Rappels des concepts de base:

modèle de données web, fonctions de similarité, recherche d’information et évaluation des performances, concepts de base des réseaux sociaux.

Introduction

L'analyse des données du Web nécessite un large éventail de concepts, de théories et d'approches ainsi qu'une variété de contextes d'application.

Nous présentons d'abord une introduction aux modèles de données Web. Ensuite, Les deux concepts et approches essentiels en matière de recherche d'information - les mesures de similarité et les métriques d'évaluation - En outre, certains concepts de base des réseaux sociaux sont abordés dans ce chapitre.

Modélisation des données Web

Pour implémenter efficacement le Web mining, il existe de nombreux types d'expressions de données pouvant être utilisées pour modéliser la co-occurrence des interactions entre les utilisateurs Web et les pages, telles que la matrice, le graphe orienté et non orienté.

Par exemple les interactives entre les utilisateurs et les pages Web, et les relations mutuelles entre les pages Web sont modélisées sous la forme d'une matrice de lien hypertexte adjacent (lien entrant ou sortant) ou d'une matrice session-page vue.

Similarité

Afin de retrouver les documents similaires, nous avons besoin d'une mesure de similarité.

La similarité est une mesure numérique permettant d’indiquer la distance entre les nœuds afin de les comparer.

Il existe deux types de similarités:

- la similarité des attributs : est calculée à l'aide des caractéristiques internes des nœuds indépendamment de la topologie du réseau.
- la similitude structurelle. calculée en se basant sur de la topologie du réseau. Pour la mesurer, certaines méthodes calculent la distance entre deux nœuds, d'autres, les chemins locaux et d'autres comptent le nombre de voisins que deux nœuds ont en commun.

La similarité dépend souvent du système ou de l'application spécifique en question.

Exemples de Similarité

- L’indice de Jaccard : (Jaccard, 1901) $S_{xy} = \frac{|\Gamma(x) \cap \Gamma(y)|}{|\Gamma(x) \cup \Gamma(y)|}$
- L’indice de cosinus : $S_{xy} = \frac{|\Gamma(x) \cap \Gamma(y)|}{\sqrt{|\Gamma(x)| \times |\Gamma(y)|}}$
- L’indice de cosinus : (Salton & McGill, 1983)

$$similarity(A,B) = \frac{A \cdot B}{\|A\| \times \|B\|} = \frac{\sum_{i=1}^{n} A_i \times B_i}{\sqrt{\sum_{i=1}^{n} A_i^2} \times \sqrt{\sum_{i=1}^{n} B_i^2}}$$

- ù $\Gamma(x)$ représente l'ensemble des voisins du nœud x et $|\Gamma(x)|$ désigne son degré.

Évaluation des performances de recherche d'informations 1/2

Un processus de recherche d'informations commence lorsqu'un utilisateur saisit une requête dans le système.(des mots-clés qui représentent les informations requises)

Chaque élément d'information est extrait du Web et stocké dans le référentiel. avec un index

La plupart des systèmes de RI calculent un score numérique sur la façon dont chaque objet de la base de données correspond à la requête, et classent les objets en fonction de cette valeur.

Évaluation des performances de recherche d'informations 2/2

Il existe plusieurs mesures pour l'évaluation des performances des systèmes de recherche d'information

On suppose qu'on a une requête particulière et un ensemble de documents pertinents et un autre non pertinent alors les mesures suivantes sont définies

$$precision = \frac{|\{retrieved\ documents\}| \cap |\{relevant\ documents\}|}{|\{retrieved\ documents\}|}$$

$$Recall = \frac{|\{retrieved\ documents\}| \cap |\{relevant\ documents\}|}{|\{relevant\ documents\}|}$$

F-measure

$$F = \frac{2 \cdot precision \cdot recall}{precision + recall}$$

Les réseaux sociaux

La notion de réseau existe dans plusieurs domaines de recherche en informatique comme dans d’autres disciplines. En sociologie par exemple, il existe les réseaux dont les nœuds sont des individus et les liens entre eux sont les relations sociales, comme la parenté, l'amitié ou les intérêts communs. Dans les réseaux de collaboration, les nœuds sont des personnes et la relation entre deux individus est le travail. En biologie, il existe les réseaux métaboliques dont les nœuds sont des protéines et les liens sont les réactions chimiques entre elles. En web les nœuds sont des pages et les liens sont les hyperliens entre ces pages.

Les études ont montré qu’avec des algorithmes, il est possible de déterminer la personnalité d’un individu mieux que son entourage (Youyou, et al., 2015)

La modélisation des réseaux par des graphes facilite l’étude et la compréhension de leur structure

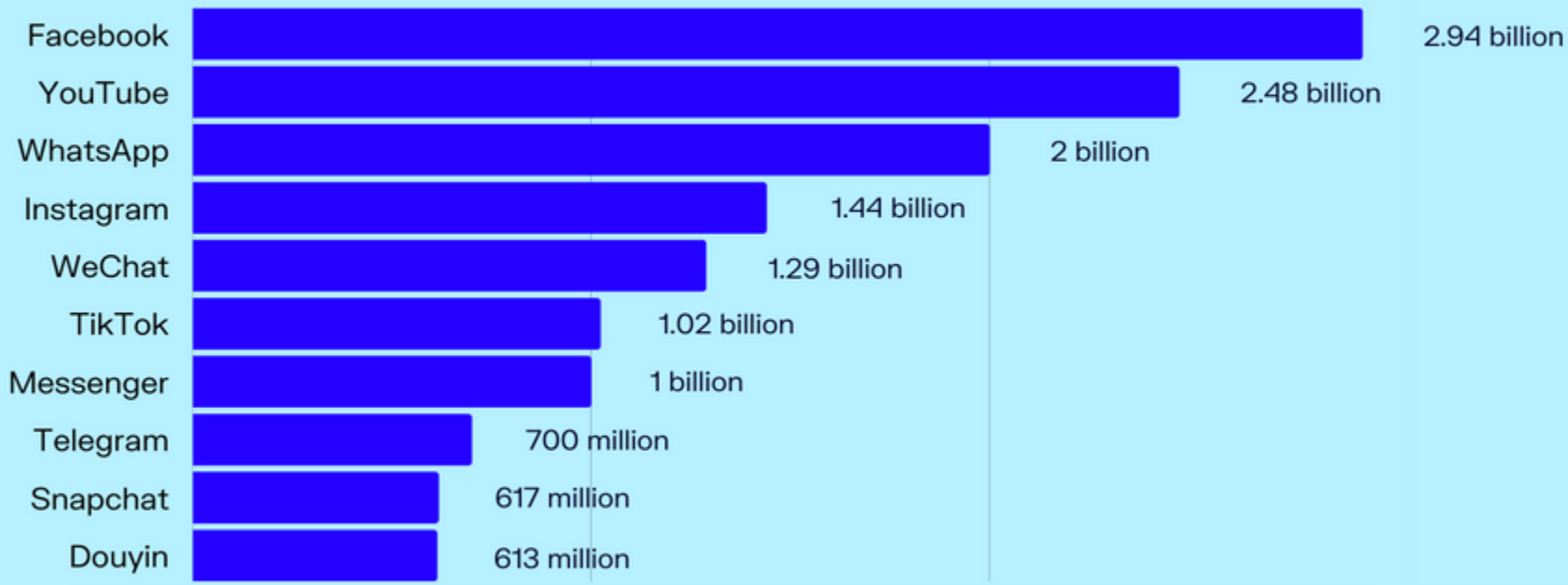
Les réseaux sociaux

En sciences humaines et sociales, un réseau social (en anglais : Social network) est une structure sociale composée d'individus et/ou d’organisations appelées « nœuds », qui sont liés par des relations sociales de plusieurs types, comme la parenté, l'amitié, les intérêts communs, l'échange de transactions financières, ou les relations de connaissance et de croyance (Tonnies & Loomis, 2002). Cette structure constitue généralement un groupement qui a un sens comme la famille, les collègues, un groupe d'amis, une communauté, etc.

en dehors du domaine des sciences sociales « médias sociaux »

« un groupe d’applications en ligne fondé sur l’idéologie et la technologie du Web 2.0 et permettent la création et l’échange du contenu généré par les utilisateurs » (Kaplan & Haenlein, 2010).

Most Popular Social Media Platforms in 2022



Source: DataReportal

shopify

Analyse des réseaux sociaux (SNA)

- Le Web peut être considéré comme une société virtuelle ou un réseau social virtuel, où chaque page peut être considérée comme un acteur social et chaque lien hypertexte comme une relation.

- Le SNA est l’ensemble de méthodes et d'outils permettant d'étudier et d'analyser les relations, les interactions et les communications des individus ou des sous-groupes dans les réseaux sociaux (Otte & Rousseau, 2002).

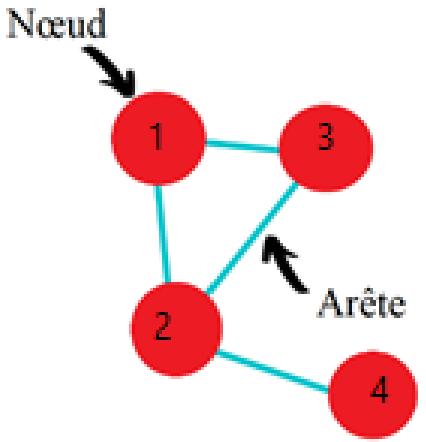
- Le SNA peut être mis en œuvre de deux manières principales : la visualisation (sociogrammes) et l'analyse mathématique (théorie des graphes).

Certaines notions

Taille : le nombre de sommets présentés dans un réseau,

La **densité** désigne le rapport entre le nombre de toutes les arêtes existantes et le nombre total possible d'arêtes dans le réseau. $\frac{m}{2n(n-1)}$

communauté : signifie que les membres d'un même groupe présentent une grande similitude sur certains aspects, tels que les croyances culturelles ou religieuses, les intérêts ou les préférences, etc.



La centralité

La centralité reflète l'importance des acteurs au sein d'un réseau. Le calcul de la centralité nous permet d'identifier les personnes les plus influentes dans un réseau social

Il existe plusieurs mesures de centralité les plus utilisées sont :

- La centralité de degré (degree centrality) est la plus simple mesure à mettre en œuvre. Il est défini comme le nombre de voisins que possède un nœud. $deg(v_i) = k_i = \sum_j A_{i,j}$

- Centralité de proximité (closeness centrality) est une mesure basée sur la somme des distances géodésiques d'un acteur donné à tous les autres nœuds. $C(v) = \frac{N-1}{\Sigma_u d(u,v)}$, où d(u,v) est la distance u et v.

- Centralité d'intermédiarité (betweenness centrality) est une mesure basée sur la fréquence à laquelle un acteur se trouve en position intermédiaire le long du chemins géodésiques reliant deux autres nœuds.

$$C_B(v) = \sum_{s \neq v \neq t \in V} \frac{\sigma_{st}(v)}{\sigma_{st}}$$

où σ_{st} est le nombre total de chemins les plus courts du s au t et $\sigma_{st}(v)$ est le nombre de ces chemins qui passent par v.