Projet Final ARS

Système de recommandation d'articles de Wikipédia

Arnaud BROSSAY
Théo HOPSORE

9 janvier 2024

Table des matières

I Introduction

En 2023, Wikipédia, une encyclopédie communautaire gratuite, était le 5e site le plus utilisé au monde. Avec plus de 30 millions d'articles recenssés, l'une des force du site est l'interconnexion de ces derniers formant ainsi une structure similaire à celle d'un réseau social.
Cette interconnexion entre les différents articles est notament représentée par la recommandations d'articles en relation avec la page visitée.
Pour ce faire, le site utilise un algorithme de recommandation qui lui est propre basé sur des métriques précises afin de créer un lien entre différentes pages.

Au cours de ce projet, nous allons tenter de proposer un algorithme égo-centré multicritère afin de recommander un article en fonction de l'article que l'on consulte actuellement.
Nous allons étudier le graphe dirigé *Wikipedia.gml* qui recense les liens entre 27000 pages du site. Dans un premier temps, nous allons définir différentes fonctions de modularité afin de construire une communauté locale.

II Exploration du graphe

Comme dit précédement, nous allons travailler sur la base du graphe dirigé *Wikipedia.gml* qui représente un regroupement de plus de 27 000 articles issus du site éponyme.

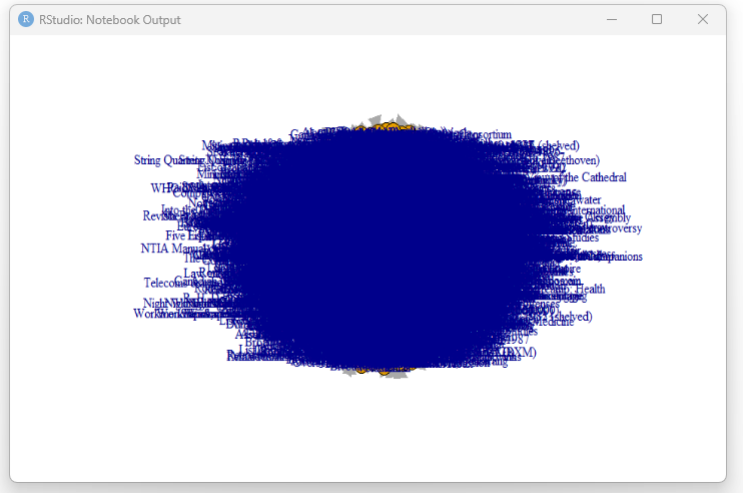


FIGURE 1 – Un premier aperçu du graphe dirigé *Wikipedia.gml* non-traité

```
Nombre de noeuds : 27475
Nombre de liaisons : 85729
Densité du graph : 0.0001135712
Diamètre du graph : 41
Transitivité du graph: 0.04908247
Graph connecté ? : FALSE
```

FIGURE 2 – Informations du graphe *Wikipedia.gml* non-traité

Nous pouvons comprendre d'après la figure et les informations ci-dessus, le graphe est assez complexe. Nous avons donc pris la décision de ne travailler que sur un sous-ensemble de celui afin d'obtenir des résultats rapides, étudier un graphe de cet envergure étant couteux en temps et en ressources informatiques.

Nous allons donc passer le graphe de orienté à non-orienté, en réduire de taille à un sous-ensemble aléatoire de 2000 sommets puis séléctionner le plus grand sous-graphe connexe issus de ce sous-ensemble. Cela permettra d'accelerer significativement l'analyse et les calculs tout en conservant une représentation significative du réseau.

```r
{r}
BiggestConnexeGraph = function(g){

  clus = clusters(g)
  nodes = which(clus$membership == which.max(clus$csize))

  return(induced.subgraph(g,nodes))
}

graph = as.undirected(graph) # On rend le graph non dirige
graph = induced.subgraph(graph, sample.int(vcount(graph), 2000)) # On reduit la
taille du graph aleatoirement pour aller plus vite
sousG = BiggestConnexeGraph(graph) # On garde le sous graph connexe de taille
maximale
vertex_attr(sousG)$id = 1:vcount(sousG) #attrbution d'id
```

FIGURE 3 – Code permettant le traitement du graphe

Nous obtenons ainsi un graphe plus plus rapide et simple à traiter, grandant des liens significatifs entre les différents sommets.

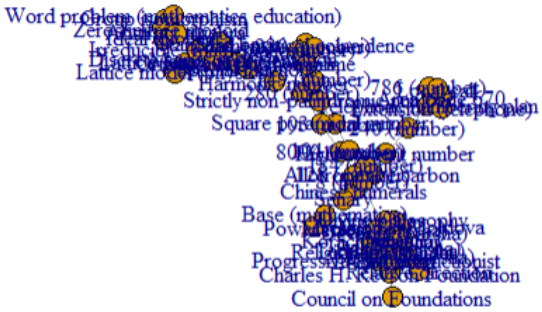


FIGURE 4 – Le sous-graphe connexe issue de *Wikipedia.gml*

```
Nombre de noeuds : 103
Nombre de liaisons : 105
Densité du graph : 0.01998858
Diamètre du graph : 14
Transitivité du graph: 0.003952569
Graph connecté ? : TRUE
```

FIGURE 5 – Informations du sous-graphe connexe de *Wikipedia.gml*

III Calcul des Modularités

Afin de trouver des articles pertinents à recommander, nous allons utiliser des fonctions de calcul de modularité. Celles-ci vont alors permettre de trouver des clusters d'article liés par des thématiques similaires.

Nous allons donc utiliser trois différentes fonctions de modularité afin de comparer différentes partitions du graphe, nous permettant ainsi d'identifier des structures significatives et d'obtenir des informations pertinentes sur la connectivité des articles.
Nous avons ainsi :

— la modularité locale R

$$R = \frac{B_{in}}{B_{in} + B_{out}}$$

— la modularité locale M

$$M = \frac{D_{in}}{D_{out}}$$

— et la modularité locale L

$$L = \frac{L_{in}}{L_{out}}$$

```r
{r}
mod_R = function(g, C, B, S){

  Bin = length(E(g)[B %--% B])
  Bout = length(E(g)[B %--% S])

  return (Bin/(Bin+Bout))
}
```

```r
{r}
mod_M <- function( g ,C, B, S ) {

D <- union(C,B)
din <- length(E(g)[D %--% D])
dout <- length(E(g)[B %--% S])

return (din /dout)
}
```

```r
{r}
neighbors_in <- function(n,g,E){
  return(length(intersect(neighbors(g,n),E)))
}

mod_L <- function(g,C,B,S){
  D <- union(C,B)
  lin <- sum(sapply(D,neighbors_in,g,D))/length(D)
  lout <- sum(sapply(B,neighbors_in,g,S))/length(B)
  return(lin/lout)
}
```

FIGURE 6 – Déclaration des différentes fonctions de modularité

Nous allons ensuite calculer la communauté locale d'un noeud en fonction des trois formules de modularité locale.

```
local_com = function(target, g){

  if (is.igraph(g) && target %in% V(g)){

    C <- c()
    B <- c(target)
    S <- c(V(g)[neighbors(g, target)]$id)
    Q <- 0
    new_Q <- 0

    while ((length(S) > 0) && (new_Q >= Q)){

      QS <- sapply(S, compute_quality, g, C, B, S)
      new_Q <- max(QS)

        if (new_Q >= Q){
          s_node = S[which.max(QS)]
          res = update(s_node, g, C, B, S)

          C = res$C
          B = res$B
          S = res$S
          Q = new_Q
        }
    }
    return(union(C, B))

  } else {

    stop('Erreur')
  }
}
```

FIGURE 7 – Fonciton *local_com* permettant de calculer la communauté locale d'un noeud

IV Découpage et explication de l'algorithme

Comme dit précédement, l'objectif est de proposer des articles pertinents par rapport au sujet de la page consultée, mais qui n'ont pas un rapport direct avec cette derniere. La recommandation d'un article en fonction d'une page fait entrer en jeu un processus en plusieurs étapes :

- Le calcul des communautées égo-centrées
- Choix des mesures dyadiques
- Classement des recommandations

IV.1 Calcul des communautées égo-centrées

Comme expliqué précédement, dans un premier temps, nous calculons les communautées égo-centrées qui sont des sous-graphes centrés sur un noeud spécifique. Celà va permettre d'analyser les structures locale autour de chaque noeud dans le réseau.
La modularité est donc une mesure cruciale qui va venir évaluer la signification d'une partition d'un réseau en communauté. Nous avons choisi d'utiliser nos 3 mesures de la modularité en même temps (modR, modM et modL) qui sont pondérés par des poids afin de mesurer la qualité de notre partitionnement en fonction de plusieurs critères à la fois. Car après plusieurs tests, et compte tenu de notre graphe et de notre sous-graphe d'essai, nous avons convenu qu'il était plus optimal d'avoir une répartition des poids identiques. Nous avons donc un poids de 0.33 pour chaque modularité, ce qui assure une mesure optimale dans la majorité des cas.

```
compute_quality = function(n, g, C, B, S){

  quality = update(n, g, C, B, S)
  C = quality$C
  B = quality$B
  S = quality$S

  modR = mod_R(g, C, B, S)
  modM = mod_M(g, C, B, S)
  modL = mod_L(g, C, B, S)

  res = (0.33 * modR) + (0.33 * modM) + (0.33 * modL)

  return(res)
}
```

FIGURE 8 – Fonction *compute_quality* permettant de calculer une moyenne pondéré des modularités

IV.2 Choix des mesures dyadiques

Les mesures dyatiques sont des indicateurs qui permettent d'évaluer les interactions entre des paires de noeuds donnés dans le réseau.

Dans notre cas, nous allons utiliser :

— la similarité de Jaccard, qui se base sur le nombre d'éléments communs et distincts entre deux ensemble de voisins de noeuds.

$$Similarité\ de\ Jaccard = \frac{A \cap B}{A \cup B}$$

— la mesure d'Adamic-Adar, prend en compte les voisins communs entre deux noeuds donnant une grande importance aux voisins rare et peu communs aux noeuds et inversemment.

$$AdamicAdar(i,j) = \sum_{k \in N(i) \cap N(j)} \frac{1}{log(|N(k)|)}$$

— et les chemins les plus courts, qui représente le trajet le plus court entre deux noeuds, calculé en fonction du nombre minimum d'arêtes nécessaires pour passer d'un noeud à l'autre.

IV.3 Classement des recommandations

On va donc venir calculer le score de chaque noeud en fonction d'un noeud donné. On obtient le score final d'un noeud ainsi :

$$score\_noeud = Similarité\ de\ Jaccard\ + AdamicAdar + chemins\ les\ plus\ cours\ * 0.3$$

On va ainsi ensuite retourner la page avec le meilleur score.

```
{r}
########## Ensemble Ranking ##########

score_tab = matrix(ncol = 2)
colnames(score_tab) <- c("Noeud","Score")

for (i in 1:length(gg$com)){
  if (sim_sp[i] > 1){
    score_i = c(gg$com[i],sim_jacc_min[1,][i+1] + sim_aa[,noeud_cible][i+1] +
sim_sp[i]*0.3)
    score_tab = rbind(score_tab, score_i)
  }
}
score_tab = score_tab[order(score_tab[,2], decreasing=FALSE),]
score_tab
if (length(score_tab) > 2) {
  print(paste("Recommandations pour le noeud ", vertex_attr(sousG, index
=noeud_cible)$label, ": "))

  for (i in 1:(nrow(score_tab)-1)){
    label_i = vertex_attr(sousG, index=score_tab[i])$label
    if (is.null(label_i)){
      print(paste("Recommandation ", i, ": ",score_tab[i]))
    } else {
      print(paste("Recommandation ", i, ": ",label_i))
    }
  }
} else {
  print(paste("Pas de recommandation pour le noeud ", vertex_attr(sousG, index
=noeud_cible)$label))
}
```

FIGURE 9 – Déclaration de la fonction *score_tab*

V Résultats

Ainsi, nous avons fait le test en partant de la page Wikipedia *Q-Q plot*, qui est une technique de visualisation probabilistique.

Q–Q plot

13 languages

Article Talk Read Edit View history Tools

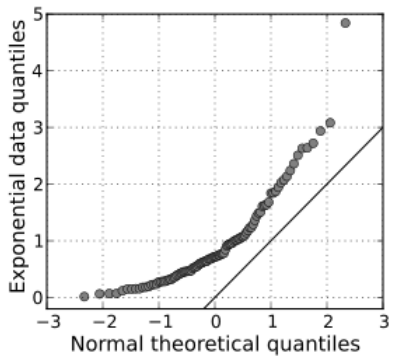
From Wikipedia, the free encyclopedia

Not to be confused with P–P plot.

In statistics, a **Q–Q plot** (**quantile–quantile plot**) is a probability plot, a graphical method for comparing two probability distributions by plotting their *quantiles* against each other.[1] A point (x, y) on the plot corresponds to one of the quantiles of the second distribution (y-coordinate) plotted against the same quantile of the first distribution (x-coordinate). This defines a parametric curve where the parameter is the index of the quantile interval.

If the two distributions being compared are similar, the points in the Q–Q plot will approximately lie on the identity line $y = x$. If the distributions are linearly related, the points in the Q–Q plot will approximately lie on a line, but not necessarily on the line $y = x$. Q–Q plots can also be used as a graphical means of estimating parameters in a location-scale family of distributions.

A Q–Q plot is used to compare the shapes of distributions, providing a graphical view of how properties such as location, scale, and skewness are similar or different in the two distributions. Q–Q plots can be used to compare collections of data, or theoretical distributions. The use of Q–Q plots to compare two samples of data can be viewed as a non-parametric approach to comparing their underlying distributions. A Q–Q plot is generally more diagnostic than comparing the samples' histograms, but is less widely known. Q–Q plots are commonly used to compare a data set to a theoretical model.[2][3] This can provide an assessment of goodness of fit that is graphical, rather than reducing to a numerical summary statistic. Q–Q plots are also used to compare two theoretical distributions to each other.[4] Since Q–Q plots compare distributions, there is no need for the values to be observed as pairs, as in a scatter plot, or even for the numbers of values in the two groups being compared to be equal.



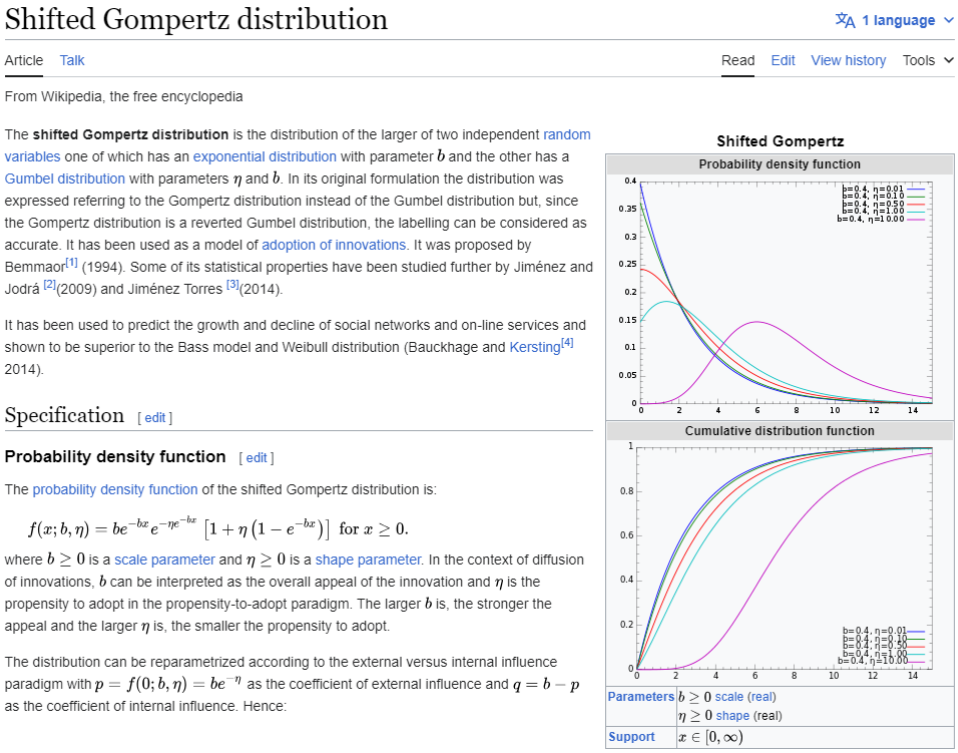
A normal Q–Q plot of randomly generated, independent standard exponential data, ($X \sim \mathrm{Exp}(1)$). This Q–Q plot compares a sample of data on the vertical axis to a statistical population on the horizontal axis. The points follow a strongly nonlinear pattern, suggesting that the data are not distributed as a standard normal ($X \sim \mathrm{N}(0,1)$). The offset between the line and the points suggests that the mean of the data is not 0. The median of the points can be determined to be near 0.7

FIGURE 10 – Page Wikipedia du *Q-Q plot*

Nous obtenons ainsi en recommendation avec un score d'approximativement 1.54 la page *Shifted Gompertz distribution* qui est une distribution statistique.

```
        Noeud    Score
score_i    53 1.542695
           NA       NA
[1] "Recommandations pour le noeud  Q-Q plot : "
[1] "Recommandation  1 :  Shifted Gompertz distribution"
```

FIGURE 11 – Score de recommendation



Shifted Gompertz distribution

1 language

Article Talk Read Edit View history Tools

From Wikipedia, the free encyclopedia

The **shifted Gompertz distribution** is the distribution of the larger of two independent random variables one of which has an exponential distribution with parameter b and the other has a Gumbel distribution with parameters η and b. In its original formulation the distribution was expressed referring to the Gompertz distribution instead of the Gumbel distribution but, since the Gompertz distribution is a reverted Gumbel distribution, the labelling can be considered as accurate. It has been used as a model of adoption of innovations. It was proposed by Bemmaor[1] (1994). Some of its statistical properties have been studied further by Jiménez and Jodrá [2](2009) and Jiménez Torres [3](2014).

It has been used to predict the growth and decline of social networks and on-line services and shown to be superior to the Bass model and Weibull distribution (Bauckhage and Kersting[4] 2014).

Specification [edit]

Probability density function [edit]

The probability density function of the shifted Gompertz distribution is:

$$f(x;b,\eta) = be^{-bx}e^{-\eta e^{-bx}}\left[1+\eta\left(1-e^{-bx}\right)\right] \text{ for } x \geq 0.$$

where $b \geq 0$ is a scale parameter and $\eta \geq 0$ is a shape parameter. In the context of diffusion of innovations, b can be interpreted as the overall appeal of the innovation and η is the propensity to adopt in the propensity-to-adopt paradigm. The larger b is, the stronger the appeal and the larger η is, the smaller the propensity to adopt.

The distribution can be reparametrized according to the external versus internal influence paradigm with $p = f(0;b,\eta) = be^{-\eta}$ as the coefficient of external influence and $q = b - p$ as the coefficient of internal influence. Hence:

Shifted Gompertz

Probability density function

Cumulative distribution function

Parameters	$b \geq 0$ scale (real) $\eta \geq 0$ shape (real)
Support	$x \in [0, \infty)$

FIGURE 12 – Page Wikipedia du *Shifted Gompertz distribution*

Les deux pages ont pour sujet commun les statistiques mais elles ne sont jamais citées l'une par rapport à l'autre.

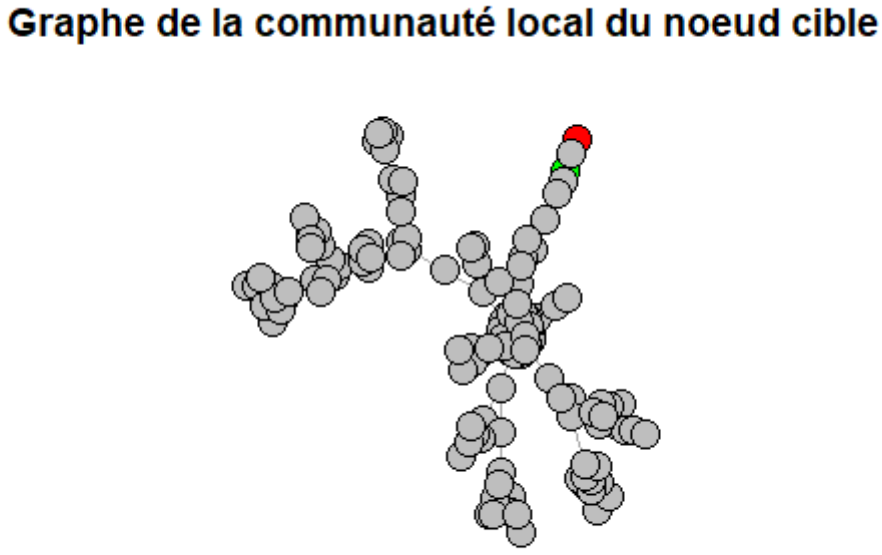


FIGURE 13 – Graphe de la communauté locale du noeud cible

VI Conclusion

Au cours de ce projet, nous avons élaboré un algorithme de recommandation pour les pages Wikipedia en se basant sur l'analyse des réseaux sociaux et des communautés présentes au sein de ces réseaux. Étant donné que le graphe utilisé pour cette étude était trop volumineux, nous avons dû travailler sur une partie restreinte de celui-ci. Malgré cette contrainte, notre algorithme a réussi à proposer des recommandations d'articles étroitement liés au nœud cible.

Par ailleurs, l'application de cet algorithme à d'autres domaines, tels que la recommandation d'amis sur des réseaux sociaux comme Facebook, pourrait également être intéressante. Les systèmes de recommandation connaissent une utilisation croissante dans divers secteurs tels que la publicité, le streaming, etc., visant à offrir des expériences personnalisées correspondant davantage à nos préférences. Il serait donc pertinent de comparer l'efficacité de notre algorithme avec des systèmes plus sophistiqués et affinés qui se perfectionnent chaque jour, ce qui nous permettrait d'approfondir nos connaissances sur ce sujet.