Web mining

Etienne G. Tajeuna

November 17, 2022

Plan

Introduction

- Le Web pourrait être vu comme un repertoire de données,

Introduction

- Le Web pourrait être vu comme un repertoire de données,
- Il est caractérisé par :

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.

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'hétérogéneité des données : Les données présentes sont de types et natures diverses.

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'hétérogéneité 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.

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'hétérogéneité 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.

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'hétérogéneité 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.

Introduction

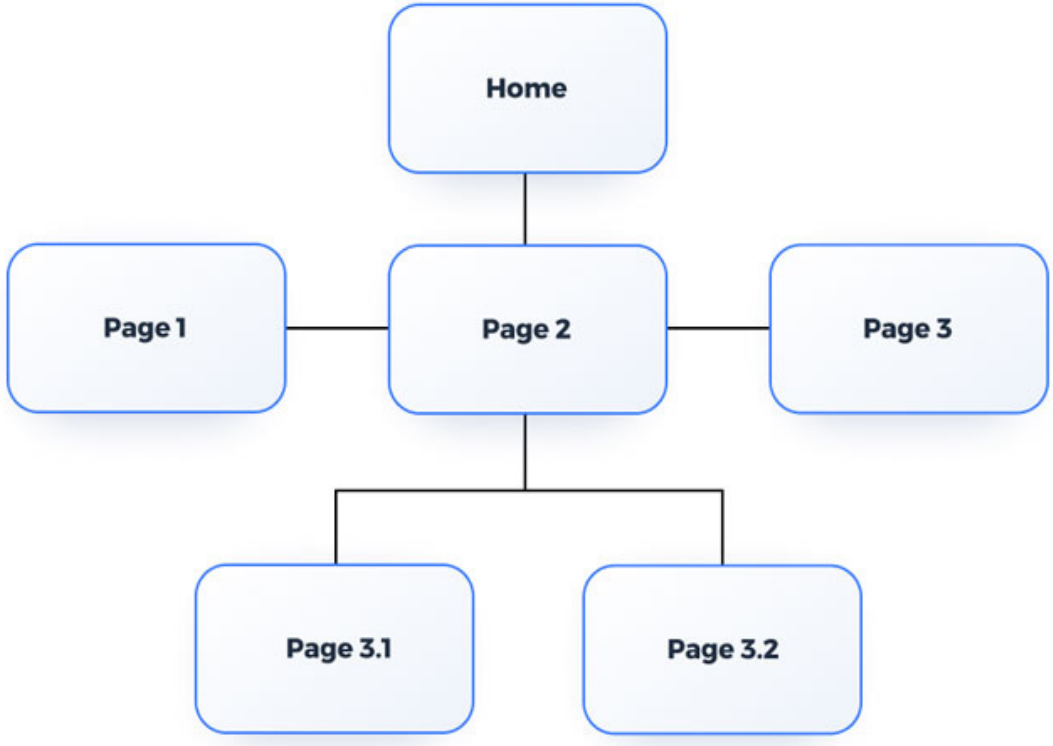
- Le Web pourrait être vu comme un repertoire de données,
- Il est caractérisé par :
  - son grand volume de données : en perpétuelle augmentation, la quantité de données disponibles est immense.
  - l'hétérogéneité des données : Les données présentes sont de types et natures diverses.
  - l'interconnexion 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

- 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.

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

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.

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.

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.

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.

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.

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

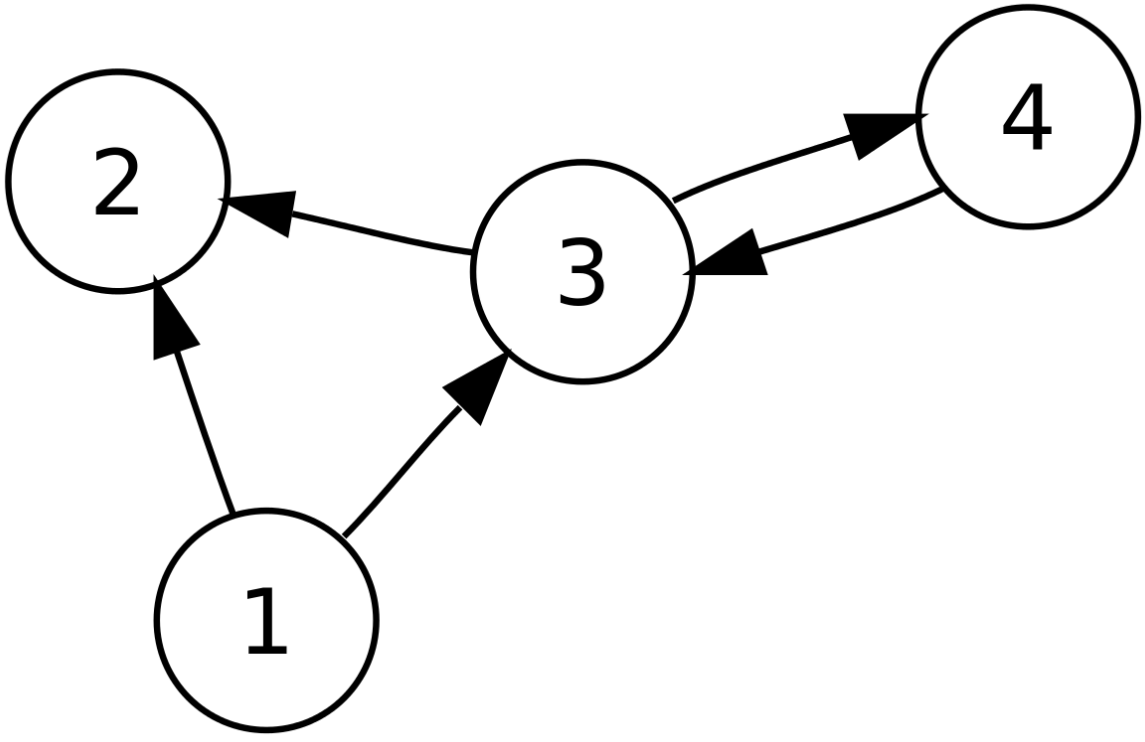
Forage de structure web

- 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.

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

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

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.

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

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:

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:

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

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.

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.

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

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:

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.

Forage de structure web – PageRank

- Se base sur le critère de prestige pour faire la classification des pages webs.

Forage de structure web – PageRank

- Se base sur le critère de prestige pour faire la classification des pages webs.
- À partir d'une matrice stochastique décrivant la relation entre les différentes pages, on effectue le calcul de classement des pages.

Forage de structure web – PageRank

- Se base sur le critère de prestige pour faire la classification des pages webs.
- À partir d'une matrice stochastique décrivant la relation entre les différentes pages, on effectue le calcul de classement des pages.
- Les étapes sont les suivantes:

Forage de structure web – PageRank

- Se base sur le critère de prestige pour faire la classification des pages webs.
- À partir d'une matrice stochastique décrivant la relation entre les différentes pages, on effectue le calcul de classement des pages.
- Les étapes sont les suivantes:
  1. Construire la matrice d'adjacence A des pages webs liées

Forage de structure web – PageRank

- Se base sur le critère de prestige pour faire la classification des pages webs.
- À partir d'une matrice stochastique décrivant la relation entre les différentes pages, on effectue le calcul de classement des pages.
- Les étapes sont les suivantes:
  1. Construire la matrice d'adjacence A des pages webs liées
  2. Transoformer cette matrice sous forme stochastique A^*

Forage de structure web – PageRank

- Se base sur le critère de prestige pour faire la classification des pages webs.
- À partir d'une matrice stochastique décrivant la relation entre les différentes pages, on effectue le calcul de classement des pages.
- Les étapes sont les suivantes:
  1. Construire la matrice d'adjacence A des pages webs liées
  2. Transoformer cette matrice sous forme stochastique A^*
  3. Résoudre l'équation $A^* \cdot X = X$ où $X = (x_1, x_2, \ldots, x_n)$ est le vecteur de probabilité estimant la chance de parcourir les différentes pages webs.

Forage de structure web – PageRank

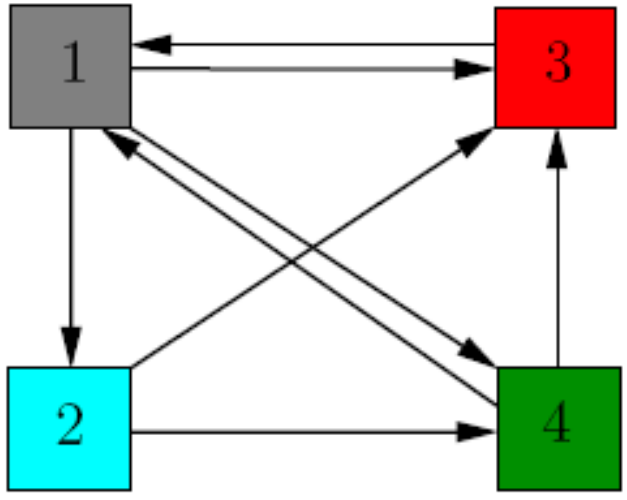
- Il existe plusieurs méthodes pour résoudre l'équation,

Forage de structure web – PageRank

- Il existe plusieurs méthodes pour résoudre l'équation,
- Méthode iterative:

Forage de structure web – PageRank

- Il existe plusieurs méthodes pour résoudre l'équation,
- Méthode iterative:



Forage de structure web – PageRank

- On détermine la matrice d'adjacence A

Forage de structure web – PageRank

- On détermine la matrice d'adjacence A
- Puis on la transforme en matrice stochastique

Forage de structure web – PageRank

- On détermine la matrice d'adjacence A
- Puis on la transforme en matrice stochastique

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

Forage de structure web – PageRank

- Marche aléatoire jusqu'à convergence:

Forage de structure web – PageRank

- Marche aléatoire jusqu'à convergence:

$$v = \begin{pmatrix} 0.25 \\ 0.25 \\ 0.25 \\ 0.25 \end{pmatrix}, \quad Av = \begin{pmatrix} 0.37 \\ 0.08 \\ 0.33 \\ 0.20 \end{pmatrix}, \quad A^2 v = A\,(Av) = A \begin{pmatrix} 0.37 \\ 0.08 \\ 0.33 \\ 0.20 \end{pmatrix} = \begin{pmatrix} 0.43 \\ 0.12 \\ 0.27 \\ 0.16 \end{pmatrix}$$

$$A^3 v = \begin{pmatrix} 0.35 \\ 0.14 \\ 0.29 \\ 0.20 \end{pmatrix}, \quad A^4 v = \begin{pmatrix} 0.39 \\ 0.11 \\ 0.29 \\ 0.19 \end{pmatrix}, \quad A^5 v = \begin{pmatrix} 0.39 \\ 0.13 \\ 0.28 \\ 0.19 \end{pmatrix}$$

$$A^6 v = \begin{pmatrix} 0.38 \\ 0.13 \\ 0.29 \\ 0.19 \end{pmatrix}, \quad A^7 v = \begin{pmatrix} 0.38 \\ 0.12 \\ 0.29 \\ 0.19 \end{pmatrix}, \quad A^8 v = \begin{pmatrix} 0.38 \\ 0.12 \\ 0.29 \\ 0.19 \end{pmatrix}$$

Forage de structure web – PageRank

- On pourrait aussi résoudre le problème en résolvant l'équation $A^* \cdot X = X$.

Forage de structure web – PageRank

- On pourrait aussi résoudre le problème en résolvant l'équation $A^* \cdot X = X$.
- Dans notre exemple, on aurait le système d'équation

Forage de structure web – PageRank

- On pourrait aussi résoudre le problème en résolvant l'équation $A^* \cdot X = X$.
- Dans notre exemple, on aurait le système d'équation

$$\begin{cases} x_1 = 1 \cdot x_3 + \frac{1}{2} \cdot x_4 \\ x_2 = \frac{1}{3} \cdot x_1 \\ x_3 = \frac{1}{3} \cdot x_1 + \frac{1}{2} \cdot x_2 + \frac{1}{2} \cdot x_4 \\ x_4 = \frac{1}{3} \cdot x_1 + \frac{1}{2} \cdot x_2 \end{cases}$$

Forage de structure web – PageRank, Calibrage

- Problème avec la matrice stochastique

- Problème avec la matrice stochastique

$$\begin{bmatrix} 0 & 0 & 1 & 0 \\ \frac{1}{3} & 0 & \frac{1}{3} & \frac{1}{3} \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{bmatrix}.$$

Forage de structure web – PageRank, Calibrage

- Problème avec la matrice stochastique

$$\begin{bmatrix} 0 & 0 & 1 & 0 \\ \frac{1}{3} & 0 & \frac{1}{3} & \frac{1}{3} \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{bmatrix}.$$

- Lorsqu'on a des valeurs nulles, on pourrait se retrouver dans une difficulté de classement

Forage de structure web – PageRank, Calibrage

- Problème avec la matrice stochastique

$$\begin{bmatrix} 0 & 0 & 1 & 0 \\ \frac{1}{3} & 0 & \frac{1}{3} & \frac{1}{3} \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{bmatrix}.$$

- Lorsqu'on a des valeurs nulles, on pourrait se retrouver dans une difficulté de classement

$$\begin{bmatrix} \frac{1}{4} & \frac{1}{4} & \frac{1}{4} & \frac{1}{4} \end{bmatrix} \begin{bmatrix} 0 & 0 & 1 & 0 \\ \frac{1}{3} & 0 & \frac{1}{3} & \frac{1}{3} \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{bmatrix}^{10000} = \begin{bmatrix} 0.5 & 0 & 0.5 & 0 \end{bmatrix}$$

Forage de structure web – PageRank, Calibrage

- Afin de limiter cet effet déplorable,

Forage de structure web – PageRank, Calibrage

- Afin de limiter cet effet déplorable,
- Il faudrait s'assurer que la matrice de transition ait toujours des valeurs non-nulles,

Forage de structure web – PageRank, Calibrage

- Afin de limiter cet effet déplorable,
- Il faudrait s'assurer que la matrice de transition ait toujours des valeurs non-nulles,
- Pour ce faire on va définir un coefficient de calibrage $q \in [0, 1]$ et ainsi calculer le facteur de *PageRank* d'une page u ($\mathcal{P}(u)$) comme suit:

Forage de structure web – PageRank, Calibrage

- Afin de limiter cet effet déplorable,
- Il faudrait s'assurer que la matrice de transition ait toujours des valeurs non-nulles,
- Pour ce faire on va définir un coefficient de calibrage $q \in [0, 1]$ et ainsi calculer le facteur de *PageRank* d'une page u ($\mathcal{P}(u)$) comme suit:

$$\mathcal{P}(u) = \frac{1-q}{N} + \sum_{v \in In_u} \frac{qP(v)}{|Out_v|}$$

Forage de structure web – PageRank, Calibrage

- Afin de limiter cet effet déplorable,
- Il faudrait s'assurer que la matrice de transition ait toujours des valeurs non-nulles,
- Pour ce faire on va définir un coefficient de calibrage $q \in [0, 1]$ et ainsi calculer le facteur de *PageRank* d'une page u ($\mathcal{P}(u)$) comme suit:

$$\mathcal{P}(u) = \frac{1-q}{N} + \sum_{v \in In_u} \frac{qP(v)}{|Out_v|}$$

Avec N le nombre de pages webs, $In_u = \{v/(u, v) \in \boldsymbol{E}\}$, $Out_v = \{w/(v, w) \in \boldsymbol{E}\}$, $P()$ le prestige d'une page.

Forage de structure web – PageRank, Calibrage

$$\begin{bmatrix} 0 & \frac{1}{3} & \frac{1}{3} & \frac{1}{3} & 0 & 0 \\ \frac{1}{3} & 0 & 0 & 0 & \frac{1}{3} & \frac{1}{3} \\ 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

Forage de structure web – PageRank, Calibrage

$$\begin{bmatrix} 0 & \frac{1}{3} & \frac{1}{3} & \frac{1}{3} & 0 & 0 \\ \frac{1}{3} & 0 & 0 & 0 & \frac{1}{3} & \frac{1}{3} \\ 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

$$\begin{bmatrix} \frac{1-q}{6} & \frac{q}{3}+\frac{1-q}{6} & \frac{q}{3}+\frac{1-q}{6} & \frac{q}{3}+\frac{1-q}{6} & \frac{1-q}{6} & \frac{1-q}{6} \\ \frac{q}{3}+\frac{1-q}{6} & \frac{1-q}{6} & \frac{1-q}{6} & \frac{1-q}{6} & \frac{q}{3}+\frac{1-q}{6} & \frac{q}{3}+\frac{1-q}{6} \\ q+\frac{1-q}{6} & \frac{1-q}{6} & \frac{1-q}{6} & \frac{1-q}{6} & \frac{1-q}{6} & \frac{1-q}{6} \\ q+\frac{1-q}{6} & \frac{1-q}{6} & \frac{1-q}{6} & \frac{1-q}{6} & \frac{1-q}{6} & \frac{1-q}{6} \\ q+\frac{1-q}{6} & \frac{1-q}{6} & \frac{1-q}{6} & \frac{1-q}{6} & \frac{1-q}{6} & \frac{1-q}{6} \\ q+\frac{1-q}{6} & \frac{1-q}{6} & \frac{1-q}{6} & \frac{1-q}{6} & \frac{1-q}{6} & \frac{1-q}{6} \end{bmatrix}$$

Forage de contenu web

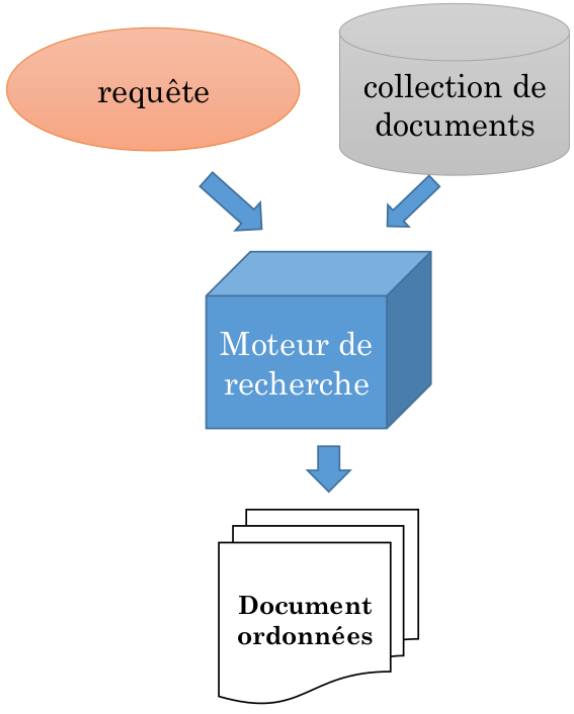
- Ici le contenu de la page web est prise en considération

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

- 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.

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:

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.

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

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

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.

Forage de contenu web

- La sélection des documents peut se faire suivant une représentation vectorielle.

Forage de contenu web

- La sélection des documents peut se faire suivant une représentation vectorielle.
  - Étant donnée la famille de termes $\boldsymbol{W} = \{W_1, W_2, \ldots W_n\}$ que l'on pourrait rencontrer sur une page web P,

Forage de contenu web

- La sélection des documents peut se faire suivant une représentation vectorielle.
  - Étant donnée la famille de termes $\boldsymbol{W} = \{W_1, W_2, \ldots W_n\}$ que l'on pourrait rencontrer sur une page web P,
  - On pourrait représenter la page web P par le vecteur $P = (\omega_1, \omega_2, \ldots, \omega_n)$ où ω_i, $1 \leq i \leq n$, est le nombre d'occurrence du mot W_i dans le contenu de la page P.

Forage de contenu web

- La sélection des documents peut se faire suivant une représentation vectorielle.
  - Étant donnée la famille de termes $\boldsymbol{W} = \{W_1, W_2, \ldots W_n\}$ que l'on pourrait rencontrer sur une page web P,
  - On pourrait représenter la page web P par le vecteur $P = (\omega_1, \omega_2, \ldots, \omega_n)$ où ω_i, $1 \leq i \leq n$, est le nombre d'occurrence du mot W_i dans le contenu de la page P.
  - La requête de l'utilisateur est aussi représenté sous le même format

Forage de contenu web

- La sélection des documents peut se faire suivant une représentation vectorielle.
  - Étant donnée la famille de termes $\boldsymbol{W} = \{W_1, W_2, \ldots W_n\}$ que l'on pourrait rencontrer sur une page web P,
  - On pourrait représenter la page web P par le vecteur $P = (\omega_1, \omega_2, \ldots, \omega_n)$ où ω_i, $1 \leq i \leq n$, est le nombre d'occurrence du mot W_i dans le contenu de la page P.
  - La requête de l'utilisateur est aussi représenté sous le même format
  - En fonction d'une mesure de similarité on peut donc faire le classement des pages en fonction de celles qui sont les plus similaires de la requête.