04. L'algorithme de PageRank

Durée
1h30
modulable

Thématique

WEB

Activité
Branchée

Description de l'activité

Dans cette activité, on présente les principes des moteurs de recherche, en particulier l'algorithme PageRank qui permet de mesurer la popularité d'un site ou d'une page web

Objectifs pédagogiques ou compétences

Objectifs généraux	Objectifs intermédiaires	Compétences
Compréhension des algorithmes	- Connaître l'algorithme de PageRank - Connaître le fonctionnement d'un moteur de recherche - Représenter les liens entre des sites par un graphe - Se représenter l'usage des algorithmes dans le quotidien et leur impact	- Analyser un graphe - Se représenter les inconvénients d'un graphe - Effectuer des recherches sur Internet et résumer

Matériel et outils

- Dés : 1 par groupe de 4 élèves
- Ordinateurs et connexion Internet
- Fiche Activité élève à imprimer
- Si besoin : Graphs à imprimer ou projeter

Tags

#moteur de recherche #PageRank #graphe

Déroulé de l'activité

Introduction : *(~15 minutes)*

- **Présenter les objectifs de la séance (contenu théorique et productions attendues)** *(2-3 minutes)*
- **Rappels – Les moteurs de recherche :** *(~10 minutes)*

Pour lancer la thématique, on peut lancer une première discussion sur les moteurs de recherche. L'objectif est de vérifier que les élèves ont assimilé les concepts suivants :

  - Fonction d'un moteur de recherche
  - Différence entre moteur de recherche et navigateur
  - Automatisation des moteurs de recherche

Dans la fiche activité, nous avons intégré une fiche « quiz », que les jeunes peuvent remplir en binôme ou petits groupes, ou bien qui peut guider une discussion au sein de la classe.

Activités : *(~45 minutes modulables)*

- **Activité 1 : Le principe du surfeur aléatoire**

Pour comprendre l'idée du surfeur aléatoire, les élèves utilisent des dés et des modèles de graphes simples. Le professeur passe dans les rangs pour vérifier l'avancée des travaux et répondre aux différentes questions.

Le professeur donne un premier modèle de graphe, donne des explications et les élèves répondent aux questions. Une mise en commun des résultats et des échanges avec le professeur a lieu à l'issu de cette première étape. La deuxième étape reprend le même principe avec un autre graphe. La troisième étape introduit des graphes qui peuvent poser des problèmes avec la méthode utilisée.

- **Activité 2 : Le principe du surfeur aléatoire**

On propose de modifier l'algorithme utilisé à la suite des problèmes mis en avant dans l'activité 1.

Cette partie est un peu plus difficile. Le professeur donne des explications en faisant un schéma au tableau ou un diaporama. Il peut proposer de faire l'étude sur un seul graphe, le puits ou la poche web. Les élèves utilisent la fonction randint de la bibliothèque random pour les probabilités.

Il peut aussi présenter un programme qui simule cet algorithme sur le graphe étudié.

Conclusion : *(~15 minutes)*

- **Bilan de la séance :** *(5 minutes)*

Pour clôturer la séance, on peut revenir sur les principales difficultés rencontrées pendant l'activité. Éventuellement, il est possible de finir sur un court échange autour de :

- **Les algorithmes :**

- Tous les moteurs n'ont pas les mêmes algorithmes. Le principe est généralement le même mais les critères peuvent varier.
    - Quels sites ont des algorithmes connus pour mettre en avant certains contenus plutôt que d'autres ?
    - Quels sont les critères qui peuvent favoriser / limiter la mise en avant d'un contenu ?

Pistes de réflexion :

- Tiktok, YouTube, Netflix, IG : Les sites de streaming et les RS ont également des algorithmes pour sélectionner le contenu à mettre en avant, en fonction des goûts de l'usager/abonné, mais pas seulement.
- Temps de visionnage, engagement (likes, partages, …), fraîcheur (valorisation des nouveaux contenus), thèmes et contenus inadaptés, taux de clic (CTR) dans les contenus suggérés, …

    - Quels effets négatifs les algorithmes peuvent avoir sur les créateurs de contenu ?

Pistes de réflexion :

- Biais de popularité : mise en avant des créateurs / thématiques / formats les plus populaire, qui laisse peu de place aux nouveaux créateurs / nouveaux types de contenus
- Risque d'uniformisation (cf. biais de popularité)
- Contenu sensationnaliste ou polarisant : certains algorithmes sont conçus pour maximiser l'engagement des utilisateurs en favorisant les contenus sensationnalistes, choquants ou polarisants. Cela peut encourager la création de contenus extrêmes, provocateurs ou controversés simplement pour attirer l'attention
- Effet de « bulle filtrée » : en mettant en avant les contenus qui vont dans notre sens, les algorithmes peuvent créer chez les usagers une vision du monde polarisée et/ou réduite.
- Manipulation des algorithmes : en décidant de certains critères pour limiter / valoriser un certain type de contenus, on peut manipuler l'opinion. La polémique autour des changements opérés par Elon Musk au sein de Twitter peut être un exemple concret.

- **Les professions en lien avec les algorithmes :**

On peut lancer la discussion en parlant des produits / personnes qui dépendent des algorithmes, pour amener la diversité de métiers en lien avec ce domaine.

Côté métiers, on peut notamment mettre en avant les compétences / principales facettes de plusieurs métiers et voir si les élèves se projettent sur un ou plusieurs de ces métiers.

EVALUATION :

Partie 2 : Pour aller plus loin peut être évaluée.

Suggestions sur l'usage des algorithmes :

- **Moteurs de recherche, RS, plateformes de streaming** (vidéos, musique, …).
- **Sites de recommandation :** TripAdvisor pour les voyages, Pandora pour la musique, et, de manière plus globale, toute fonctionnalité basée sur la recherche personnalisée.
- **L'e-commerce :** Les sites de commerce électronique comme Amazon, eBay et Alibaba utilisent des algorithmes pour recommander des produits aux utilisateurs en fonction de leurs achats antérieurs, de leurs recherches et de leur comportement de navigation. De même, ces sites (sans oublier les entreprises qui ne sont pas directement liées à l'e-commerce) nécessitent de savoir utiliser les algorithmes des moteurs de recherche pour être bien référencés, ou encore ceux des RS pour être visibles et gagner en clients.
- **Des domaines du quotidien :** la médecine peut utiliser des algorithmes pour la détection précoce des maladies ou pour l'aide à la décision, et la finance les utilise pour prévoir des tendances ou analyser le marché.
- **Les objets connectés et autonomes** (assistants vocaux, voitures autonomes, …) utilisent également les algorithmes pour aider à automatiser des tâches complexes, prendre des décisions basées sur des données et fournir des recommandations personnalisées.

Suggestions sur les métiers :

- **Data Scientist :** Les data Scientist analysent les données, développent des modèles statistiques et construisent des algorithmes pour extraire des informations et des recommandations à partir des données collectées.
- **Ingénieur en algorithmes de recommandation :** Ces ingénieurs conçoivent, développent et optimisent des algorithmes de recommandation utilisés dans divers domaines tels que le commerce électronique, les médias sociaux et les services de streaming.
- **Analyste SEO (Search Engine Optimization) :** Les analystes SEO travaillent sur l'optimisation du référencement des sites Web. Ils utilisent des algorithmes de moteurs de recherche pour améliorer la visibilité et le classement des sites dans les résultats de recherche.
- **Développeur d'algorithmes de recherche :** Ces développeurs conçoivent et mettent en œuvre des algorithmes utilisés dans les moteurs de recherche pour trier, classer et fournir des résultats pertinents aux utilisateurs.
- **Expert en marketing numérique :** Les experts en marketing numérique utilisent des algorithmes de recommandation et de référencement pour optimiser les campagnes publicitaires en ligne, les stratégies de contenu et les efforts de marketing afin d'atteindre un public cible spécifique.
- **Analyste de données marketing :** Ces analystes utilisent des algorithmes pour analyser les données marketing, mesurer l'efficacité des campagnes publicitaires, identifier les tendances et les comportements des consommateurs, et fournir des recommandations pour améliorer les performances marketing.
- **Spécialiste en optimisation des taux de conversion (CRO) :** Ces spécialistes utilisent des algorithmes et des techniques d'analyse pour optimiser les sites Web, les pages de destination et les parcours des utilisateurs afin d'augmenter les taux de conversion et améliorer l'expérience utilisateur.
- **Consultant en référencement :** Les consultants en référencement travaillent avec des entreprises pour optimiser leur présence en ligne, améliorer leur visibilité dans les moteurs de recherche et utiliser des algorithmes pour augmenter leur trafic organique.
- **Architecte de systèmes de recommandation :** Ces professionnels conçoivent et développent des systèmes de recommandation personnalisés pour des applications spécifiques, en utilisant des algorithmes et des techniques d'apprentissage automatique pour fournir des recommandations pertinentes aux utilisateurs.
- **Analyste de données de streaming :** Les analystes de données de streaming utilisent des algorithmes pour analyser les habitudes de visionnage des utilisateurs, les modèles de consommation de contenu et les données d'engagement pour fournir des recommandations personnalisées et améliorer l'expérience utilisateur sur les plateformes de streaming.

URL, protocole HTTP et modèle client-serveur

Fiche activité - Correction

Introduction - Faisons le point … Répondez aux questions suivantes :

- **1. À quoi sert un moteur de recherche ?**

  - ☐ À naviguer sur le web de manière totalement sécurisée.
  - ☐ Il permet de trouver et d'installer des programmes ou applications sur notre ordinateur
  - ☐ Il sert à retrouver une ressource (image, document, site web, vidéo, etc.)

- **2. Donnez quelques exemples de moteurs de recherche connus.**

Google, Qwant, Safari, DuckDuckGo Go, Ecosia, Bing, …

- **3. Connaissez-vous d'autres types de moteurs de recherche ?**

Ceux des …

- Sites et applications de streaming (YouTube, Netflix)
- Réseaux sociaux (Facebook Twitter, Instagram, …)
- Internes à un site (Reddit, …)
- Propre à une messagerie (Gmail, …)
- Propre à son système d'exploitation (recherche de fichiers)

- **4. La recherche sur ces moteurs est …**

  - ☐ Manuelle : elle est réalisée en temps réel par un humain qui, en fonction d'étiquettes décrivant les différents sites, choisit le plus adapté.
  - ☐ Semi-Automatique : des humains vérifient, de manière asynchrone, les propositions des moteurs de recherche pour une question donnée et les améliore.
  - ☐ Automatique : les moteurs indexent régulièrement les sites en fonction de critères, ce qui leur permet de sélectionner

- **5. Voici deux méthodes manuelles pour hiérarchiser la pertinence de sites web sur un domaine :**

- Faire appel à un.e spécialiste (personne ou groupe de personnes) pour classer les pages. Le classement serait certes très pertinent pour ce domaine, mais il faudrait solliciter un grand nombre de spécialistes de différents domaines pour obtenir un algorithme capable de traiter des requêtes dans tous les domaines possibles et imaginables.
- Demander aux internautes eux- mêmes de voter, pour chaque domaine, et de choisir le classement sur la base de ce vote, considérant, que compte-tenu du grand nombre de votants, le classement obtenu serait pertinent.

  - Quels sont les inconvénients de ces méthodes ?

  - La difficulté de gérer le très grand nombre de pages évoquant une requête (plusieurs millions au minimum

- L'actualisation fréquente de ce classement
- La réelle fiabilité des jugements (critères, subjectivité, …)
  - Connaissez-vous un site qui utilise cette méthode ?

Wikipédia

Dans l'optique de développer des protocoles entièrement automatisés et efficaces, **Google** a cherché un modèle de hiérarchisation qui soit :

- **Exploitable dans tous les domaines**
- **Utilisable pour tous les mots-clés**
- **Adaptable à un très grand nombre de données**, même **évolutives**, tout en étant **automatisable** et **suffisamment efficace**.

C'est en répondant à ce cahier des charges que ce nouveau venu a réussi l'exploit, **en quelques mois** et malgré l'émergence de Bing ou encore Qwant, **à obtenir le quasi-monopole de la recherche thématique sur le web**.

- **Test :**

Choisissez 3 ou 4 moteurs de recherche différents et effectuez la même recherche, avec les mêmes mots-clés. Choisissez un élément plutôt précis. Par exemple, plutôt « qu'Intelligence Artificielle », préférez « L'Intelligence Artificielle peut-elle créer de l'art ? ».

Que constatez-vous en analysant la première page ? Prenez en compte plusieurs paramètres :

- Les résultats indexés sont-ils les mêmes ?
- Présence de publicité ?
- Pertinence des sites ? (Sites officiels/reconnus ou non, etc.)
- Pertinence des pages par rapport à la requête

L'idée à la base du modèle de Larry Page et Sergey Brin, fondateurs de Google, revient à attribuer à chaque page un nombre positif entre 0 et 1, appelé score **(en anglais "PageRank")** de la page**, qui caractérisera** la pertinence **de cette page. Ils proposent alors de déterminer ce score à partir des deux règles suivantes :**

- R1 : **Le score attribué à une page doit être** d'autant plus élevé que celle-ci est référencée dans une page faisant autorité **(dont le score est élevé).**
- R2 : Le score attribué à une page doit être d'autant moins élevé que celle-ci est référencée dans une page contenant un grand nombre de références.

Leur idée : utiliser un **surfeur aléatoire.**

1. Le principe du surfeur aléatoire

- **Expliquez le principe du surfeur aléatoire** – Vous pouvez chercher sur Internet.

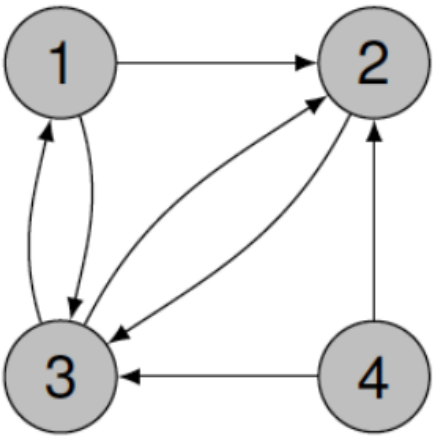
Après avoir dressé la liste (sans classement) de tous les sites traitant de la requête, le surfeur aléatoire en choisit au hasard un. Puis il s'intéresse aux liens hypertexte du site sur lequel il se trouve, vers les autres sites qu'il a listés. Il en choisit alors un au hasard et répète cette opération sans s'arrêter, en comptant pour chacun des sites combien de fois il l'a visité. Les sites sont alors affichés dans l'ordre décroissant de leur nombre de visites.

Ainsi, pour un certain mot-clé entré, il s'intéresse aux sites qui évoquent ce mot-clé, mais également aux liens hypertexte qui permettent de passer d'un site à l'autre.

- **Étape 1 : Quand un internaute effectue une recherche sur un mot-clé, il dresse la liste de tous les sites relatifs à ce mot-clé (mais il ne classe rien à cette étape).**
- **Étape 2 : Il dresse ensuite la liste de tous les liens entre les sites de sa liste.**
- **Étape 3 : Il choisit un site au hasard et parcourt les liens, en comptant combien de fois il visite chaque site.**
- **Étape 4 : À la fin de son parcours, il classe les sites du plus visité au moins visité.**

- **À votre tour de tester !**

Pour illustrer comment un algorithme de calcul peut être mis en place à partir de ces règles, nous allons prendre l'exemple du classement de quatre pages. Le problème de l'attribution du score peut être représenté par un graphe orienté : les quatre pages sont représentées par les quatre sommets d'un graphe, dont les arêtes orientées représentent les références (liens) pouvant exister entre ces différentes pages.



On observe notamment que la page 1 référence les pages 2 et 3, et que la page 4 référence, elle, les pages 2 et 3, mais n'est référencée par aucune de ces trois autres pages.

Règles du jeu : On imagine qu'un internaute a effectué une recherche sur un mot-clé. En réalisant l'étape 1, le surfeur a trouvé quatre sites relatifs à ce mot-clé (que nous noterons 1, 2, 3 et 4). En réalisant l'étape 2, il a découvert que :

- Le site 1 a des liens vers les sites 2 et 3.
- Le site 2 a un lien vers le site 3.
- Le site 3 a des liens vers les sites 1 et 2.
- Le site 4 a des liens vers les sites 2 et 3.

Le graphe au-dessus schématise cette situation. Pour rappel, voici les différentes étapes à suivre :

- **Étape 1** : Quand un internaute effectue une recherche sur un mot-clé, il dresse la liste de tous les sites relatifs à ce mot-clé (mais il ne classe rien à cette étape).
- **Étape 2 :** Il dresse ensuite la liste de tous les liens entre les sites de sa liste.
- **Étape 3 :** Il choisit un site au hasard et parcourt les liens, en comptant combien de fois il visite chaque site.
- **Étape 4 :** À la fin de son parcours, il classe les sites du plus visité au moins visité.

Le graphe inclut les étapes 1 et 2. À l'étape 3, vous devez choisir un point de départ au hasard. Pour cela, lancez le dé. Si vous obtenez 1, vous commencerez sur le site 1. Si vous obtenez 2, vous commencerez sur le site 2, etc. Si vous obtenez 5 ou 6, recommencez jusqu'à obtenir 1, 2, 3 ou 4. **Votre groupe fera le processus 30 fois en alternant les lancers.**

Placez votre pion sur le site de départ et ajoutez une visite au**.x** site**.s** pouvant être visité**.s** dans le tableau ci-dessous

Site n°	**1**	**2**	**3**	**4**
Nombre de visites				

Classez les sites du meilleur au moins bon, selon le surfeur : .....

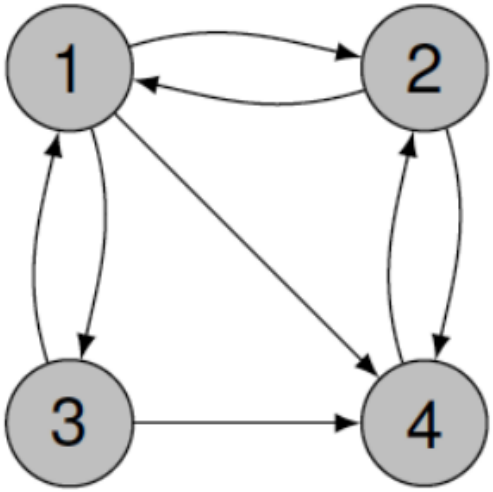
Comparez vos résultats avec ceux des autres groupes. Que constatez-vous ?

Le premier exercice a pour but de faire manipuler le surfeur aléatoire aux élèves au moyen d'un dé. Il permet de réexpliquer la consigne éventuellement à des groupes qui l'auraient mal comprise. En effet, la convergence des fréquences étant assez rapide, avec une trentaine de sauts, les différents groupes devraient avoir le même classement. Un groupe qui aurait un classement divergent est vraisemblablement un groupe qui a mal compris le principe.

D'autre part, sans autres consignes, notamment sur le nombre de surfs aléatoires à exécuter, la comparaison des effectifs n'a que peu de sens (même si l'ordre de classement sera le même). L'idée est de faire émerger que la bonne quantité à comparer est la fréquence de visite de chaque site.

- **Un simple coup de chance ?**

Expérimentez le même protocole, avec le graphe ci-dessous. Il faudra prévoir 50 lancers.



Site n°	1	2	3	4
Nombre de visites				

Classez les sites du meilleur au moins bon, selon le surfeur : .....

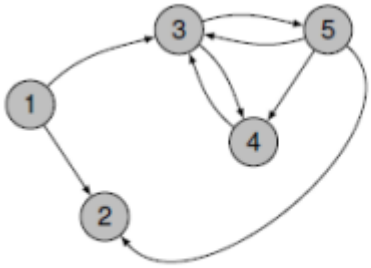
Comparez vos résultats avec ceux des autres groupes. Que constatez-vous ?

Après avoir pris un temps de remédiation pour les groupes qui n'avaient pas tout à fait compris le principe du surf, cet exercice, où l'on invitera les différents groupes à donner la fréquence plutôt que l'effectif, permet d'illustrer le principe de l'algorithme. Les fréquences sont très similaires, et cela indépendamment du site de départ.

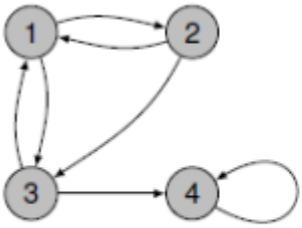
Cet exercice permet aussi une deuxième remédiation pour des groupes qui n'auraient toujours pas compris le principe du surf aléatoire et la modélisation proposée pour l'illustrer.

- **Les limites du surfeur**

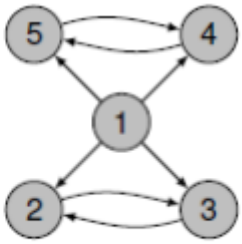
Les trois graphs ci-dessous représentent chacun un problème qui bloque le surfeur. Pour chaque graph, et sans faire de lancer de dé, trouvez-le et expliquez-le. Vous pourrez ensuite tester vos hypothèses en « calculant » avec les dés leur PageRank et en comparant vos résultats avec les autres groupes.



Graph A



Graph B



Graph C

- Graph A :

Le problème :

Cet exercice présenter le premier écueil de l'algorithme : que faire quand un site n'a aucun lien vers d'autres sites ?

La solution :

Assez spontanément, les élèves ont tendance à rajouter des flèches vers les autres sites. Cette solution : le surf équiprobable permet de résoudre cette difficulté.

Le PageRank :

Site n°	1	2	3	4	5
Nombre de visites					

- Graphs B & C

Le problème :

Ces deux exercices présentent le principal écueil de l'algorithme : le blocage dans un puits ou une poche du web.

La solution :

Même si les élèves peuvent avoir de nombreuses idées pour s'en sortir, la solution choisie par Google ne peut être trouvée en raison de la trop grande difficulté mathématique : les chaînes de Markov et le théorème de Perron-Frobenius se trouvent en dehors de leurs connaissances. C'est le moment de leur présenter la solution mise en place par Google.

Le PageRank :

Site n°	1	2	3	4
Nombre de visites				

2. Pour aller plus loin …

- **Heureusement, il y a les maths !**

Un théorème complexe d'algèbre linéaire qui peut s'adapter à notre cas : le théorème de Perron-Frobenius, qui nous dit que si de chaque sommet part un lien vers chacun des autres sommets, alors les fréquences des positions au cours de notre surf aléatoire vont toujours converger vers la même valeur, et cela indépendamment de la position de départ.

- **Idée brillante de L. Page et S. Brin**

Préalable : Si une page ne comporte aucun lien vers l'extérieur, comme le sommet 2 du graphe A, on crée artificiellement un lien de la page vers toutes les autres pages. Détail de l'idée :

- À chaque étape, on continue la promenade aléatoire précédente avec une probabilité de 0,85 ;
- Et avec une probabilité de 0,15, on fait un saut aléatoire (vers n'importe quelle page) avec une probabilité 1/n de tomber sur une page donnée, où n est le nombre de pages.

On peut démontrer mathématiquement que cette façon de procéder permet de faire systématiquement rentrer le graphe considéré dans le champ d'application du théorème de Perron-Frobenius. Cela garantit donc la convergence des fréquences vers des valeurs limites qui seront considérées comme étant les PageRank des pages. Leur méthode résout au passage le cas des puits car elle empêche de se retrouver dans une poche du web sans pouvoir en sortir. De plus, plutôt que de calculer les valeurs limites, calculs qui se révèlent dans la pratique très longs, l'algorithme du PageRank simule comme nous venons de le faire un surfeur aléatoire et prend les fréquences trouvées comme estimation des valeurs limites.

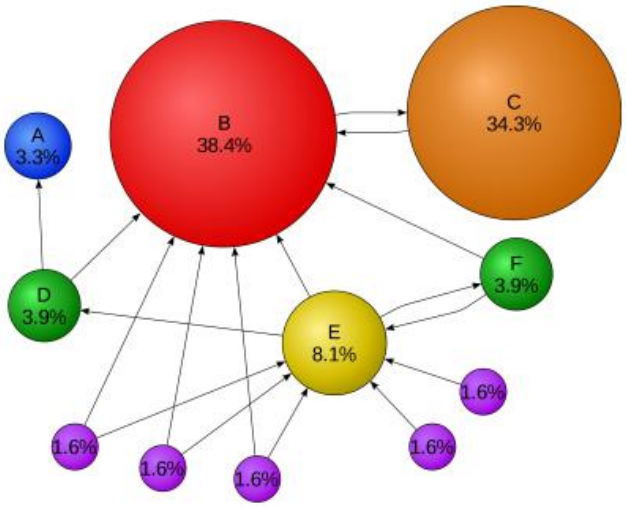
a. Grâce à la méthode des fondateurs de Google, proposez un classement des pages des Graphes B et C.
b. Leur méthode aboutie modifie-t-elle le classement des pages pour lesquelles il n'y avait pas de problèmes ?

À retenir :

Le PageRank ou PR est l'algorithme d'analyse des liens concourant au système de classement des pages Web utilisé par le moteur de recherche Google. Il mesure quantitativement la popularité d'une page web.

Le PageRank n'est qu'un indicateur parmi d'autres dans l'algorithme qui permet de classer les pages du Web dans les résultats de recherche de Google.

Ce système a été inventé par Larry Page, cofondateur de Google.



Liens avec PIX

- **T1. Informations et données**
  - 1.1 Mener une recherche et une veille d'information

Mener une recherche ou une veille d'information pour répondre à un besoin d'information, se tenir au courant de l'actualité d'un projet (... par abonnement à des flux...).

- **T2. Communication et collaboration**
  - 2.2 Partager et publier

Partager et publier des informations et des contenus pour communiquer ses propres productions ou opinions.

  - 2.4 S'insérer dans le monde numérique

Maîtriser les stratégies et enjeux de la présence en ligne, et choisir ses pratiques pour se positionner en tant qu'acteur social, économique et citoyen dans le monde numérique (...)

- **T3. Création de contenu**
  - 3.1 Développer des documents textuels

Produire des documents à contenu majoritairement textuel pour communiquer des idées, rendre compte et valoriser ses travaux (...)

  - 3.2 Développer des documents multimédias

Développer des documents à contenu multimédia pour créer ses propres productions multimédia (...).

  - 3.3 Adapter des documents à leur finalité

Adapter des documents de tous types en fonction de l'usage envisagé (...)

- **T4. Protection et sécurité**
  - 4.1 Sécuriser l'environnement numérique

Sécuriser les équipements, les communications et les données pour se prémunir contre les attaques, pièges, désagréments et incidents susceptibles de nuire au bon fonctionnement des matériels, logiciels, sites internet, et de compromettre les transactions et les données (avec des logiciels de protection, des techniques de chiffrement, la maîtrise de bonnes pratiques, etc.).

  - 4.2 Protéger la vie personnelle et la vie privée

Maîtriser ses traces et gérer les données personnelles pour protéger sa vie privée et celle des autres, et adopter une pratique éclairée (avec le paramétrage des paramètres de confidentialité, la surveillance régulière de ses traces par des alertes ou autres outils, etc.).

URL, protocole HTTP et modèle client-serveur

Fiche activité

Introduction - Faisons le point … Répondez aux questions suivantes :

- **1. À quoi sert un moteur de recherche ?**

  - ☐ À naviguer sur le web de manière totalement sécurisée.
  - ☐ Il permet de trouver et d'installer des programmes ou applications sur notre ordinateur
  - ☐ Il sert à retrouver une ressource (image, document, site web, vidéo, etc.)

- **2. Donnez quelques exemples de moteurs de recherche connus.**

……

……

- **3. Connaissez-vous d'autres types de moteurs de recherche ?**

……

……

……

- **4. La recherche sur ces moteurs est …**

  - ☐ Manuelle : elle est réalisée en temps réel par un humain qui, en fonction d'étiquettes décrivant les différents sites, choisit le plus adapté.
  - ☐ Semi-Automatique : des humains vérifient, de manière asynchrone, les propositions des moteurs de recherche pour une question donnée et les améliore.
  - ☐ Automatique : les moteurs indexent régulièrement les sites en fonction de critères, ce qui leur permet de sélectionner

- **5. Voici deux méthodes manuelles pour hiérarchiser la pertinence de sites web sur un domaine :**

- Faire appel à un.e spécialiste (personne ou groupe de personnes) pour classer les pages. Le classement serait certes très pertinent pour ce domaine, mais il faudrait solliciter un grand nombre de spécialistes de différents domaines pour obtenir un algorithme capable de traiter des requêtes dans tous les domaines possibles et imaginables.
- Demander aux internautes eux- mêmes de voter, pour chaque domaine, et de choisir le classement sur la base de ce vote, considérant, que compte-tenu du grand nombre de votants, le classement obtenu serait pertinent.
  - Quels sont les inconvénients de ces méthodes ?

……

……

- Connaissez-vous un site qui utilise cette méthode ?

.....

Dans l'optique de développer des protocoles entièrement automatisés et efficaces, **Google** a cherché un modèle de hiérarchisation qui soit :

- **Exploitable dans tous les domaines**
- **Utilisable pour tous les mots-clés**
- **Adaptable à un très grand nombre de données**, même **évolutives**, tout en étant **automatisable** et **suffisamment efficace**.

C'est en répondant à ce cahier des charges que ce nouveau venu a réussi l'exploit, **en quelques mois** et malgré l'émergence de Bing ou encore Qwant, **à obtenir le quasi-monopole de la recherche thématique sur le web**.

- **Test :**

Choisissez 3 ou 4 moteurs de recherche différents et effectuez la même recherche, avec les mêmes mots-clés. Choisissez un élément plutôt précis. Par exemple, plutôt « qu'Intelligence Artificielle », préférez « L'Intelligence Artificielle peut-elle créer de l'art ? ».

Que constatez-vous en analysant la première page ? Prenez en compte plusieurs paramètres :

- Les résultats indexés sont-ils les mêmes ?
- Présence de publicité ?
- Pertinence des sites ? (Sites officiels/reconnus ou non, etc.)
- Pertinence des pages par rapport à la requête

L'idée à la base du modèle de Larry Page et Sergey Brin, fondateurs de Google, revient à attribuer à chaque page un nombre positif entre 0 et 1, appelé score **(en anglais "PageRank")** de la page**, qui caractérisera** la pertinence **de cette page. Ils proposent alors de déterminer ce score à partir des deux règles suivantes :**

- R1 : **Le score attribué à une page doit être** d'autant plus élevé que celle-ci est référencée dans une page faisant autorité **(dont le score est élevé).**
- R2 : Le score attribué à une page doit être d'autant moins élevé que celle-ci est référencée dans une page contenant un grand nombre de références.

Leur idée : utiliser un **surfeur aléatoire.**

1. Le principe du surfeur aléatoire

- **Expliquez le principe du surfeur aléatoire** – Vous pouvez chercher sur Internet.

.....

.....

.....

.....

.....

.....

.....

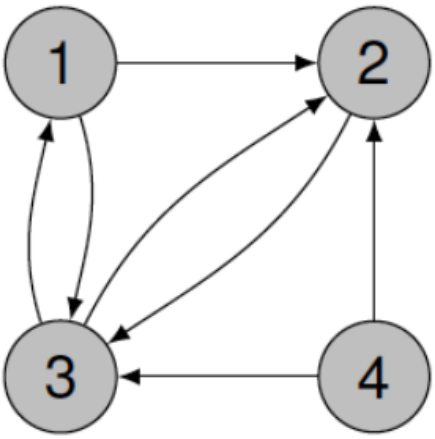
.....

.....

.....

- **À votre tour de tester !**

Pour illustrer comment un algorithme de calcul peut être mis en place à partir de ces règles, nous allons prendre l'exemple du classement de quatre pages. Le problème de l'attribution du score peut être représenté par un graphe orienté : les quatre pages sont représentées par les quatre sommets d'un graphe, dont les arêtes orientées représentent les références (liens) pouvant exister entre ces différentes pages.



On observe notamment que la page 1 référence les pages 2 et 3, et que la page 4 référence, elle, les pages 2 et 3, mais n'est référencée par aucune de ces trois autres pages.

Règles du jeu : On imagine qu'un internaute a effectué une recherche sur un mot-clé. En réalisant l'étape 1, le surfeur a trouvé quatre sites relatifs à ce mot-clé (que nous noterons 1, 2, 3 et 4). En réalisant l'étape 2, il a découvert que :

- Le site 1 a des liens vers les sites 2 et 3.
- Le site 2 a un lien vers le site 3.
- Le site 3 a des liens vers les sites 1 et 2.
- Le site 4 a des liens vers les sites 2 et 3.

Le graphe au-dessus schématise cette situation. Pour rappel, voici les différentes étapes à suivre :

- **Étape 1** : Quand un internaute effectue une recherche sur un mot-clé, il dresse la liste de tous les sites relatifs à ce mot-clé (mais il ne classe rien à cette étape).
- **Étape 2 :** Il dresse ensuite la liste de tous les liens entre les sites de sa liste.
- **Étape 3 :** Il choisit un site au hasard et parcourt les liens, en comptant combien de fois il visite chaque site.
- **Étape 4 :** À la fin de son parcours, il classe les sites du plus visité au moins visité.

Le graphe inclut les étapes 1 et 2. À l'étape 3, vous devez choisir un point de départ au hasard. Pour cela, lancez le dé. Si vous obtenez 1, vous commencerez sur le site 1. Si vous obtenez 2, vous commencerez sur le site 2, etc. Si vous obtenez 5 ou 6, recommencez jusqu'à obtenir 1, 2, 3 ou 4. **Votre groupe fera le processus 30 fois en alternant les lancers.**

Placez votre pion sur le site de départ et ajoutez une visite au**.x** site**.s** pouvant être visité**.s** dans le tableau ci-dessous

Site n°	**1**	**2**	**3**	**4**
Nombre de visites				

Classez les sites du meilleur au moins bon, selon le surfeur : .....

Comparez vos résultats avec ceux des autres groupes. Que constatez-vous ?

.....

.....

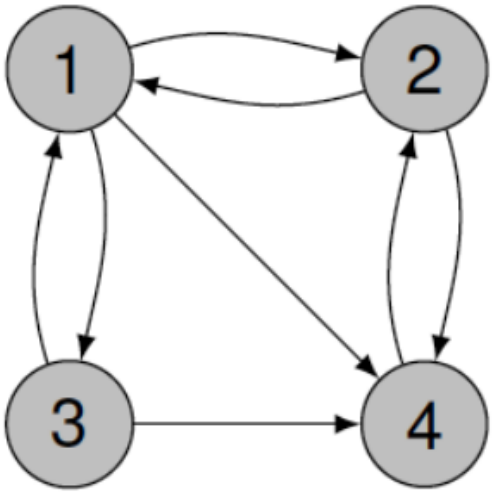
.....

.....

.....

- **Un simple coup de chance ?**

Expérimentez le même protocole, avec le graphe ci-dessous. Il faudra prévoir 50 lancers.



Site n°	1	2	3	4
Nombre de visites				

Classez les sites du meilleur au moins bon, selon le surfeur : .....

Comparez vos résultats avec ceux des autres groupes. Que constatez-vous ?

.....

.....

.....

.....

.....

.....

.....

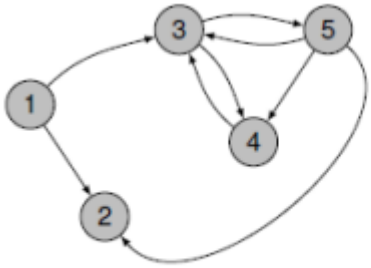
.....

.....

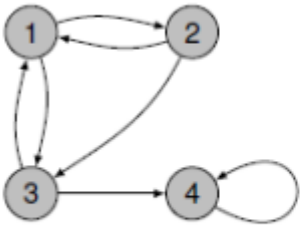
.....

- **Les limites du surfeur**

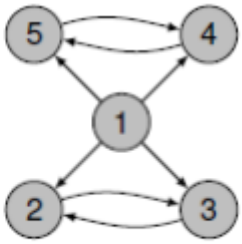
Les trois graphs ci-dessous représentent chacun un problème qui bloque le surfeur. Pour chaque graph, et sans faire de lancer de dé, trouvez-le et expliquez-le. Vous pourrez ensuite tester vos hypothèses en « calculant » avec les dés leur PageRank et en comparant vos résultats avec les autres groupes.



Graph A



Graph B



Graph C

- Graph A :

Le problème :

.....

.....

La solution :

.....

.....

.....

Le PageRank :

Site n°	1	2	3	4	5
Nombre de visites					

- Graphs B & C

Le problème :

.....

.....

La solution :

.....

.....

.....

Le PageRank :

Site n°	1	2	3	4
Nombre de visites				

2. Pour aller plus loin …

- **Heureusement, il y a les maths !**

Un théorème complexe d'algèbre linéaire qui peut s'adapter à notre cas : le théorème de Perron-Frobenius, qui nous dit que si de chaque sommet part un lien vers chacun des autres sommets, alors les fréquences des positions au cours de notre surf aléatoire vont toujours converger vers la même valeur, et cela indépendamment de la position de départ.

- **Idée brillante de L. Page et S. Brin**

Préalable : Si une page ne comporte aucun lien vers l'extérieur, comme le sommet 2 du graphe A, on crée artificiellement un lien de la page vers toutes les autres pages. Détail de l'idée :

- À chaque étape, on continue la promenade aléatoire précédente avec une probabilité de 0,85 ;
- Et avec une probabilité de 0,15, on fait un saut aléatoire (vers n'importe quelle page) avec une probabilité 1/n de tomber sur une page donnée, où n est le nombre de pages.

On peut démontrer mathématiquement que cette façon de procéder permet de faire systématiquement rentrer le graphe considéré dans le champ d'application du théorème de Perron-Frobenius. Cela garantit donc la convergence des fréquences vers des valeurs limites qui seront considérées comme étant les PageRank des pages. Leur méthode résout au passage le cas des puits car elle empêche de se retrouver dans une poche du web sans pouvoir en sortir. De plus, plutôt que de calculer les valeurs limites, calculs qui se révèlent dans la pratique très longs, l'algorithme du PageRank simule comme nous venons de le faire un surfeur aléatoire et prend les fréquences trouvées comme estimation des valeurs limites.

a. Grâce à la méthode des fondateurs de Google, proposez un classement des pages des Graphes B et C.
b. Leur méthode aboutie modifie-t-elle le classement des pages pour lesquelles il n'y avait pas de problèmes ?

……

……

……

……

……

……

……

……

À retenir :

Le PageRank ou PR est l'algorithme d'analyse des liens concourant au système de classement des pages Web utilisé par le moteur de recherche Google. Il mesure quantitativement la popularité d'une page web.

Le PageRank n'est qu'un indicateur parmi d'autres dans l'algorithme qui permet de classer les pages du Web dans les résultats de recherche de Google.

Ce système a été inventé par Larry Page, cofondateur de Google.