Extraction des connaissances dans le BD (Data Mining)

Sommaire

Le Data Mining est un ensemble de techniques qui permet de récupérer de la base de données des données utiles.

Analyse Multicritères : SAW

a) Le problème

R(A1,A2, …,An) dont (Ai) numérique

b) Id0(identifiant)
c) Préférence $\in\{$ 'MIN', 'MAX '$\}$

Hotel	Id	Prix	Distance	NbEtoile
	H1	30	800	3
	H2	35	800	3
	H3	60	400	3
	H4	60	400	3
	H5	50	300	4
	H6	60	300	4

SELECT * FROM HOTEL WHERE (Prix,Distance,NbEtoile)=(SELECT Min(Prix),Min(Distance),Max(NbEt) FROM HOTEL) ;

I = <30,100,5> # appelle le tuple parfait

Sauf que le tuple parfait n'existe pas dans notre table, il va donc falloir chercher autrement quel est le meilleur tuple selon certains critères de préférences.

1)La méthode SAW

SAW : Simple Aggregative Weighting

Etape 1 : Normalisation des critères : r-Norm

$\forall$ u $\in$r -norm

$\forall$A $\in${ A1,A2,…,An} = R

U(A) = {

T(A)/Max(A) si A est à Maximiser

Min(A) / t(A) si A est à Minimiser

}

Hotel-norm	IdH	Prix-normalisé	Distance-normalisé	NbEtoile-normalisé
	H1	1	0.125	0.6
	H2	0.827	0.125	0.6
	H3	0.5	0.25	0.25
	H4	0.5	1	1
	H5	0.6	0.833	0.8
	H6	0.5	0.333	0.8

Etape 2 : Pondération des critères normalisés : r -Pond

W(A) = Coeff(A)/Somme Coeff(Ai) (Si on a des coefficients)

Si on décide de trier par poids, il faut définir des poids , généralement c'est l'utilisateur qui décide de ses critères

Poid : 0.5 , 0.25 , 0.25

H1 Pondéré : 1x0.5 , 0.125x0.25 , 0.6x0.25

0.5, 0.03, 0.15

Etape 3 : Aggregation des critères en un seul critère

Score : r-score (IdH,Score)

U(Score) = $\sum$ n, i = 1 +(Ai) t $\in$ r-Pond

(En gros on additionne chaque valeur de chaque tuple pour donner le score finale)

Etape 4 : Trie et affichage des objets selon le score en ordre décroissant

r-score	IdH	Score
	H1	0.68
	H2	0.608
	H3	0.46
	H4	0.75
	H5	0.58
	H6	0.53

2) Implémentation avec SQL

Etape 1

CREATE VIEW Hotel-Norm

As

SELECT H.IdH , Trunc (M.min-prix / H.prix,3), Prix-Norm, Trunc(M.Min-distance/H.distance,3) Distance -norm, Trunc(H.NbEt/M.max-nbet,3) NbEt-norm

FROM HOTEL H,(SELECT Min(Prix) MinPrix, Min(distance) Min-Distance, Max(NbEt) MaxNbEt FROM HOTEL) M ;

Etape 2

CREATE VIEW Hotel-Pond

As

SELECT IdH, Trunc(0.5xPrix-norm,3)

Prix-Pond,

Trunc(0.25x Distance-Norm,3) Distance -Pond,

Trunc(0.25 x NbEt-norm,3) NbEt-Pond

FROM Hotel-Norm ;

Etape 3

CREATE VIEW HOTEL-Score

As

SELECT IdH, Trunc(Prix-Pond + Distance-Pond + NbEt-Pond, 3) score

FROM Hotel-Pond ;

Etape 4

```
SELECT H.IdH, H.Prix, H.Distance,H.NbEt,Score
FROM HOTEL H,Hotel-Score S
WHERE H.IdH = S.Idh
ORDER BY SCORE DESC ;
```

Analyse des liens web : l'algorithme PageRank de Google

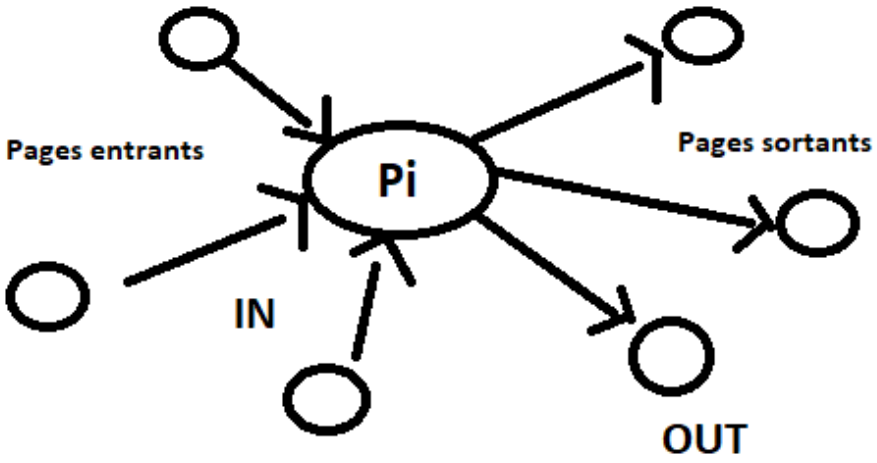
1) Le Graphe Web

W = (R,L)

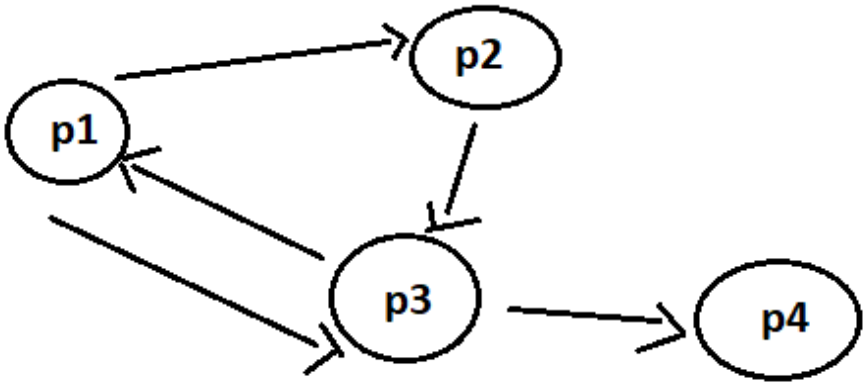
R = { P1,P2,…,Pn} l'ensemble des pages

Une page web est importante si elle est pointée par des pages importantes

	P1	P2	P3	P4
P1	0	1	1	0
P2	0	0	0	0
P3	1	0	0	1
P4	0	0	0	0



Nous prendrons ce schéma pour exemple :



PR(Pi) = α $\sum$pj $\in$ In(Pi) PR(Pj)/ | out(Pj) | + (1 - α)

PR(P1) = PR(P3)/2

PR(P4) = PR(P3)/2

PR(P3) = PR(P1)/2 + PR(P2)

PageRank(M,α,$\sum$)

Pr0 = (1/n , 1/n , … , 1/n)

K=0

Repeter

K = K+1

Pour i = 1 à n faire

PR k(Pi) = α $\sum$pj $\in$ In(Pi) PR k-1 (Pi) /| out(pj) | + (1- α)

Jusqu'à ||Pr k-1 – PR k ||1 < $\sum$

Renvoyer PR k

I(nombre d'itérations à faire) >= -P/log10(α)

Exemple :

$\alpha = 0.5$; P = 2 ; I = 7

$\alpha = 0.95$; P = 2 ; I $\approx$ 90

On remarque que plus delta est grand, plus le nombre d'itérations nécessaire est grand

Analyse d'associations entre items : l'algorithme APRIORI

1) Bases de transactions (Bd non structuré)

   Soit I un ensemble fini d'items

2) Extractions des motifs Fréquents

   2.1) Fréquence d'un motif
   Soit $X \subseteq V$, Freq (X, D) =| {y $\in$ D | $X \subseteq y$ } | / |D|

   II = {A,B ,C,D,E}

IdT	Motifs , Transations
1	ACD
2	BCE
3	ABCE
4	BE

Ex

X	Freq(X,D)
BD	0
A	2/4 = 50%
BE	2/4
CD	¼
AC	2/4

2.2) Motifs Frequents

Soit MinFreq $\in$ [0,1] un seuil de Fréquence donné par l'utilisateur

X est un motif Frequent ssi Freq (X,D) > Minfreq

L = { $X \subseteq$ II | Freq (X,D) > Min Freq } (L'ensemble des motifs Frequents de D avec MinFreq comme seuil de fréquence

Propriétés des motifs Frequents

P1) Tout sous-ens propre d'un motif Frequent est un motif Frequent

P2) Tout sur-ens d'un motif non Frequent est un motif non Frequent

--SUITE lors du prochain AMPHI --