Modélisations mathématiques

BUT3 Info Fontainebleau

2023

Introduction

On considère un ensemble de pages web traitant d'un sujet donné (par exemple, de la modélisation mathématique). On suppose que ses administrateurs peuvent y intégrer des hyperliens vers d'autres pages, s'ils les trouvent intéressantes. Pour découvrir ces pages, ils utilisent un moteur de recherche.

Objectif

Le but de ce projet est de modéliser la création de liens entre les pages web sous l'influence de moteurs de recherche.

Model

Graphe

On voit le web comme un graphe orienté. Chaque page est un nœud du graphe, chaque lien entre les pages est un arc entre deux nœuds.

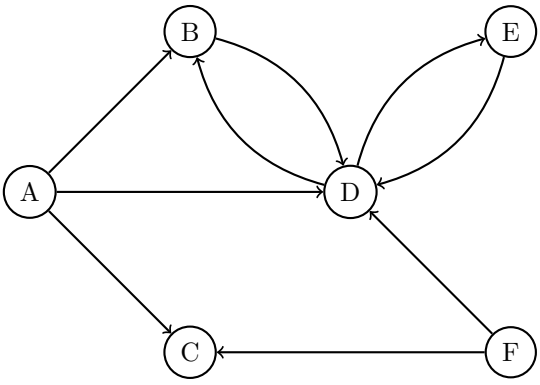

Figure 1: Pages et hyperliens

Pertinence

On suppose que chaque page x a une valeur de pertinence $p(x) \in [0; 1]$:

x	A	B	C	D	E	F
$p(x)$	0.1	0.4	0.2	0.8	0.3	0.1

Ces valeurs peuvent représenter, par exemple, la probabilité qu'un visiteur trouve sur cette page l'information recherchée. Elles dépendent du contenu de la page et sont supposées fixes.

Score

Les moteurs de recherche attribuent à chaque page x un score $s(x)$ utilisé pour trier les résultats. On suppose qu'ils n'analysent pas le contenu et n'ont donc pas l'accès à $p(x)$, mais calculent les scores en fonction de liens entre les pages, par exemple :

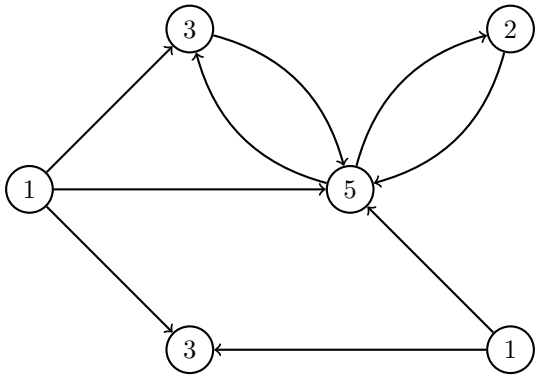

Figure 2: Score = nombre des liens entrants plus un

On normalise souvent les scores de façon que leur somme soit égale à 1, mais ce n'est pas une obligation.

Dynamique

Un administrateur qui veut rajouter un lien à sa page web, consulte un moteur de recherche. Il choisit généralement une page parmi les mieux classées et rajoute son lien s'il la trouve intéressante.

Les moteurs de recherche recalculent alors les scores, et le processus continue.

Au bout d'un moment chaque page contient suffisamment de liens, et le processus s'arrête.

Projet

Vous devez écrire un programme (en Python ?) qui simule et visualise le processus de création de liens.

Plusieurs points sont laissés à votre appréciation. Vous pouvez vous inspirer de votre propre comportement sur internet.

Il est possible de travailler en binôme.

Initialisation

Vous devez générer un graphe de n nœuds (pages), en attribuant à chaque page x une valeur de pertinence $p(x) \in [0; 1]$. Cette valeur peut être choisie

- uniformément entre 0 et 1 (plus simple)
- ou en suivant une loi plus réaliste (à réfléchir)

Aucun lien n'existe entre les pages à l'instant $t = 0$ (ou chaque page se réfère à elle-même). Le moteur de recherche attribue donc à chaque page x le même score $s(x) = 1/n$.

Rajout d'un lien

A l'instant $t+1$, on prend une page x au hasard. Son administrateur va consulter le moteur de recherche et choisir une page y à visiter, pour éventuellement rajouter un lien de x vers y. Le choix de y doit être fait en fonction du score actuel $s(y)$, par exemple

- avec la probabilité $s(y)$ (en supposant les scores normalisés)
- ou en privilégiant les pages les mieux classées d'une façon plus réaliste (à réfléchir)

En visitant la page y choisie, l'administrateur de x décide de créer (ou pas) un lien vers y en fonction de l'intérêt de y,

- avec la probabilité $p(y)$
- peut-être en fonction des liens actuels (à réfléchir). Par exemple, si x contient déjà un lien vers une page plus pertinente que y, il décide de ne pas rajouter un lien vers y. Il peut aussi décider de remplacer un lien moins pertinent par y, etc.

Calcul des scores

Quand des liens sont modifiés, le moteur de recherche recalcule les scores. Algorithmes possibles :

- nombre de liens entrants (plus simple)

- PageRank (plus fin/fun)
- une autre fonction de votre choix (plus valorisant)

Itérations

On répète les étapes précédentes plusieurs fois, afin que l'ensemble des liens se "stabilise". Les conditions d'arrêt possibles :

- chaque page contient suffisamment de liens sortants (par exemple, $\ln n$).
- ou la pertinence totale des ces liens est assez grande (à réfléchir)
- aucun nouveau lien n'est créé pendant plusieurs itérations
- une autre condition de votre choix

Visualisation

A l'issue des simulations, vous devez afficher les pages avec leurs liens :

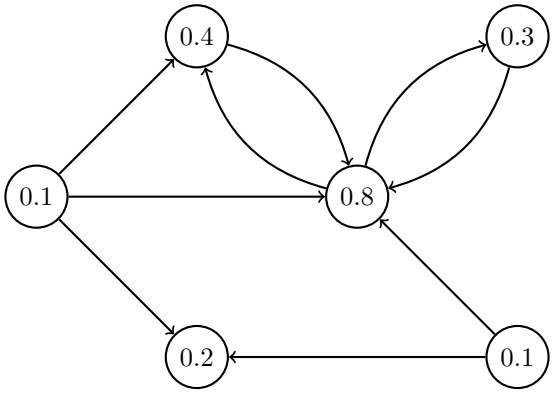

Figure 3: Pertinences et liens

Analyse

On s'attend à ce que les pages les plus pertinentes soient les plus référencées et donc mieux classées par les moteurs de recherche.

- Est-ce toujours le cas ? Qu'en est-il en absence de moteurs de recherche ?
- Si un moteur de recherche attribue un grand score à la page (même nulle) d'un sponsor, va-t-elle finir par recevoir un grand nombre de liens ?
- Autres remarques ?