LOG6308 — Systèmes de recommandation

Systèmes de recommandations 2, Approches contenu et popularité

Michel C. Desmarais

Génie informatique et génie logiciel
École Polytechnique de Montréal

Automne, 2017
(version 29 août 2017)

Approches collaboratives vs. approches contenu

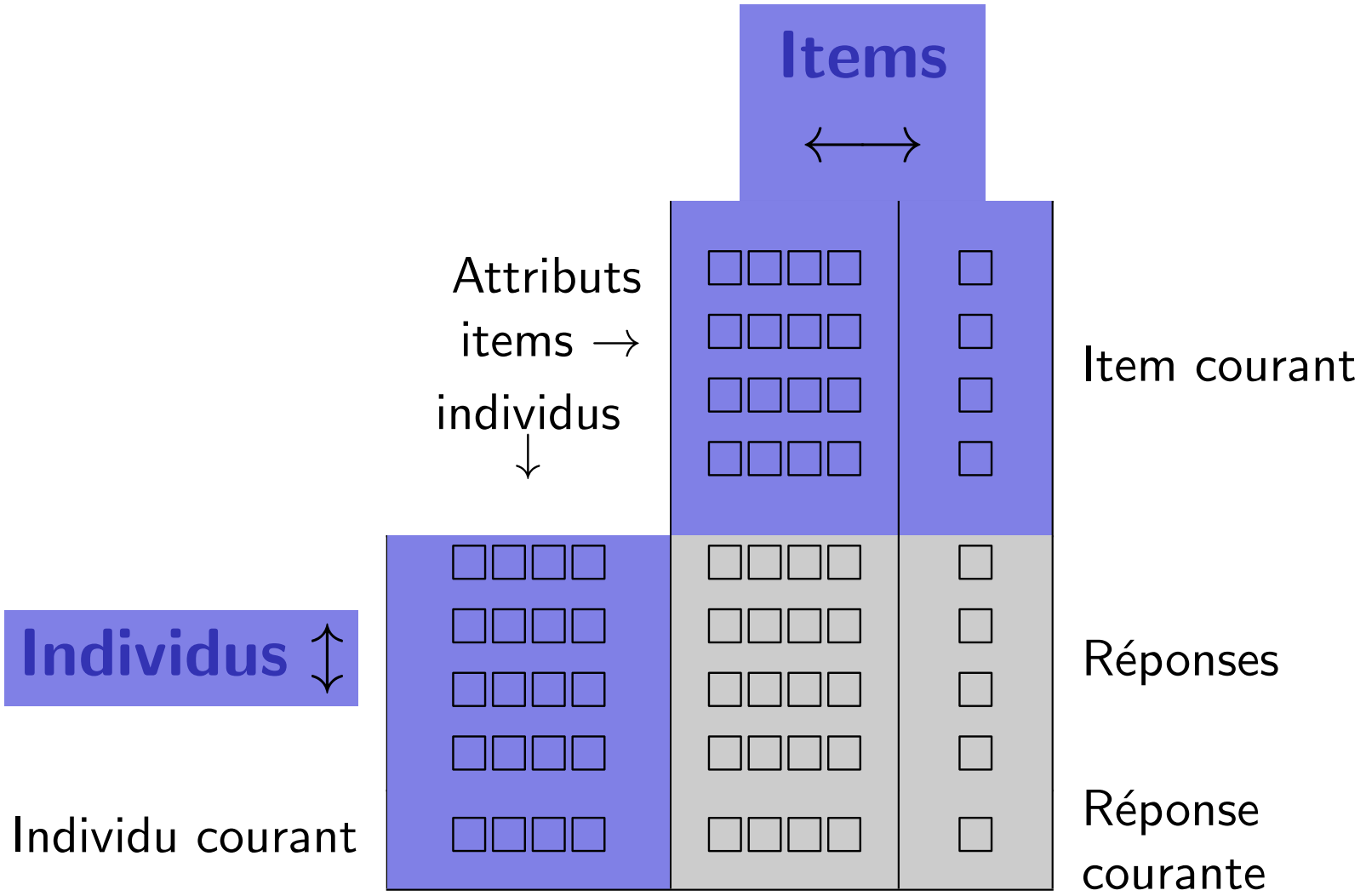
- Les approches collaboratives utilisent des mesures directes et indirectes de **l'intérêt** pour mesurer des similarités.
- Les approches contenu utilisent plutôt des attributs **sémantiques**, **démographiques**, ou autres attributs associés aux items ou aux utilisateurs afin d'établir la similarité.
- Les approches contenu ont plusieurs d'affinités avec les techniques de recherche d'information.

Basé modèle contre *basé mémoire*

- Les approches items-items et utilisateurs-utilisateurs sont dites **basé mémoire**.
- Leur complexité est de l'ordre de $\mathcal{O}(mn^2)$, où n est le nombre d'utilisateurs et m le nombre d'items (pour l'approche item-item—l'inverse pour u-u).
- Pour des applications où le temps de réponse est critique et où la BD est grande, le calcul *basé mémoire* doit s'effectuer au préalable et le résultat stocké en mémoire pour une consultation rapide.
- Si l'information pertinente change, comme c'est le cas pour les utilisateurs qui ont peu d'historique, l'approche n'est pas idéale.
- L'approche **basée sur un modèle** permet de combler cette lacune et le modèle probabiliste est un très bon candidat.

ÉCOLE POLYTECHNIQUE MONTRÉAL

Types d'information



Systèmes de recommandations 2, Approches contenu et popularité

Principes

- Utilisation de facteurs propres aux items ou aux utilisateurs :
  - **utilisateurs** : âge, profession, code postal, etc.
  - **items** : auteur, mots-clés, année de production, etc.
- Probabilités conditionnelles :
  - certains items s'adressent à des populations spécifiques et ils sont reconnaissables par leur attributs de contenu ;
  - les votes les plus probables peuvent s'exprimer sous la forme de probabilité conditionnelle :
    probabilité du vote v étant donné les attributs utilisateurs et items

Un exemple I

Vote au film *Toy Story* par un ingénieur

Quelle est la probabilité d'aimer le film *Toy Story* étant donné que je suis un ingénieur :

$$P(v_{ij} = x | A_{ik} = \text{“ing”}) = \frac{P(v_{ij} = x,\ A_{ik} = \text{“ing”})}{P(A_{ik} = \text{“ing”})}$$

où v_{ij} est le vote pour le film j (*Toy story*) de l'utilisateur i et A_{ik} est l'attribut *profession* de l'utilisateur i.

Cet estimé se calcule sur la base de ratio de fréquences où l'on ajoutera la **correction de Laplace** :

$$\frac{x_1 + 1}{(x_1 + x_2) + 2}$$

Cette correction intègre une notion de probabilité antérieure et évite les biais des petits échantillons de même que les ratios singuliers de 1/1, 0/1, 0/0.

Un exemple II

Vote au film *Toy Story* par un ingénieur

Supposons maintenant le vote d'un homme ingénieur.

On a alors deux conditions :

$$P(v_{ij} = x | A_{ik_1} = \text{“ing”},\ A_{ik_2} = \text{“H”})$$

$$= \frac{P(v_{ij}=x,\ A_{ik_1}=\text{“ing”},\ A_{ik_2}=\text{“H”})}{P(A_{ik_1}=\text{“ing”},\ A_{ik_2}=\text{“H”})}$$

Le problème avec ce calcul est que la combinaison de facteurs divise les fréquences à chaque facteur que l'on ajoute. On peut rapidement se retrouver avec aucun cas. **Solution** :On postule alors l'indépendance des facteurs.

Un exemple III

Vote au film *Toy Story* par un ingénieur

En postulant l'indépendance des facteurs, on obtient :

$$P(v_{ij}|A_{ik_1},\ A_{ik_2}) \propto P(v_{ij})P(A_{ik_1}|v_{ij})P(A_{ik_2}|v_{ij})$$

De manière plus générale, on a :

$$p(C|F_1,\ldots,F_n) = \frac{1}{Z}p(C)\prod_{i=1}^{n} p(F_i|C)$$

où C est une classe et F_i sont des attributs. Z est une valeur qui ne dépend pas de C, de sorte que lorsqu'il s'agit d'estimer la classe la plus probable on peut l'ignorer.

Systèmes de recommandations 2, Approches contenu et popularité

Approches modèles

- Plusieurs modèles ont été proposés pour fournir une estimation probabiliste des votes.
- Nous révisons ici quelques uns proposés et évalués dans Breese, Heckerman et Kadie, 1998.

Calcul de la valeur attendue du vote

La valeur attendue d'un vote (sur une échelle ordonnée) est la somme pondérée de la probabilité de chaque valeur par sa probabilité :

$$E(v) = \sum_i P(v = i)\ i$$

La formulation de cette équation dans le contexte d'estimer le vote de l'utilisateur actif, a pour un item j est :

$$E(v_{a,j}) = \sum_i P(v_{a,j} = i | v_{a,k}, k \in I_a)\ i \tag{1}$$

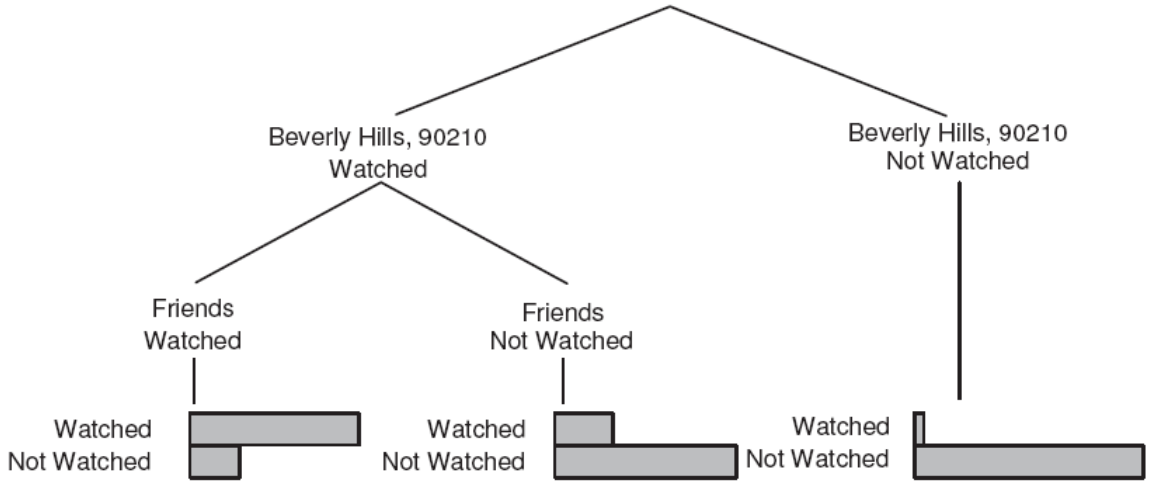
c.-à-d. que $E(v_{a,j})$ est la somme pondérée des probabilités de chaque vote étant donnée les votes aux autres items pour l'utilisateur actif.

Modèle par classes d'utilisateurs (cluster)

- *Définition de classes*
  Une première approche repose sur la définition de classes d'utilisateurs. Ces classes sont définies en fonction de leur capacité à discriminer les préférences entre les utilisateurs.
- $P(C{=}c, v_1, \ldots, v_n) = P(C{=}c) \prod_i P(v_i | C{=}c)$
  Ensuite, on calcule les probabilités qu'un utilisateur appartienne à chaque classe respective à la lumière de ses votes et on utilise ces probabilités dans l'équation (1) (voir Breese, Heckerman et Kadie, 1998, section 2.3.1).
- Cette approche nécessite le recours à l'algorithme EM pour obtenir probabilités et de déterminer un nombre adéquat de classes par des techniques d'exploration statistiques.

Modèle basé sur les noeuds observables (1)

- Une autre approche probabiliste consiste à dériver des probabilités conditionnelles entre les noeuds eux-mêmes.
- Breese et coll. utilisent par exemple un arbre de décision comme celui ci-dessous. Les noeuds terminaux représentent la probabilité de visionner “Melrose Place” conditionnellement à avoir visionné les noeuds parents.



Modèle basé sur les noeuds observables (2)

- Plutôt qu'un arbre de décision, on pourrait aussi faire un calcul de probabilité postérieure :

$$P(v_i|v_j \in S) \propto \prod_{j \in S} P(v_j|v_i)$$

- Il faut faire une sélection de noeuds pertinents, $v_j \in S$, puis établir les valeurs de $P(v_j|v_i)$ à partir des données.
- Un test d'indépendance entre v_i et tous les autres noeuds v_j permet d'effectuer une première sélection de façon simple.
  - Par exemple le χ^2
  - On conserve les noeuds qui démontrent un lien fort
- Cette approche est semblable aux réseaux bayésiens à la différence qu'on ne créé que des réseaux bayésiens dits naïfs.

ÉCOLE POLYTECHNIQUE MONTRÉAL

Hypothèses et entrainement

- *Hypothèse d'indépendance*
  Les calculs reposent sur une hypothèse d'indépendance des votes étant donné l'appartenance à une classe, ou l'indépendance conditionnelle des votes entre eux.
- *Entraînement du modèle*
  Dans le cas de l'approche avec les classes, le calcul nécessite un entraînement avec des données et repose sur (1) un algorithme pour définir les classes optimales (les plus discriminantes) et (2) estimer les paramètres du modèle (les probabilités conditionnelles). Breese et al. utilisent respectivement la vraisemblance maximale et l'algorithme EM (*expectation maximization*) pour (1) et (2).

Modèles graphiques (réseau bayésien et réseau de dépendances)

- *Approches de modèles graphiques*
  D'autres approches se fondent sur des modèles graphiques indiquant les interdépendances entre les votes et les items pour inférer la probabilité d'un vote à un item donné. Breese et al. (2001) proposent une approche basée sur les modèles bayésiens et Heckerman et al. suggèrent un modèle basé sur les réseaux de dépendances (Heckerman, Chickering, Meek, Rounthwaite et Kadie, 2000).
- *Acuité et performance*
  Les deux approches sont comparables en termes d'acuité et performent de 2% à 4% mieux que les méthodes d'espaces vectoriels dans les expériences rapportées. L'approche de réseaux de dépendance est cependant moins gourmande en ressources mémoire et de calcul.

Avantages et désavantages de l'approche *modèle*

- *Avantages*
  Ces modèles ont l'avantage d'être plus précis pour prédire les votes. Ils peuvent aussi être plus rapides dans la mesure où l'on peut créer un modèle avec des paramètres plutôt que prendre une approche où l'on doit faire des calculs basés sur l'ensemble des données (approche dite *basée mémoire*).
- *Désavantages*
  Cependant, leur déploiement est nettement plus complexe et nécessite une maîtrise des techniques de modélisation bayésienne et d'estimation de paramètres relativement sophistiquées.

Fréquence inverse des votes et amplification

- **Fréquence inverse des votes.** Comme discuté précédemment, on peut donner plus de poids aux votes moins fréquents en modulant le poids par l'équivalent de la fréquence inverse des documents.
- **Amplification des votes faibles.** Une seconde transformation du vote consiste à pénaliser les votes qui sont plus faibles pour compenser le fait que les utilisateurs ont moins tendance à voter pour les items qu'ils n'aiment pas.

Les résultats de Breese et coll.

Les algorithmes utilisés

BN	Réseau bayésien avec un arbre de décision.
CR+	Corrélation avec la fréquence inverse utilisateur et amplification
VSIM	Cosinus avec la fréquence inverse utilisateur
BC	Modèle par classe
POP	Vote majoritaire (prédiction fixe)
Rd	Aléatoire

Les résultats de Breese et coll.

Les données utilisées

MS Web : visites de sections du site web de Microsoft (vote binaire implicite).

Nielsen : programmes de télé regardés durant une période de deux semaines (vote binaire implicite).

EachMovie : vote explicite variant de 0 à 5 et portant sur des films.

Les résultats de Breese et coll.

Performances spécifiques

Résultats moyens des recommandations en utilisant 5 votes, la mesure d'utilité et 10 simulations.

	Neilsen	MSWeb	EachMovie
Algorithme			
BN	42,2	59,8	42,3
CR+	43,3	58,9	42,1
VSIM	40,0	56,1	36,7
BC	18,9	54,8	33,2
POP	19,5	46,9	28,9
Rd	1,8	1,8	0,8

Les résultats de Breese et coll.

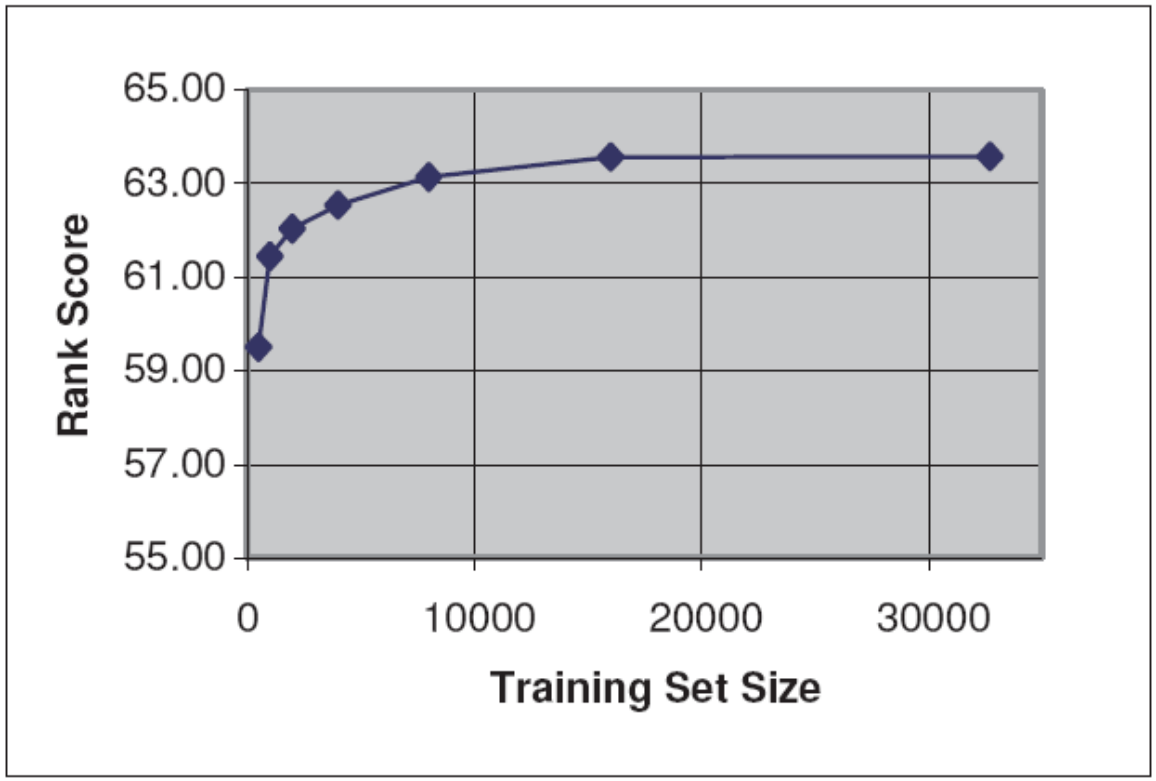
Autres résultats

- L'utilisation de la fréquence inverse utilisateur améliore les résultats mais de façon variable.
  - 1 à 3% pour la corrélation et le cosinus pour la prédiction des recommandations
  - 3 à 25% pour la prédiction d'un vote individuel et plus grande amélioration pour le cosinus que la corrélation ; ne s'applique qu'aux données EachMovie
- L'amplification du score améliore aussi les résultats de 1 à 10% mais surtout pour les données de télévision et très peu pour EachMovie

Les résultats de Breese et coll.

Taille du corpus

La taille du corpus d'entrainement influence la performance jusqu'à atteindre un seuil maximal.

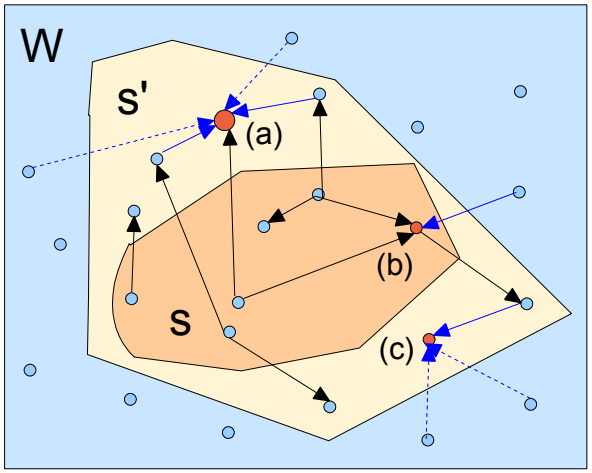


Systèmes de recommandations 2, Approches contenu et popularité

Popularité, pertinence et références

- Les items les plus populaires intéressent plus d'individus, par définition.
- La popularité peut se mesurer par les références à un document.
- Les références dans le cadre du Web sont les liens.
- L'algorithme PageRank est particulièrement bien adapté à la mesure de popularité.

Les liens entrants comme facteur de pertinence



- Il existe plusieurs variantes possibles selon qu'on différencie entre les types de liens ou qu'on ne tient compte que de s et s', par exemple.

- Soit une requête dans W (le Web recensé) qui retourne l'ensemble de pages s correspondant à la requête de mots clés.
- Soit s', les pages référencées par les pages de s
- Selon l'algorithme PageRank, (b) serait la page la plus pertinente car elle est la plus référencée de s. Cependant, (a) pourrait aussi l'être selon d'autres algorithmes qui incluraient aussi s' (cf. *miserable failure*).

ÉCOLE POLYTECHNIQUE MONTRÉAL

Algorithme de base

L'algorithme de base se fonde sur le calcul suivant :

$$\texttt{PageRank}(A) = (1-d) + d \sum_{D_1 \ldots D_n} \frac{\texttt{PageRank}(D_i)}{C(D_i)}$$

$D_1 \ldots D_n$ sont les pages qui réfèrent à A (liens entrants) ;

$C(D_i)$ est le nombre de références de la page i (liens extrants) ;

d est un facteur d'ajustement dans l'intervalle $[0, 1]$, notamment pour permettre aux pages sans liens entrants d'avoir une cote et pour faciliter la convergence pour les calculs des pages référencées par de telles pages. Une valeur courante est 0,85.

Cet algorithme est itératif et les valeurs convergent après quelques d'itérations (souvent en deça de 10). On peut initialiser les PageRank et ajouter 1 à $C(D_i)$ pour éviter des divisions par 0.

ÉCOLE POLYTECHNIQUE MONTRÉAL

Interprétation et normalisation

- Le modèle calcule la probabilité que l'on accède à une page en suivant les liens de façon aléatoire.
- Le facteur d'atténuation, d, représente la probabilité qu'un individu continu d'accéder à des pages ; donc $(1 - d)$ est la probabilité qu'il arrête.

Pour que la somme des PageRank donne 1, on divise le facteur d'atténuation par N, le nombre total de pages :

$$\texttt{PageRank}(A) = \frac{(1-d)}{N} + d \sum_{D_1 \ldots D_n} \frac{\texttt{PageRank}(D_i)}{C(D_i)}$$

Si on considère le réseau de liens comme une *chaîne de Markov* où les pages sont des états, le PageRank représente en fait la probabilité qu'un individu soit sur une page donnée par une navigation aléatoire.

Expression matricielle de l'équation PageRank

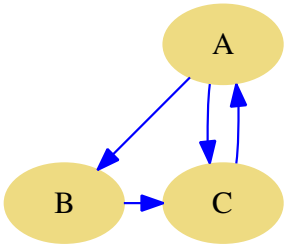
Posons les variables :

- $\mathbf{r}$ vecteur des valeurs de PR
- $\mathbf{s}$ vecteur du nombre de liens sortants
- $\mathbf{A}$ matrice d'adjance
- d le facteur d'amortissement (*damping*)
- $\mathbf{d}$ un vecteur d'amortissement de longueur n contenant
- n le nombre de de documents

Alors, la valeur de R est la suivante :

$$\mathbf{r} = (1 - \mathbf{d})/n + d\mathbf{A}(\mathbf{r}/\mathbf{s})$$

Un exemple



Le graphe ci-dessus correspond à la matrice d'adjacence :

	a	b	c
a	0	0	1
b	1	0	0
c	1	1	0

En supposant un valeur $d = 0,85$ et en initialisant le PR de chaque page à 1, les valeurs des PR à chaque itération sont alors :

Itération	A	B	C
0	1,000	1,000	1,000
1	0,900	0,475	1,325
2	1.176	0.432	0.836
3	0.760	0.549	0.917
4	0.829	0.373	0.840
5	0.764	0.402	0.720
6	0.662	0.374	0.717
7	0.659	0.331	0.650
8	0.602	0.330	0.612
9	0.570	0.306	0.586
10	0.548	0.292	0.552
...	...		
30	0.393	0.217	0.403

ÉCOLE POLYTECHNIQUE MONTRÉAL

Révision

- L'algorithme PageRank a révolutionné la recherche d'information sur le web, pourquoi semble-t-il mieux fonctionner que l'approche de l'espace vectoriel ou l'approche probabiliste ?
- Comment PageRank pourrait-il être utilisé pour effectuer des recommandations d'articles ? Expliquez comment et en quoi les résultats trouvés seraient différents des autres approches ?

Graphes et matrices

En représentant un graphe sous la forme d'une matrice d'adjacence (ou *de transition*), il est possible de déterminer les chemins transitifs entre les noeuds par la puissance de la matrice. Ainsi, supposons une matrice de transition, $\mathbf{M}$, entre quatre noeuds (a,b,c,d) et où $(i, j) = 1$ indique l'existence d'un lien $i \rightarrow j$:

$$\mathbf{M} = \begin{array}{c|cccc} & \mathbf{a} & \mathbf{b} & \mathbf{c} & \mathbf{d} \\ \mathbf{a} & 0 & 0 & 0 & 0 \\ \mathbf{b} & 1 & 0 & 0 & 0 \\ \mathbf{c} & 0 & 1 & 0 & 1 \\ \mathbf{d} & 1 & 1 & 0 & 0 \end{array} \quad \text{et} \quad \mathbf{M}^2 = \begin{array}{c|cccc} & \mathbf{a} & \mathbf{b} & \mathbf{c} & \mathbf{d} \\ \mathbf{a} & 0 & 0 & 0 & 0 \\ \mathbf{b} & 0 & 0 & 0 & 0 \\ \mathbf{c} & 2 & 1 & 0 & 0 \\ \mathbf{d} & 1 & 0 & 0 & 0 \end{array}$$

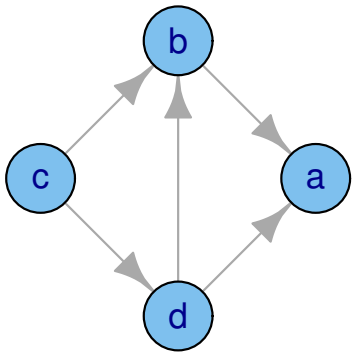
$\mathbf{M}^2$ représente ainsi les chemins de deux arcs possibles. Par exemple, la valeur de 2 pour indique que deux chemins de deux arcs sont possibles entre $c \rightarrow a$.

ÉCOLE POLYTECHNIQUE MONTRÉAL

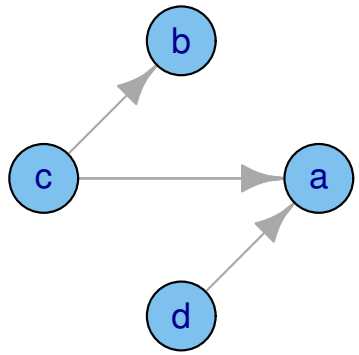
Graphes et matrices (suite)

Les graphiques suivants qui correspondent respectivement à $\mathbf{M}$ et $\mathbf{M}^2$:

M



M * M



Noter aussi que $\mathbf{M}^3$ n'aurait plus qu'un seul lien : $c \longrightarrow a$

ÉCOLE POLYTECHNIQUE MONTRÉAL