M1 Informatique – UE Projet Carnet de bord : les coulisses de la recherche documentaire

Noms, prénoms et spécialité

ZHAO Wenzhuo, Science et Technologies du Logiciel (STL)

LU Zhaojie, Science et Technologies du Logiciel (STL)

HOU Zhen, Science et Technologies du Logiciel (STL)

YANG Chengyu, Science et Technologies du Logiciel (STL)

Sujet

Scikit network : bibliothèque Python pour les grands graphes

Introduction

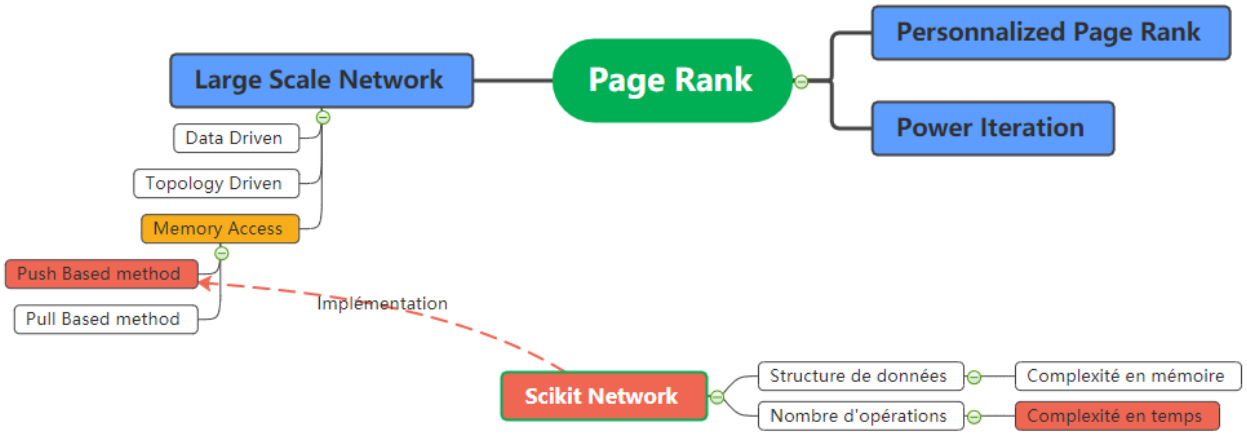
Dans la recherche des grands graphes dans le domaine d'informatique, l'algorithme de Page Rank est un indicateur de mesurer la popularité de pages web sur le réseau du web. Ce système a été inventé par Larry Page, cofondateur de Google donc est utilisé par le moteur de recherche Google.

Le principe de cet algorithme est de faire un classement sur les pages web en attribuant à chaque page une valeur proportionnelle au nombre de fois que passerait par cette page un utilisateur parcourant le graphe du Web en cliquant aléatoirement, sur un des liens apparaissant sur chaque page. Après de nombreuses corrections et améliorations apportées à cet algorithme, il peut traiter des conditions plus compliquées et prendre moins de mémoire et temps pendant le calcul.

Après avoir cherché des articles scientifiques, nous avons mieux compris cet algorithme et ses versions améliorées avec l'aide de l'encadrant de ce projet. L'objectif de notre projet est de comprendre et comparer plusieurs versions améliorées de ce classique algorithme. En faisant la recherche sur cet algorithme, nous proposons une implémentation efficace de la version "Push based Page Rank" en faisant une contribution au Scikit-network qui est une librairie Python pour l'analyse de graphes de grande taille développé à Télécom Paris.

Les mots clés retenus

- Page Rank
- Personalized Page Rank
- Power iteration
- Topology-driven PageRank
- Data-driven PageRank
- Memory Access pattern
- Pull-based PageRank
- Push-based PageRank
- Large scale Network
- Scikit Network
- Structure de données
- Nombre d'opérations
- Complexité en temps
- Complexité en mémoire



Descriptif de la recherche documentaire

Dans notre recherche documentaire, nous avons utilisé plusieurs outils de recherche pour notre recherche bibliographique. Nous nous sommes concentrés sur l'utilisation du moteur de recherche google scholar filtré sur le catalogue de Sorbonne Université et le Système universitaire de documentation. Nous avons également utilisé les ressources de la bibliothèque en ligne de l'université plus particulièrement la base de données web of science. Nous avons sélectionné et enregistré sur Zotero les ressources qui nous ont semblé les plus enrichissantes pour notre projet.

Nous avons recouru aux bases de données scientifiques comme Web of Science, Europress qui se trouvent dans la bibliothèque de la Sorbonne. Ils nous ont permis d'accéder à une multitude de sources scientifiques telles que des thèses ou des articles scientifiques particulier par exemple : « Scalable Data-Driven PageRank: Algorithms, System Issues, and Lessons Learned » qui sont des synthèses bibliographiques écrit par des chercheurs.

Nous avons également cherché sur Google Scholar et nous avons trouvé un mémoire traitant de l'algorithme pour bien améliorer l'algorithme de Page Rank.

Bibliographie produite dans le cadre du projet

Dans le cadre de notre projet et comme mentionné précédemment, nous nous sommes référés à plusieurs outils pour produire notre bibliographie aux normes IEEE.

[1] J. J. Whang, A. Lenharth, I. S. Dhillon, et K. Pingali, « Scalable Data-Driven PageRank: Algorithms, System Issues, and Lessons Learned », in *Euro-Par 2015: Parallel Processing*, vol. 9233, J. L. Träff, S. Hunold, et F. Versaci, Éd. Berlin, Heidelberg: Springer Berlin Heidelberg, 2015, p. 438‑450, doi: 10.1007/978-3-662-48096-0_34

[2] R. Andersen, F. Chung and K. Lang, « Local Graph Partitioning using PageRank Vectors », *2006 47th Annual IEEE Symposium on Foundations of Computer Science (FOCS'06)*, Berkeley, CA, USA, 2006, pp. 475-486, doi: 10.1109/FOCS.2006.44.

[3] J. Sauvola et M. Pietikäinen, « Adaptive document image binarization », *Pattern Recognition*, vol. 33, nº 2, p. 225‑236, févr. 2000, doi: 10.1016/S0031-3203(99)00055-2.

[4] S. Brin et L. Page, « The anatomy of a large-scale hypertextual Web search engine », *Computer Networks and ISDN Systems*, vol. 30, nº 1‑7, p. 107‑117, avr. 1998, doi: 10.1016/S0169-7552(98)00110-X.

[5] P. Berkhin, « A Survey on PageRank Computing », *Internet Mathematics*, vol. 2, nº 1, p. 73‑120, janv. 2005, doi: 10.1080/15427951.2005.10129098.

[6] P. Berkhin, « Bookmark-Coloring Algorithm for Personalized PageRank Computing », *Internet Mathematics*, vol. 3, nº 1, p. 41‑62, janv. 2006, doi: 10.1080/15427951.2006.10129116.

[7] T.Haveliwala, *Efficient computation of PageRank*. Stanford. 1999

[8] A. Langville et C. Meyer, « Deeper Inside PageRank », *Internet Math*, vol. 1, nº 3, p. 335‑380, janv. 2004, doi: 10.1080/15427951.2004.10129091.

[9] A.Arasu, J.Novak, A.Tomkins, «PageRank computation and the structure of the web: Experiments and algorithms », *Proceedings of the Eleventh International World Wide Web Conference, Poster Track*. 2002. p. 107-117.

[10] Q. Liu *et al.*, « An Influence Propagation View of PageRank », *ACM Trans. Knowl. Discov. Data*, vol. 11, nº 3, p. 1‑30, avr. 2017, doi: 10.1145/3046941.

Evaluation des sources

1. *Scalable Data-Driven PageRank: Algorithms, System Issues, and Lessons Learned*

Ce document a été trouvé suite à une recherche sur le site Web of Science. C'est une thèse écrite par Joyce Jiyoung Whang, Andrew Lenharth,Inderjit S. Dhillon et Keshav Pingal en 2015. Ils étaient tous Ph.D à University of Texas at Austin. Cette thèse a été publiée dans le journal « Lecture Notes in Computer Science » qui est une série d'ouvrages informatiques publiés par Springer Science + Business Media depuis 1973. Le facteur d'impact de ce journal est 0.402.

Nous avons utilisé beaucoup de résultats dans cet article dans la partie théorique de notre rapport de projet.

2. *The anatomy of a large-scale hypertextual Web search engine*

Ce document a été trouvé suite à une recherche sur le site Web of Science. C'est une thèse écrite par Sergey Brin et Lawrence Page en 1998. Ils étaient tous Ph.D en informatique à l'université Stanford en Californie. Dans le cadre de leurs recherches, ils vont lancer un projet de moteur de recherche internet qui deviendra par la suite la société Google en 1998. Cette thèse a été publiée dans le journal « Computer Networks and ISDN Systems ». Son nombre total de citations est de 4893 fois. Le facteur d'impact de ce journal est 0.352.

Cet article explique l'algorithme de pagerank original et c'est également le document le plus ancien dans ce domaine.

3. *An Influence Propagation View of PageRank*

Ce document a été trouvé suite à une recherche sur le site Web of Science. C'est une thèse écrite par Liu Qi, Xiang Biao, Yuan Nicholas Jing, Chen Enhong, Xiong Hui,

Zheng Yi et Yang Yu. Son nombre total de citations est de 33 fois. Cette thèse a été publiée dans le journal « ACM Transactions on Knowledge Discovery from Data ». Le facteur d'impact de ce journal est 2.01.

Cet article explique et analyse principalement l'impact des algorithmes pagerank appliqués sur la société.