Web mining

Shengrui Wang

22 Novembre 2024

Université de Sherbrooke

Plan

Introduction

- Le Web pourrait être vu comme un repertoire de données,
- Il est caractérisé par :
  - **son grand volume de données** : en perpétuelle augmentation, la quantité de données disponibles est immense.
  - **l'hétérogénéité des données** : Les données présentes sont de types et natures diverses.
  - **l'interconnexion du contenu** : Les liens d'une page vers d'autres pages créent une toile facilitant la navigation et la découverte des contenus.
  - **le bruit** : Vu comme un média libre, toute personne est donc capable d'y poster une information qui pourrait être importante ou pas.
  - **sa vélocité** : En perpétuelle croissance dû à l'ajout perpétuel de contenus par les utilisateurs.
  - **sa sociabilité** : Les utilisateurs peuvent collaborer entre eux.

Introduction

- Bien qu'il soit vu comme un repertoire *fourre-tout*, où l'on peut incorporer tout type d'information, il n'en demeurre pas moins que les liens entre les différentes pages de contenus constituent une structure topologique bien définie.

Introduction

- Ainsi présenté, le Web n'est plus seulement vu comme un repertoire *fourre-tout*, mais aussi une *toile* organisée où l'on pourrait naviguer.
- La fouille de données a pour objectif l'extraction de connaissances à partir
  - **de la structure formée par les liens entre pages/objets/communautés** : Forage de structure. Ce type de fouille vise essentiellement à analyser la structure du Web.
  - **du contenu des pages Web** : Forage de contenu, Ce type de fouille vise essentiellement à analyser le contenu des pages.
  - **des données d'usages** (typiquement des logs de serveurs web) : forage de données d'usage.

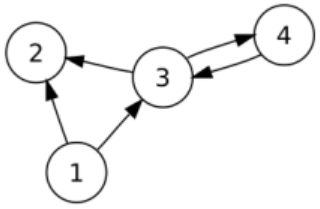
Forage de structure web

- Les premiers moteurs de recherche retrouvaient les documents pertinents **seulement en fonction de la similarité** du contenu avec la requête de l'utilisateur.
- l'augmentation rapide du nombre de pages Web a rendu **très difficile l'ordonnancement** des documents retrouvés.
- il est **très facile de corrompre** le contenu d'une page de telle sorte à la rendre similaire à des très nombreuses requêtes.
- Pour contourner ce problème, il s'est donc posé la question d'**exploiter les liens entre les pages**. En d'autres termes, exploiter la structure topologique du web.

Forage de structure web

- La structure topologique du Web peut-etre représentée par une structure de graphe
- où les noeuds représenteraient les différentes pages, tandis que les liens seraient les hyperliens.
- De cette manière, on pourrait donc exploiter les propriétés statistiques du graphe et ainsi déduire de la pertinence de certaines pages Webs.

Forage de structure web



Forage de structure web

- Formellement, la structure topologique du Web ou d'un site Web pourrait être donnée par le graphe $\boldsymbol{G} = (\boldsymbol{V}, \boldsymbol{E})$, où
- $\boldsymbol{V}$ serait l'ensemble des noeuds représentant les différentes pages webs et
- $\boldsymbol{E} = \{(u, v) / u \in \boldsymbol{V}, v \in \boldsymbol{V}\}$ l'ensemble des liens representant les hyperliens entre les pages webs.
- Lorsque le graphe est non-dirigé on a, $(u, v) = (v, u)$ et lorsqu'il est dirigé on pourrait avoir $(u, v) \neq (v, u)$.

Forage de structure web – Importance d'une page web

- En fonction de sa position dans la structure topologique, on pourrait évaluer sa pertinence en utilisant les mesures de centralités
- Degré d'un noeud:

$$D(u) = |\{(u, v) \in \boldsymbol{E} / v \in \boldsymbol{V}\}|$$

- Prestige d'un noeud (fonctionnel uniquement dans les graphes dirigés), Il représente encore le dégré entrant.

$$P(u) = |\{(v, u) \in \boldsymbol{E} / v \in \boldsymbol{V}\}|$$

- L'intermédiarité:

$$I(u) = \sum_{v \in \boldsymbol{V} \setminus \{u,w\}} \sum_{w \in \boldsymbol{V} \setminus \{u,v\}} SP_{v,w}(u)$$

avec $SP_{v,w}(u)$ le nombre de plus court chemins entre v et w en passant par u.

PageRank

Papier :

Lawrence Page and Sergey Brin

"The PageRank citation ranking: Bringing order to the Web"

http://coblitz.codeen.org:3125/citeseer.ist.psu.edu/cache/papers/cs/7144/http:zSzSzwww-db.stanford.eduzSz~backrubzSzpageranksub.pdf/page98pagerank.pdf

PageRank

Définition

➢ Le PageRank est une valeur numérique (une note qui varie sur la barre d'outil Google de 0 à 10) que Google attribut a une page web pour représenter (refléter) son importance par rapport à d'autres pages du même site ou par rapport à n'importe quelle autre page sur le web.

➢ Le PageRank sera bénéfique à une page Web pour le positionnement sur google, ce n'est pas le seul facteur mais il a une grande importance. Google calcul cette note (le PR) en analysant principalement le nombre et la qualité des liens entrant et des liens sortant pour chaque page sur le web.

Chaque lien placé sur une page **A** vers une page **B** est interprété par Google comme un VOTE de **A** à **B**.

le plus de liens entrant une page, plus haute sa note sera (son PageRank). Les liens sortant la feront baisser.

PageRank

Remarque

➢ **Un vote émis par la page d'accueil d'un site majeur tel que Microsoft pèse beaucoup plus lourd qu'un vote émis par la page perso de votre site.**

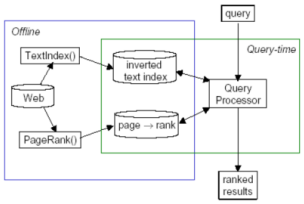
➢ **Le PR est une mesure pour déterminer l'importance d'une page et non d'un site en entier**

➢ **Le PR d'une page peut être visualisé par les utilisateurs de « Google toolbar »**

En résumé …

Le PageRank d'une page Web, c'est-à-dire son poids dans l'index de Google, dépend donc non seulement de la qualité du contenu d'un site, mais aussi, par extension, de celle des sites qui le lient.

Cycle de vie d'une requête



➢ Les calculs les plus complexes et les plus longs se font offline.

➢ Le moteur n'a plus que des calculs très simples à effectuer pour construire les pages de résultats

PageRank

Calcule du PageRank

Une page A reçoit des liens émis par les page T1, …, Tn

$$PR(A) = (1-d) + d * \left[\frac{PR(T1)}{C(T1)} + \cdots + \frac{PR(Tn)}{C(Tn)} \right]$$

d : un facteur d'amortissement « damping factor » pouvant être ajusté entre 0 et 1. Généralement *d* est entre 0.8 et 0.9, la valeur suggérée $d = 0.85$

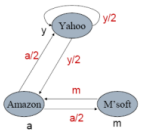
C(Ti) : le nombre de liens émis par la page Ti (liens sortant) « Out-links »

➢ Le PageRank peut être calculé en utilisant un simple algorithme itératif

➢ On remarque que le PageRank d'une page ne peut être inférieur a 0.15 même s'il n'y a aucune page qui mène à elle. Donc une page aurait toujours un minimum de 0.15 à partager.

Intuition derrière la formule de PageRank

Estimation d'une matrice des liens Web



➢ Les lignes et les colonnes de la matrice sont les page Web. Soit X_{ij} est un élément de la matrice

➢ **Construction de la matrice**

Si la page *j* a *n* liens sortants **alors**
$X_{ij} = 1/n$ s'il y a un lien sortant de la page *i* vers *j*
Sinon $X_{ij} = 0$

⇩

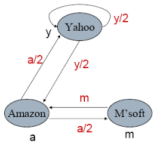
➢ les liens entre les page Web sont présentés dans une matrice stochastique (La somme des éléments de chaque colonne = 1)

	y	a	m
y	1/2	1/2	0
a	1/2	0	1
m	0	1/2	0

➢ On remarque que Yahoo! divise sont importance avec lui-même et Amazon. Microsoft donne toute sont importance à Amazon

Intuition derrière la formule de PageRank

Estimation d'une matrice des liens Web



➢ Chaque vote est proportionnel à l'importance de sa page source.

➢ Soit *P* une page Web avec *n* liens sortant et une valeur d'importance = *x*
Chaque lien sortant de *P* aura un poids = *x*/*n*

➢ Résultat : 3 équations à 3 inconnus (en supposant *d*=1)

$$y = y/2 + a/2$$
$$a = y/2 + m$$
$$m = a/2$$

Estimation de l'importance des pages Web (en supposant *d*=1)

❑ Soit [y, a, m] un vecteur qui représente l'importance de chaque page: Yahoo!, Amazon, Microsoft dans cet ordre.

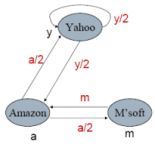
➢ L'équation qui décrive ces trois variable est définie par:

$$\begin{bmatrix} y \\ a \\ m \end{bmatrix} = \begin{bmatrix} 1/2 & 1/2 & 0 \\ 1/2 & 0 & 1 \\ 0 & 1/2 & 0 \end{bmatrix} \begin{bmatrix} y \\ a \\ m \end{bmatrix}$$

$$y = y/2 + a/2$$
$$a = y/2 + m$$
$$m = a/2$$

Solution de cette équation

y		1/3	1/3	5/12	3/8	2/5
a	=	1/3	1/2	1/3	11/24	… 2/5
m		1/3	1/6	1/4	1/6	1/5



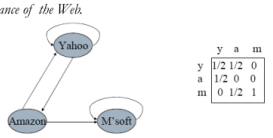
➢ **Yahoo! et Amazon ont la même valeur d'importance, et le double de la valeur de l'importance de Microsoft**

Problème … Spider trap

Spider traps : un groupe d'une ou plusieurs pages qui n'ont aucun lien sortant vont accumuler toute les valeurs d'importance des pages web.

***Spider traps**: a group of one or more pages that have no links out of the group will eventually accumulate all the importance of the Web.*

Exemple



	y	a	m
y	1/2	1/2	0
a	1/2	0	0
m	0	1/2	1

y		1	1	3/4	5/8		0
a	=	1	1/2	1/2	3/8	...	0
m		1	3/2	7/4	2		3

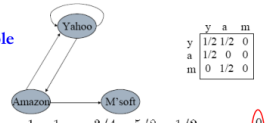
➢ **Microsoft devient « a spider trap »**

Problème … Dead end

Dead ends : une page qui n'a aucun lien sortant.

***Dead ends**: a page that has no successors has nowhere to send its importance. Eventually, all importance will "leak out of" the Web.*

Exemple



	y	a	m
y	1/2	1/2	0
a	1/2	0	0
m	0	1/2	0

y	1	1	3/4	5/8	1/2		0
a	1	1/2	1/2	3/8	5/16	...	0
m	1	1/2	1/4	1/4	3/16		0

➢**Microsoft devient « a dead end »**

Solution de Google … PageRank

L'idée

"Instead of applying the matrix directly, "tax" each page some fraction of its current importance, and distribute the taxed importance equally among all pages"

$$PR(A) = (1-d) + d * \left[\frac{PR(T1)}{C(T1)} + \cdots + \frac{PR(Tn)}{C(Tn)} \right]$$

Exemple : tax = d = 20%

$$\begin{bmatrix} y \\ a \\ m \end{bmatrix} = 0.8 \begin{bmatrix} 1/2 & 1/2 & 0 \\ 1/2 & 0 & 0 \\ 0 & 1/2 & 1 \end{bmatrix} \begin{bmatrix} y \\ a \\ m \end{bmatrix} + \begin{bmatrix} 0.2 \\ 0.2 \\ 0.2 \end{bmatrix}$$

y = 7/11; a = 5/11; m = 21/11

Cette distribution de l'importance des pages Web est plus raisonnable que La première fois sans le d : (y = 0; a = 0; m = 3)

Exemple de calcul de PageRank

$$PR(A) = (1-d) + d * \left[\frac{PR(T1)}{C(T1)} + \cdots + \frac{PR(Tn)}{C(Tn)} \right]$$

➢ Le PR d'une page A dépend du PR des pages T1, …, Tn qui émettent un lien vers A et ne peut donc pas être déterminé sans connaître le PR de ces dernières et de toutes celles qui émettent un lien vers elles et ainsi de suite …

➢ Selon Lawrence Page and Sergey Brin :
- Le PR peut être calculé en utilisant un simple algorithme itératif.
- Le PR correspond au vecteur propre principal de la matrice normalisée des liens du Web.

Exemple de calcul de PageRank

Exemple 1: **2 pages A et B pointant l'une vers l'autre**

Chaque page a un lien sortant donc C(A) = C(B) = 1

➢ Nous ne connaissons pas le PR des deux page, donc il ne faut une valeur de départ: 1 par exemple

PR(A)	= (1 - d) + d(PR(B)/1)
PR(B)	= (1 - d) + d(PR(A)/1)

Soit, avec un facteur d'amortissement de 0.85 :

PR(A)	= 0.15 + 0.85 * 1 = **1**
PR(B)	= 0.15 + 0.85 * 1 = **1**

PR(A) = PR(B) = 1

Exemple de calcul de PageRank

➢ Prenons une autre valeur de départ, cette fois = 0

Première itération

PR(A)	= 0.15 + 0.85 * 0 = **0.15**
PR(B)	= 0.15 + 0.85 * 0.15 = **0.2775**

Deuxième itération

PR(A)	= 0.15 + 0.85 * 0.2775 = **0.385875**
PR(B)	= 0.15 + 0.85 * 0.385875 = **0.47799375**

Troisième itération

PR(A)	= 0.15 + 0.85 * 0.47799375 = **0.5562946875**
PR(B)	= 0.15 + 0.85 * 0.5562946875 = **0.622850484375**

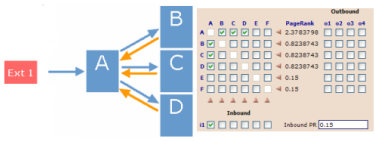
➢ On remarque bien que les valeurs de PR augmente à chaque itération

➢ Les valeurs vont continuer à augmenter jusqu'à ils seront proche ou égale à 1

Exemple de calcul de PageRank

Exemple 5: Effet des liens entrant

Exemple 5-1 : Structure hiérarchique avec lien entrant



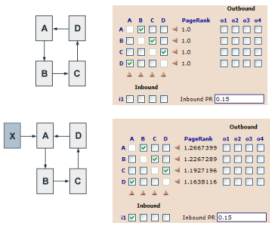
Un webmaster d'un site extérieur à émis un lien de sa page vers la page A.

➔ ceci va augmenter automatiquement le PageRank de A

Exemple de calcul de PageRank

Exemple 5: Effet des liens entrant

Exemple 5-2 : Le cas d'une structure circulaire



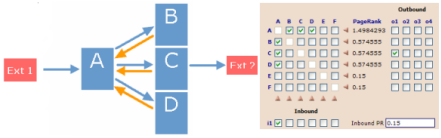
PR(A) = PR(B) = PR(C) =PR(D) =1

PR(A) > PR(B) > PR(C) > PR(D)

Exemple de calcul de PageRank

Exemple 6: Effet des liens sortant

Exemple 6-1 : Structure hiérarchique avec lien entrant et lien sortant

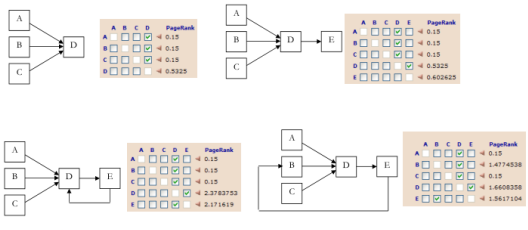


➢ Le simple fait d'émettre un lien vers une page extérieure au départ de la page C fait chuter le PR de manière conséquente sur toutes les pages du site.

➢ La seule cause de la chute de PR vient du fait que la page C qui redistribuait dans l'exemple 5-1 tout son PR à la page A, ne lui renvoie plus que la moitié de celui-ci.

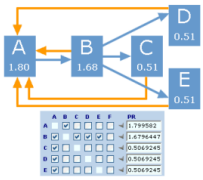
Exemple de calcul de PageRank

Exemple 6-2



Exemple de calcul de PageRank

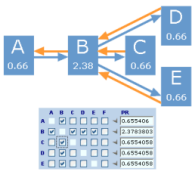
Exemple 7 : Un plan de site



➢ Dans cet exemple, la page d'accueil (A) émet un lien vers le plan du site (B). Celui-ci, en plus du lien retour vers la page d'accueil émet un lien vers chacune des pages du site (C, D et E). Pour éviter les "fuites" de PageRank, celles-ci émettent un lien en retour vers la page d'accueil.

Exemple de calcul de PageRank

Exemple 8 : Un plan de site lié différemment



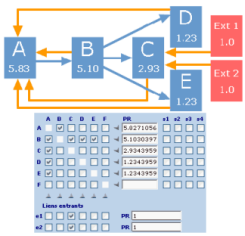
❑ La seule différence avec l'exemple précédent vient du lien en retour des pages internes du site. Plutôt que d'émettre un lien vers la page d'accueil, elles émettent le lien en retour vers le plan du site (B).

❑ Ceci a pour effet de favoriser le PageRank de la page B, ce qui peut être souhaité si elle est riche en mots clés.

❑ Par contre, la page d'accueil A voit son PageRank diminuer fortement comme elle n'a plus qu'un lien entrant.

Exemple de calcul de PageRank

Exemple 9 : Un plan de site avec des liens entrants sur une page interne



- Le contenu intéressant de la page C lui vaut des liens spontanés de la part de deux sites extérieurs.
- On suppose le PR des pages entrantes = 1
  La page C voit tout naturellement sont PR augmenté.
- Grace au chainage interne la page d'accueil et le plan du site font toutes deux un bond vers le haut.
- Dans une moindre mesure les pages D et E profitent du gain en PR du plan du site B.

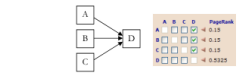
PageRank - Discussion

❑ Améliorer la pertinence des résultats constitue un objectif technologique majeur pour la plupart des moteurs de recherche. Le problème, c'est que cette « pertinence » est une notion subjective, qu'il est donc difficile de manier avec des algorithmes purement mathématiques.

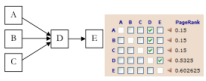
❑ La valeur du PageRank part du principe intuitif que l'importance d'une page dépend du nombre de liens entrants pointant vers cette page, mais aussi de l'importance des pages d'où partent ces liens.

PageRank - Discussion

- **Problème : propagation du PageRank**



Plusieurs pages pointent vers la page D
La page D à un bon contenu
Le PR de D va augmenter



Il est toujours possible de voir un lien sortant d'une page avec un bon contenu vers une page non importante (ex. D➔E). Conséquence, cette page (E) va avoir son PR augmenté alors qu'elle ne contient aucune information pertinente.

PageRank - Discussion

❑ Le calcul du PageRank pour un index de la taille de Google nécessite du temps et une puissance de calcul phénoménale.

➢ Il est possible que ce calcul est géré avec une architecture basée sur un grand nombre de machines calculant chacune les valeurs pour une petite zone de la colossale matrice du Web.

➢ L'algorithme itératif de calcul du PageRank converge au bout d'un certain nombre d'itérations vers une valeur fixe.

❑ La convergence de l'algorithme n'est pas prouvée théoriquement. Cependant, les expérimentations montrent que la convergence est assurée en pratique.

Forage de contenu web

- Ici le contenu de la page web est prise en considération
- Un classement des pages web est effectué en fonction de leurs scores de pertinence par rapport à la requête.

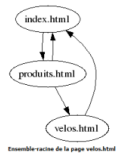
Forage de contenu web

- Étant donné D l'ensemble des contenus des pages webs.
- La sélection des documents peut se faire suivant une représentation booléenne:
  - Ici la requête de l'utilisateur est prise comme une suite logique.
  - Exemple: $((xANDy)AND(NOTz))$ signifie les documents contenants les termes x et y et non z
  - le système récupère chaque document qui rend la requête logiquement vraie
  - Ici l'appariement est exact, on ne tient pas compte de la pertinence d'un mot dans le document recherché.

Exemple 1

➢ Supposons que, à la suite d'une recherche par mots clés : « rabais postal sur vélo », on obtient comme pages correspondantes seulement la page « velos.html ».

➢ Il faut alors ajouter toutes les pages qui font des liens vers « velos.html » et les pages qui reçoivent des liens en provenance de « velos.html ».



Exemple 1

Définissons alors la matrice A par Aij = 1 si j → i et par Aij = 0 sinon.

La première colonne correspond à la page « index.html », la seconde colonne correspond à la page « produits.html », et la troisième colonne correspond à la page « velos.html ».

$$A = \begin{bmatrix} 0 & 1 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix}.$$

On fixe h(1) = h(2) = h(3) = 1

$$a(p) = \sum_{q \to p} h(q) = A \begin{bmatrix} h(q_1) \\ h(q_2) \\ h(q_3) \end{bmatrix} = \begin{bmatrix} 0 & 1 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} = \begin{bmatrix} 2 \\ 1 \\ 1 \end{bmatrix}$$

Exemple 1

Pour calculer h à partir de a il suffit de prendre la transposée de A

$$\begin{bmatrix} h(p_1) & h(p_2) & h(p_3) \end{bmatrix} = A^T \begin{bmatrix} 2 \\ 1 \\ 1 \end{bmatrix} = \begin{bmatrix} 0 & 1 & 0 \\ 1 & 0 & 1 \\ 1 & 0 & 0 \end{bmatrix} \begin{bmatrix} 2 \\ 1 \\ 1 \end{bmatrix} = \begin{bmatrix} 1 \\ 3 \\ 2 \end{bmatrix}$$

Normalisation

$$a = \begin{bmatrix} 2 & 1 & 1 \end{bmatrix} \Big/ \sqrt{6} \quad , \quad h = \begin{bmatrix} 1 & 3 & 2 \end{bmatrix} \Big/ \sqrt{14}$$

On répète ensuite une seconde fois.

Estimation de a à partir de h

$$a = \begin{bmatrix} 0 & 1 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} 1/\sqrt{14} \\ 3/\sqrt{14} \\ 2/\sqrt{14} \end{bmatrix} = \begin{bmatrix} 5/\sqrt{14} \\ 1/\sqrt{14} \\ 3/\sqrt{14} \end{bmatrix}$$

Estimation de h à partir de a

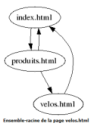
$$h = \begin{bmatrix} 0 & 1 & 0 \\ 1 & 0 & 1 \\ 1 & 0 & 0 \end{bmatrix} \begin{bmatrix} 5/\sqrt{14} \\ 1/\sqrt{14} \\ 3/\sqrt{14} \end{bmatrix} = \begin{bmatrix} 1/\sqrt{14} \\ 8/\sqrt{14} \\ 5/\sqrt{14} \end{bmatrix}$$

Exemple 1

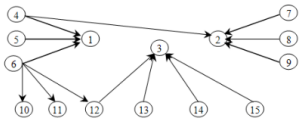
a = [0.85, 0, 0.53]

h = [0, 0.85, 0.53]

L'algorithme HITS recommande de visiter d'abord la page « index.html » et ensuite, seulement, la page « velos.html ». La page « produits.html » ne serait pas recommandée.



Exemple 2



Cette représentation graphique suggère que :

➢ Les nœuds qui doivent recevoir un grand « authority scores » sont : 1, 2 et 3.

➢ Les nœuds 1, 2 et 3 doivent avoir un « authority score » très élevé par rapport aux nœuds 10, 11 et 12.

Exemple 2

Avec HITS on trouve :

Les valeurs de « authority score » pour les nœuds 1, 2, 10, 11, 12 sont très élevés alors que le nœud 3 a un « authority score » très proche de zéro.

Cette performance de HITS peut être expliquée par le fait que

✓ Le « hub score » du nœud 6 augmente puisque il pointe vers plusieurs nœuds. Parmi eux il y a le nœud 1 qui a un « authority score » élevé.

✓ Le fait que le nœud 6 pointe vers 1 contribue à augmenter sont « hub score » par conséquent « authority score » des nœuds 10, 11 et 12 va augmenter.

✓ Les nœuds qui pointent vers 3 ont relativement un faible « hub score » en comparaison avec les autres nœuds qui pointent vers 1 et 2.
Résultat, le nœud 3 va avoir un faible « authority score ».