ensemble fréquent, exemple :

Génération de règles - 1

- Approche descendante de génération fondée sur deux propriétés :
  1. **Redondance simple** : Nous testons les règles ayant le nombre de conditions minimal pour un sous-ensemble fréquent, exemple :
     R1 $A \Rightarrow B, C$

Génération de règles - 1

- Approche descendante de génération fondée sur deux propriétés :
  1. **Redondance simple** : Nous testons les règles ayant le nombre de conditions minimal pour un sous-ensemble fréquent, exemple :

     R1 $A \Rightarrow B, C$

     R2 $A, B \Rightarrow C$

Génération de règles - 1

- Approche descendante de génération fondée sur deux propriétés :
  1. **Redondance simple** : Nous testons les règles ayant le nombre de conditions minimal pour un sous-ensemble fréquent, exemple :
     R1 $A \Rightarrow B, C$
     R2 $A, B \Rightarrow C$
- confiance(R1)=support(ABC)/support(A),

Génération de règles - 1

- Approche descendante de génération fondée sur deux propriétés :
  1. **Redondance simple** : Nous testons les règles ayant le nombre de conditions minimal pour un sous-ensemble fréquent, exemple :
     R1 $A \Rightarrow B, C$
     R2 $A, B \Rightarrow C$
- confiance(R1)=support(ABC)/support(A),
- confiance(R2)=support(ABC)/support(AB),

Génération de règles - 1

- Approche descendante de génération fondée sur deux propriétés :
  1. **Redondance simple** : Nous testons les règles ayant le nombre de conditions minimal pour un sous-ensemble fréquent, exemple :
     R1 $A \Rightarrow B, C$
     R2 $A, B \Rightarrow C$
- confiance(R1)=support(ABC)/support(A),
- confiance(R2)=support(ABC)/support(AB),
- confiance(R2)>confinace(R1).

Génération de règles - 2

- Approche descendante de génération fondée sur deux propriétés :

Génération de règles - 2

- Approche descendante de génération fondée sur deux propriétés :
  1. **Redondance stricte** : Nous commençons par la recherche par les ensembles fréquents les plus grands, exemple :

Génération de règles - 2

- Approche descendante de génération fondée sur deux propriétés :
  1. **Redondance stricte** : Nous commençons par la recherche par les ensembles fréquents les plus grands, exemple :
     R1 $A \Rightarrow B, C, D$

Génération de règles - 2

- Approche descendante de génération fondée sur deux propriétés :
  1. **Redondance stricte** : Nous commençons par la recherche par les ensembles fréquents les plus grands, exemple :

     R1 $A \Rightarrow B, C, D$

     R2 $A \Rightarrow B, C$

Génération de règles - 2

- Approche descendante de génération fondée sur deux propriétés :
  1. **Redondance stricte** : Nous commençons par la recherche par les ensembles fréquents les plus grands, exemple :
     R1 $A \Rightarrow B, C, D$
     R2 $A \Rightarrow B, C$
- confiance(R1)=support(ABCD)/support(A),

Génération de règles - 2

- Approche descendante de génération fondée sur deux propriétés :
  1. **Redondance stricte** : Nous commençons par la recherche par les ensembles fréquents les plus grands, exemple :
     R1 $A \Rightarrow B, C, D$
     R2 $A \Rightarrow B, C$
- confiance(R1)=support(ABCD)/support(A),
- confiance(R2)=support(ABC)/support(A),

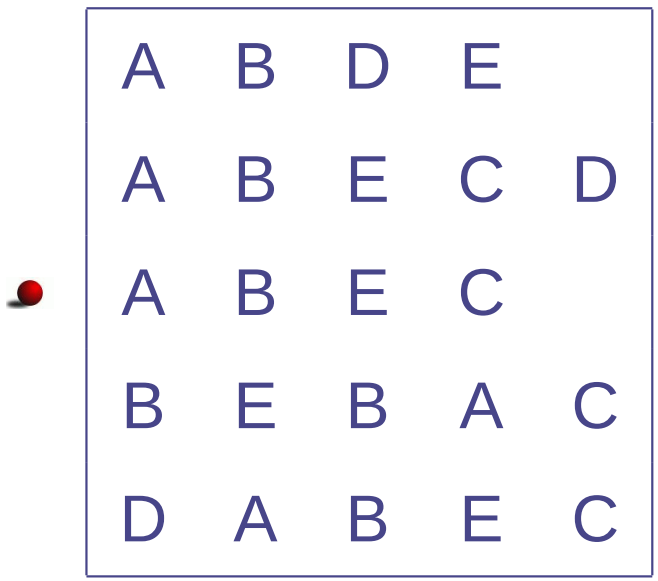
Génération de règles - 2

- Approche descendante de génération fondée sur deux propriétés :
  1. **Redondance stricte** : Nous commençons par la recherche par les ensembles fréquents les plus grands, exemple :
     R1 $A \Rightarrow B, C, D$
     R2 $A \Rightarrow B, C$
- confiance(R1)=support(ABCD)/support(A),
- confiance(R2)=support(ABC)/support(A),
- confiance(R2)>confinace(R1)

Analyse des associations - 2

- Système de Recommandation
  - Exemple : A partir de l'algorithme Apriori : trouver les sous ensembles fréquents.
  - Dans le contexte d'une navigation actuelle, essayer à chaque étapes de prédire (recommander à l'utilisateur) la page suivante de la navigation.

Système de Recommandation - Exemple



Système de Recommandation - Exemple

A	B	D	E	
A	B	E	C	D
A	B	E	C	
B	E	B	A	C
D	A	B	E	C

- Les itemsets fréquents : ABCE(4), ABC(4), ABE(5), ACE(4), BCE(4), AB(5),AC(4), AE(5), BC(4), BE(5),CE(4).

Système de Recommandation - Utilisation

- Les sous ensembles fréquents : ABCE(4), ABC(4), ABE(5), ACE(4), BCE(4), AB(5),AC(4), AE(5), BC(4), BE(5),CE(4).

Système de Recommandation - Utilisation

- Les sous ensembles fréquents : ABCE(4), ABC(4), ABE(5), ACE(4), BCE(4), AB(5),AC(4), AE(5), BC(4), BE(5),CE(4).
- Utilisateur a effectué une partie de chemin, lui recommander la page suivante :
  - <B,E> les deux recommandations possibles sont A (1), C(4/5) selon les règles :
  -

$$B, E \rightarrow A$$

$$B, E \rightarrow C$$

Fouille des Données de la Toile

- **Contenu** *(Web Content Mining)* Analyse des contenus des pages web :
  - Classer et/ou segmenter les pages selon le thème.
  - Chercher des descriptions de produits, etc.

Fouille des Données de la Toile

- **Contenu** *(Web Content Mining)* Analyse des contenus des pages web :
  - Classer et/ou segmenter les pages selon le thème.
  - Chercher des descriptions de produits, etc.
- **Comportement** *(Web Usage Mining)* Analyse les traces de navigations des internautes (logs)
  - Algorithmes d'analyse et de traitement de séquences
  - Besoin d'une phase de pré-traitement

Fouille des Données de la Toile

- **Contenu** *(Web Content Mining)* Analyse des contenus des pages web :
  - Classer et/ou segmenter les pages selon le thème.
  - Chercher des descriptions de produits, etc.
- **Comportement** *(Web Usage Mining)* Analyse les traces de navigations des internautes (logs)
  - Algorithmes d'analyse et de traitement de séquences
  - Besoin d'une phase de pré-traitement
- **Structure** *(Web Structure Mining)* Découverte des connaissances à partir des hyperliens.

Les moteurs de recherche

- **Requête utilisateur posée :** algorithmes de recherches de contenus similaires.

Les moteurs de recherche

- **Requête utilisateur posée :** algorithmes de recherches de contenus similaires.
- **Problèmes à partir de 1996 :** besoins de nouvelles techniques de tri ("ranking")
  - Nombre de pages Web a très largement augmenté, *exemple :* pour la requête **classification technique** on estime qu'il exits 10 millions de pages résultats.
  - Les méthodes de recherche par le contenus sont facilement *attaqués* par les concepteurs des pages.

Moteurs de recherche & liens

- Alors : **prenons en compte les liens** dans les techniques de "ranking"
  - Liens sortants : indiquent une croyance implicite à l’autorité des pages pointues;
  - Liens entrants.

Moteurs de recherche & liens

- Alors : **prenons en compte les liens** dans les techniques de "ranking"
  - Liens sortants : indiquent une croyance implicite à l'autorité des pages pointues;
  - Liens entrants.
- Algorithmes 1997-1998 : PageRank & HITS

Analyse de réseaux sociaux

- Etude des entités sociales, les interactions et les relations ;

Analyse de réseaux sociaux

- Etude des entités sociales, les interactions et les relations ;
- Modélisation par un graphe :
  - un nœud représente un acteur,
  - un lien représente une relation,
  - détecter des communautés : *sous graphes*.

Analyse de réseaux sociaux

- Etude des entités sociales, les interactions et les relations ;
- Modélisation par un graphe :
  - un nœud représente un acteur,
  - un lien représente une relation,
  - détecter des communautés : *sous graphes*.
- *La toile est un réseau social virtuel ..*

Analyse de réseaux sociaux

- Etude des entités sociales, les interactions et les relations ;
- Modélisation par un graphe :
  - un nœud représente un acteur,
  - un lien représente une relation,
  - détecter des communautés : *sous graphes*.
- *La toile est un réseau social virtuel ..*
- Deux mesures pour les réseaux sociaux
  - Notion de centralité,
  - Le prestige.

Notion de centralité

- Une personne ayant beaucoup de relations est considéré plus important,

Notion de centralité

- Une personne ayant beaucoup de relations est considéré plus important,
- un *acteur central* correspond à un nœud ayant de nombreux liens,

Notion de centralité

- Une personne ayant beaucoup de relations est considéré plus important,
- un *acteur central* correspond à un nœud ayant de nombreux liens,
- les mesures de centralités :
  - le degré de centralité.
  - centralité de proximité
  - centralité d'intermédiarité.

Degré de centralité

- *Acteur central* est l'acteur le plus actif de point de vue communication,

Degré de centralité

- *Acteur central* est l'acteur le plus actif de point de vue communication,
- *Graphe non dirigé*, contenant n nœuds, $d(i)$ étant le degré du nœud i :

$$C_D(i) = \frac{d(i)}{n-1}$$

Degré de centralité

- *Acteur central* est l'acteur le plus actif de point de vue communication,
- *Graphe non dirigé*, contenant n nœuds, $d(i)$ étant le degré du nœud i :

$$C_D(i) = \frac{d(i)}{n-1}$$

- *Graphe dirigé*, contenant n nœuds, $d_o(i)$ étant le degré (liens sortants) du nœud i :

$$C'_D(i) = \frac{d_o(i)}{n-1}$$

Centralité de proximité

- La notion de centralité utilise la distance.

Centralité de proximité

- La notion de centralité utilise la distance.
- *Acteur central* est l'acteur qui communique *facilement* avec les autres,

Centralité de proximité

- La notion de centralité utilise la distance.
- *Acteur central* est l'acteur qui communique *facilement* avec les autres,
- $d(i, j)$ est la distance entre deux acteurs mesurée en nombre minimal de liens.

Centralité de proximité

- La notion de centralité utilise la distance.
- *Acteur central* est l'acteur qui communique *facilement* avec les autres,
- $d(i, j)$ est la distance entre deux acteurs mesurée en nombre minimal de liens.
- *Graphe non dirigé*, contenant n nœuds :

$$C_C(i) = \frac{n-1}{\sum_{j=1}^{n} d(i,j)}$$

Centralité de proximité

- La notion de centralité utilise la distance.
- *Acteur central* est l'acteur qui communique *facilement* avec les autres,
- $d(i, j)$ est la distance entre deux acteurs mesurée en nombre minimal de liens.
- *Graphe non dirigé*, contenant n nœuds :

$$C_C(i) = \frac{n-1}{\sum_{j=1}^{n} d(i,j)}$$

- *Graphe dirigé* La distance doit prendre en compte les sens des liens.

Centralité d'intermédiarité-1

- Deux nœuds *non adjacents* k & j qui se communiquent et si le nœud i se trouve sur le chemin de communication : *i est un acteur important*.

Centralité d'intermédiarité-1

- Deux nœuds *non adjacents* k & j qui se communiquent et si le nœud i se trouve sur le chemin de communication : *i est un acteur important*.
- *Graphe non dirigé* :
  - p_{jk} le nombre des chemins les plus cours entre j et k,
  - $p_{jk}(i)$ le nombre des chemins les plus cours entre j et k passant par i,
  - $C_B(i) = \sum_{j<k} \frac{p_{jk}(i)}{p_{jk}}$
  - $0 < C_B(i)$ & $C_B(i) < \frac{(n-1)(n-2)}{2}$

Centralité d'intermédiarité-2

- *Graphe dirigé* :

$$C'_B(i) = \frac{\sum_{j<k} \frac{p_{jk}(i)}{p_{jk}}}{(n-1)(n-2)}$$

Le prestige

- Le prestige d’un acteur est mesuré par les liens entrants.

Le prestige

- Le prestige d’un acteur est mesuré par les liens entrants.
- Les mesures :
  - le degré de prestige,
  - le prestige de proximité,
  - le prestige de tri (rank prestige).

Le prestige - mesures

- Le degré :

$$P_D(i) = \frac{d_I(i)}{n-1}$$

Le prestige - mesures

- Le degré :

$$P_D(i) = \frac{d_I(i)}{n-1}$$

- La proximité :
  - I_i est l'ensemble d'acteurs qui atteignent l'acteur i: il existe un chemin de j vers i.
  - $d(j,i)$ le chemin le plus court de j vers i.
  - La distance moyenne est $\frac{\sum_{j \in I_i} d(j,i)}{|I_i|}$
  - Le prestige de proximité : $P_p(i) = \frac{\frac{|I_i|}{n-1}}{\frac{\sum_{j \in I_i} d(j,i)}{|I_i|}}$

"Rank Prestige"

- Considérons la réputation des personnes choisissant l'acteur i

"Rank Prestige"

- Considérons la réputation des personnes choisissant l'acteur i
- $P_R(i) = A_{1i}P_R(1) + A_{2i}P_R(2) + .. + A_{ni}P_R(n)$
  - A_{ij} vaut 1 si i pointe vers j,sinon 0.
  - On pose P le vecteur contenant les valeurs de tri :

$$P = (P_R(1), P_R(2), .., P_R(n))^T$$

  - A est la matrice exprimant la propriété adjacente

$$P = A^T P$$

    .
  - Recherche de valeurs propres de A^T.

La co-citation

- Mesure de similarités entre deux documents :

La co-citation

- Mesure de similarités entre deux documents :
- Si les papiers i et j sont cités par k alors ils sont liés.

La co-citation

- Mesure de similarités entre deux documents :
- Si les papiers i et j sont cités par k alors ils sont liés.
- Si les papiers i et j sont cités par plusieurs papiers alors ils sont *similaires*.
  - Soit L la matrice de citation, L_{ij} vaut 1 si i cite j, sinon 0.
  - La co-citation est une mesure qui est définie par :

$$C_{ij} = \sum_{k=1}^{n} L_{ki} L_{kj}$$

  - La co-citation est symétrique $C_{ij} = C_{ji}$

Le couplage biblio

- Mesure de similarités entre deux documents

Le couplage biblio

- Mesure de similarités entre deux documents
- Si les papiers i et j citent le même papier k alors ils sont liés.

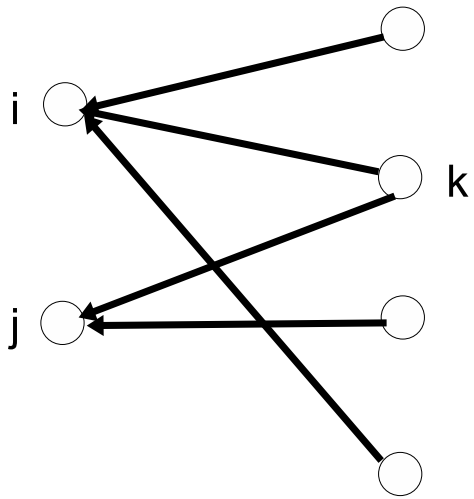
Le couplage biblio

- Mesure de similarités entre deux documents
- Si les papiers i et j citent le même papier k alors ils sont liés.
- Si les papiers i et j citent plusieurs papiers alors ils sont similaires.
  - Soit L la matrice de citation, L_{ij} vaut 1 si i cite j, sinon 0.
  - Le couplage biblio est une mesure qui est définie par :

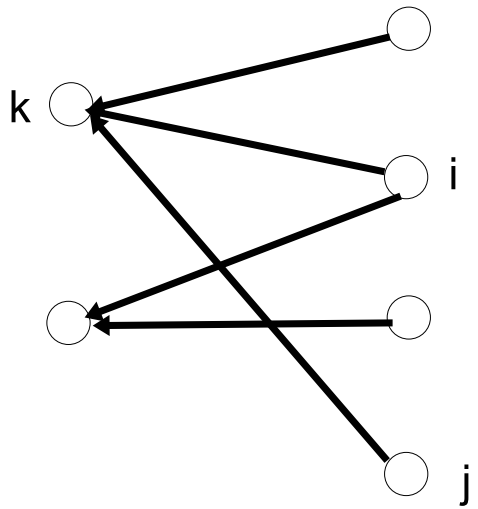
$$B_{ij} = \sum_{k=1}^{n} L_{ik} L_{jk}$$

  - Le couplage biblio est symétrique $B_{ij} = B_{ji}$

Co-citation : exemple



Couplage Bibligraphique : exemple



L’algorithme "PageRank"

- 1998, by Sergey Brin & Larry Page.

L’algorithme "PageRank"

- 1998, by Sergey Brin & Larry Page.
- Tri statique des pages Web.

L’algorithme "PageRank"

- 1998, by Sergey Brin & Larry Page.
- Tri statique des pages Web.
- Le web est traité comme étant un graphe dirigé(V,E).

L’algorithme "PageRank"

- 1998, by Sergey Brin & Larry Page.
- Tri statique des pages Web.
- Le web est traité comme étant un graphe dirigé(V,E).
- la mesure associée à une page i est :

$$P(i) = \sum_{(j,i)\in E} \frac{P(j)}{O_j}$$

où O_j étant le nombre de liens sortant de la page j

L'algorithme "PageRank"

- 1998, by Sergey Brin & Larry Page.
- Tri statique des pages Web.
- Le web est traité comme étant un graphe dirigé(V,E).
- la mesure associée à une page i est :

$$P(i) = \sum_{(j,i)\in E} \frac{P(j)}{O_j}$$

où O_j étant le nombre de liens sortant de la page j

- On pose *P* le vecteur contenant les valeurs de tri :

$$P = (P(1), P(2), .., P(n))^T$$

L'algorithme "PageRank"

- A est la matrice exprimant la propriété adjacente :

$$A_{ij} = \begin{cases} \frac{1}{O_j} & \text{if } (i,j) \in E \\ 0 & \text{otherwise} \end{cases} \quad (1)$$

$$P = A^T P$$

L'algorithme "PageRank"

- A est la matrice exprimant la propriété adjacente :

$$A_{ij} = \begin{cases} \frac{1}{O_j} & \text{if } (i,j) \in E \\ 0 & \text{otherwise} \end{cases} \tag{2}$$

$$P = A^T P$$

- Recherche de valeurs propres de A^T

Le point de vue Markovien - 1

- Une page web est un nœud ou un état,

Le point de vue Markovien - 1

- Une page web est un nœud ou un état,
- Un hyper-lien est une transition,

Le point de vue Markovien - 1

- Une page web est un nœud ou un état,
- Un hyper-lien est une transition,
- A_{ij} est la probabilité de passer de la page i à la page j.

Le point de vue Markovien - 1

- Une page web est un nœud ou un état,
- Un hyper-lien est une transition,
- A_{ij} est la probabilité de passer de la page i à la page j.
- Étant donnée une distribution initiale :

$$p_0 = (p_0(1), p_0(2), .., p_0(n))$$

avec A la matrice de transition (nxn), nous avons :

$$\sum_{i=1}^{n} p_0(i) = 1 \text{ \& } \sum_{j=1}^{n} A_{ij} = 1$$

Le point de vue Markovien - 2

$$p_1(j) = \sum_{i=1}^{n} A_{ij}(1) p_0(i)$$

Le point de vue Markovien - 2

- $$p_1(j) = \sum_{i=1}^{n} A_{ij}(1) p_0(i)$$
- Et plus généralement : $P_k = A^T P_{k-1}$

Le point de vue Markovien - 2

- $$p_1(j) = \sum_{i=1}^{n} A_{ij}(1) p_0(i)$$
- Et plus généralement : $P_k = A^T P_{k-1}$
- **théorème de chaînes de Markov** Si A est irréductible et non périodique alors π existe et unique :

$$\lim_{k \to \infty} P_k = \pi$$

Le point de vue Markovien - 2

- $$p_1(j) = \sum_{i=1}^{n} A_{ij}(1) p_0(i)$$
- Et plus généralement : $P_k = A^T P_{k-1}$
- **théorème de chaînes de Markov** Si A est irréductible et non périodique alors π existe et unique :

$$\lim_{k \to \infty} P_k = \pi$$

- $$P = A^T P$$

Face à la réalité de la toile - 1

- **I. A n’est pas stochastique** : Il y a des pages qui ne contiennent pas de liens sortants.

Face à la réalité de la toile - 1

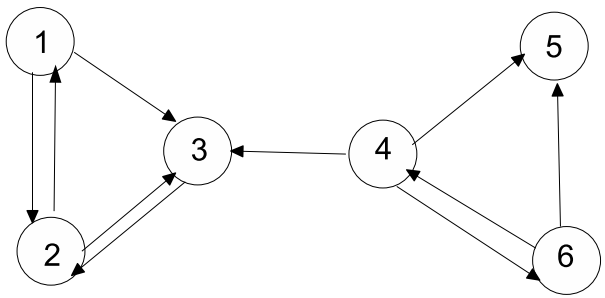
- **I. A n'est pas stochastique** : Il y a des pages qui ne contiennent pas de liens sortants.
- Soit la matrice suivante :

$$A = \begin{bmatrix} 0 & 1/2 & 1/2 & 0 & 0 & 0 \\ 1/2 & 0 & 1/2 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1/3 & 0 & 1/3 & 1/3 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1/2 & 1/2 & 0 \end{bmatrix}$$

Exemple de liens

- Soit la matrice suivante :

$$A = \begin{bmatrix} 0 & 1/2 & 1/2 & 0 & 0 & 0 \\ 1/2 & 0 & 1/2 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1/3 & 0 & 1/3 & 1/3 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1/2 & 1/2 & 0 \end{bmatrix}$$



Face à la réalité de la toile - 2

- **I. A n’est pas stochastique** : On ajoute des liens sortant et on obtient :

Face à la réalité de la toile - 2

- **I. A n'est pas stochastique** : On ajoute des liens sortant et on obtient :
- On transforme A ainsi:

$$\overline{A} = \begin{bmatrix} 0 & 1/2 & 1/2 & 0 & 0 & 0 \\ 1/2 & 0 & 1/2 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1/3 & 0 & 1/3 & 1/3 \\ 1/6 & 1/6 & 1/6 & 1/6 & 1/6 & 1/6 \\ 0 & 0 & 0 & 1/2 & 1/2 & 0 \end{bmatrix}$$

Face à la réalité de la toile - 3

$$\overline{A} = \begin{bmatrix} 0 & 1/2 & 1/2 & 0 & 0 & 0 \\ 1/2 & 0 & 1/2 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1/3 & 0 & 1/3 & 1/3 \\ 1/6 & 1/6 & 1/6 & 1/6 & 1/6 & 1/6 \\ 0 & 0 & 0 & 1/2 & 1/2 & 0 \end{bmatrix}$$

Face à la réalité de la toile - 3

- $$\overline{A} = \begin{bmatrix} 0 & 1/2 & 1/2 & 0 & 0 & 0 \\ 1/2 & 0 & 1/2 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1/3 & 0 & 1/3 & 1/3 \\ 1/6 & 1/6 & 1/6 & 1/6 & 1/6 & 1/6 \\ 0 & 0 & 0 & 1/2 & 1/2 & 0 \end{bmatrix}$$

- **II.** $\overline{A}$ **n'est pas irréductible** : Le Web n'est pas un graphe fortement connecté.

Face à la réalité de la toile - 3

$$\overline{A} = \begin{bmatrix} 0 & 1/2 & 1/2 & 0 & 0 & 0 \\ 1/2 & 0 & 1/2 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1/3 & 0 & 1/3 & 1/3 \\ 1/6 & 1/6 & 1/6 & 1/6 & 1/6 & 1/6 \\ 0 & 0 & 0 & 1/2 & 1/2 & 0 \end{bmatrix}$$

- **II.** $\overline{A}$ **n'est pas irréductible** : Le Web n'est pas un graphe fortement connecté.
- **III.** $\overline{A}$ **n'est pas apériodique.**

Face à la réalité de la toile - 4

- *On s'arrange* : On ajoute (vers toutes les pages) un lien sortant ayant une petite probabilité contrôlé par d.

Face à la réalité de la toile - 4

- *On s'arrange* : On ajoute (vers toutes les pages) un lien sortant ayant une petite probabilité contrôlé par d.
- $$P = [(1-d)\frac{E}{n} + dA^T]P$$

Face à la réalité de la toile - 4

- *On s'arrange* : On ajoute (vers toutes les pages) un lien sortant ayant une petite probabilité contrôlé par d.
- $$P = [(1-d)\frac{E}{n} + dA^T]P$$
- $$[(1-d)\frac{E}{n} + dA^T] = \begin{bmatrix} 1/60 & 7/15 & 1/60 & 1/60 & 1/6 & 1/60 \\ 7/15 & 1/60 & 11/12 & 1/60 & 1/6 & 1/60 \\ 7/15 & 7/15 & 1/60 & 19/60 & 1/6 & 1/60 \\ 1/60 & 1/60 & 1/60 & 1/60 & 1/6 & 7/15 \\ 1/60 & 1/60 & 1/60 & 19/60 & 1/6 & 7/15 \\ 1/60 & 1/60 & 1/60 & 19/60 & 1/6 & 1/60 \end{bmatrix}$$

L'algorithme PageRank (G)

- $P_0 \leftarrow \frac{e}{n}$

L'algorithme PageRank (G)

- $P_0 \leftarrow \frac{e}{n}$
- $k \leftarrow 1$

L'algorithme PageRank (G)

- $P_0 \leftarrow \frac{e}{n}$
- $k \leftarrow 1$
- repeat
  - $P_k = ((1-d)e + dA^T)P_{k-1}$
  - $k \leftarrow k+1$

L'algorithme PageRank (G)

- $P_0 \leftarrow \frac{e}{n}$
- $k \leftarrow 1$
- repeat
  - $P_k = ((1-d)e + dA^T)P_{k-1}$
  - $k \leftarrow k+1$
- until $||P_k - P_{k-1}|| < \epsilon$

L'algorithme PageRank (G)

- $P_0 \leftarrow \frac{e}{n}$
- $k \leftarrow 1$
- repeat
  - $P_k = ((1-d)e + dA^T)P_{k-1}$
  - $k \leftarrow k+1$
- until $||P_k - P_{k-1}|| < \epsilon$
- return P_k

Intégration du temps

- Privilégier les pages récentes.

Intégration du temps

- Privilégier les pages récentes.
- $d = f(t)$

Intégration du temps

- Privilégier les pages récentes.
- $d = f(t)$
- Si la page i est ancienne alors :
  - les pages citées par i sont encore plus anciennes
  - $1 - f(t)$ doit être assez élevée pour permettre de passer à d'autres pages non citées par i.

Intégration du temps

- Privilégier les pages récentes.
- $d = f(t)$
- Si la page i est ancienne alors :
  - les pages citées par i sont encore plus anciennes
  - $1 - f(t)$ doit être assez élevée pour permettre de passer à d'autres pages non citées par i.

Intégration du temps

- Privilégier les pages récentes.
- $d = f(t)$
- Si la page i est ancienne alors :
  - les pages citées par i sont encore plus anciennes
  - $1 - f(t)$ doit être assez élevée pour permettre de passer à d'autres pages non citées par i.

Conclusion et Perspectives

- Exploration de comportement utilisateurs (analyse de Logs).

Conclusion et Perspectives

- Exploration de comportement utilisateurs (analyse de Logs).
- Analyse de la structure (hyperliens).

Conclusion et Perspectives

- Exploration de comportement utilisateurs (analyse de Logs).
- Analyse de la structure (hyperliens).
- Web 2.0 : (web participatif et social).

Conclusion et Perspectives

- Exploration de comportement utilisateurs (analyse de Logs).
- Analyse de la structure (hyperliens).
- Web 2.0 : (web participatif et social).
- Besoins : détection de communautés, prédiction de liens.
  - Techniques de réseaux sociaux.
  - Fouille et exploration de liens.

Conclusion et Perspectives

- Exploration de comportement utilisateurs (analyse de Logs).
- Analyse de la structure (hyperliens).
- Web 2.0 : (web participatif et social).
- Besoins : détection de communautés, prédiction de liens.
  - Techniques de réseaux sociaux.
  - Fouille et exploration de liens.