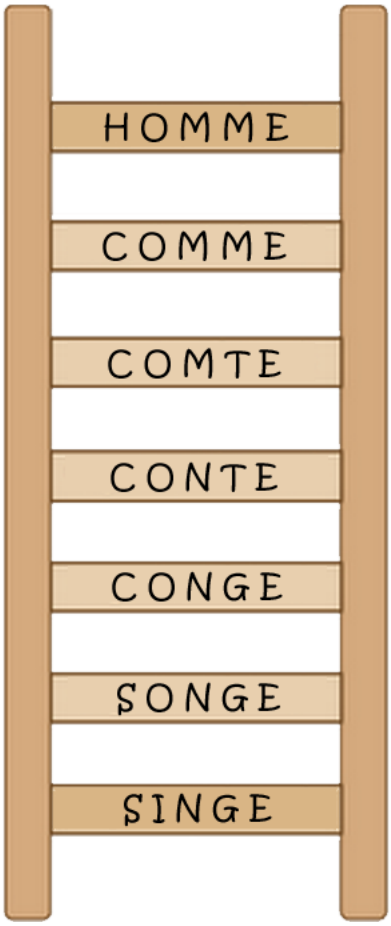
Une application des graphes : les échelles de mots

Didier Müller
Lycée cantonal de Porrentruy

La théorie des graphes est très intéressante, et on peut l'aborder au lycée déjà. J'ai d'ailleurs écrit un cahier de la CRM sur le sujet cette année [7]. Un souci est que les exercices sont souvent soit trop simples (et l'on ne voit pas forcément l'utilité d'utiliser cette théorie), soit trop longs (et les élèves se découragent). Or, voici que je suis tombé par hasard au début de l'année sur un article de Jon McLoone : *The Longest Word Ladder Puzzle Ever* [5]. Un superbe problème, où la théorie des graphes n'apparaît pas immédiatement, et qui n'est résoluble qu'avec un ordinateur, moyennant des programmes assez simples à écrire.



Un **doublet**, ou **échelle de mots** (***Word Ladder Puzzle*** en anglais) est un jeu inventé par Lewis Carroll [1]. La première mention de ce jeu apparaît dans son journal le 12 mars 1878. Le jeu est publié pour la première fois le 29 mars 1879 dans le magazine britannique *Vanity Fair*. Il s'agit de trouver une chaîne de mots reliant deux mots donnés, où à chaque étape les mots ne diffèrent que d'une seule lettre, sans changer la place des lettres. Par exemple, pour relier EXOS à MATH, on peut créer une chaîne de 8 échelons :

EXOS, E**R**OS, **G**ROS, GR**I**S, G**A**IS, **M**AIS, MA**T**S, MAT**H**.

Ce que nous allons chercher ici, ce sont les chaînes les plus courtes qui relient deux mots. Dans la suite de cet article, pour alléger le texte, nous appellerons la chaîne la plus courte une « échelle ».

Donald Knuth, le précurseur

Donald Knuth, le célèbre informaticien de l'université de Stanford, inventeur entre autres de LaTeX, s'est intéressé à ce jeu avec les mots anglais de cinq lettres. Il a construit un graphe dans lequel 5757 mots étaient représentés par des sommets. Deux sommets étaient reliés par une arête s'ils ne différaient que d'une lettre, selon la règle établie par Lewis Carroll. Ce graphe comprenait 14'135 arêtes. Il a ensuite écrit un programme permettant de trouver les échelles entre deux mots donnés en entrée [2].
Nous allons refaire le travail de Knuth, mais avec des mots français, et nous ne nous contenterons pas des mots de cinq lettres...

Mots français de quatre lettres

Commençons notre étude avec les mots de quatre lettres. J'ai utilisé la liste des 2441 mots autorisés au Scrabble®, qui est disponible sur le web [3].

Il faut maintenant représenter le graphe dans le programme : les sommets seront bien évidemment les mots et deux sommets seront reliés par une arête si ces mots ne diffèrent que d'une lettre. Un programme en Python assez facile à écrire (disponible sur [6]) a généré les listes d'adjacences : pour chaque mot on donne la liste des mots voisins :

```
ABAT: [ABOT,AFAT,AXAT,EBAT]
ABBE: [ABEE,ABLE,AUBE]
ABEE: [ABBE,ABER,ABLE,AGEE,ANEE,AXEE]
...
```

```
ZOOS: [ZOBS,ZOES,ZOOM]
ZOUK: [SOUK]
ZUPS: [OUPS,ZIPS]
```

En analysant cette liste, on voit vite que le graphe n'est pas *connexe*, ce qui veut dire que l'on ne peut pas forcément trouver une chaîne entre deux mots quelconques. Il existe en effet 94 mots sans voisins :

```
ACUL AFRO AGHA AIGU AMOK ARUM AZUR BAHT BODY CIAO CLUB COSY DAHU DAUW DESK EDEN ENOL ENVI
EPAR ETOC EVOE EXPO FIQH FISC FOLK FUGU GIRL GLEY GOAL GOLF GOTH GUNZ GYMS HADJ HOPI INFO
INOX INTI INUK ITOU IVRE IXIA JAZZ JEEP KEPI KERN KHOL KICK LABO LAKH LEHM LULU LYNX MAAR
MAMY MUON NOEL OEIL OGAM OHMS ORAL OUAH OUED OUZO OVNI PRAO RUMB SIKH SMOG SNOB SUMO THUG
THYM UBAC UGNI ULNA ULVE UMMA UNAU VOEU VOMI WASP WATT WITZ WURM YAWL YEYE YORK YUAN YUKO
ZARB ZINC ZIZI ZOZO
```

Il existe aussi seize petits groupes de mots qui ne sont pas reliés au plus grand nombre :

quatorze couples

```
TAEL-TAAL, NECK-TECK, ORYX-ONYX, DIBI-BIBI, SLOW-SHOW, SOAP-SWAP, INCH-INCA, YOGA-YOGI,
AFIN-ASIN, JUDO-JUDD, ORDI-ORDO, OBEL-OBEI, EDAM-EXAM, APEX-APAX.
```

et deux triplets

```
AMUI-AGUI-AMMI, BIRR-GRRR-BRRR.
```

Le reste des mots non isolés forme une seule composante connexe de 2313 mots, ce qui signifie que depuis n'importe quel mot de ce groupe, on peut trouver une chaîne allant jusqu'à un autre mot de ce groupe.
En tout, il y a 10'226 arêtes dans ce graphe.

Pour compter le nombre de composantes connexe, nous avons utilisé l'algorithme de marquage récursif suivant (c'est en fait un parcours en profondeur d'un arbre) :

```
PROCEDURE marquer(s,m):
    # marque le sommet s et ses voisins avec la marque m
    marque de s := m
    POUR tous les voisins v de s FAIRE
        SI v n'est pas marqué ALORS
            marquer(v,m)

marque = 0
POUR tous les mots s FAIRE
    SI s n'est pas marqué ALORS
        marque := marque + 1
        marquer(s,marque)
```

Il suffit ensuite de regarder les marques des mots pour dénombrer facilement les composantes du graphe.

Le mot ayant le plus de voisins est `PAIS` ; il en a 29 : `BAIS DAIS FAIS GAIS HAIS JAIS LAIS MAIS NAIS PACS PAFS PAIE PAIN PAIR PAIT PAIX PALS PANS PARS PATS PAYS PLIS POIS PRIS PUIS RAIS SAIS TAIS VAIS`

Voici la répartition du nombre *N* de voisins des mots, pour *N* compris entre 0 et 29 :

0	1	2	3	4	5	6	7	8	9	10	11	12	13	14
94	142	131	183	170	185	160	170	152	140	137	115	107	71	82
15	16	17	18	19	20	21	22	23	24	25	26	27	28	29
89	60	55	50	34	34	25	20	7	16	7	1	2	1	1

Pour trouver une échelle entre deux sommets, il faut implémenter *l'algorithme de Dijkstra*. C'est un algorithme classique de la théorie des graphes [7], que je ne décrirai pas ici. On peut le trouver déjà implémenté sur le web, et l'adapter à nos besoins particuliers. Dans le cas qui nous intéresse, toutes les arêtes

ont un poids de 1, ce qui simplifie (un peu) les choses.

Avec les mots de quatre lettres du Scrabble®, la plus longue échelle compte 20 échelons et relie les mots ISBA et GNOU (en théorie des graphes, on dit que le *diamètre* du graphe est 20). Voici une des 46 possibilités pour cette échelle (on peut facilement créer les 45 autres possibilités sur [4]) :

ISBA, ISSA, ISSU, INSU, INDU, INDE, IODE, RODE, RIDE, AIDE, AIDS, AIES, AGES, AGAS, ADAS, ADOS, ADON, ANON, GNON, **GNOU**.

Il n'y a pas d'autres échelles à 20 échelons que celles reliant GNOU à ISBA.

Pour trouver le diamètre du graphe, nous avons utilisé l'algorithme de *Dijkstra* légèrement modifié. On pourrait naïvement penser qu'il suffit d'appliquer cet algorithme entre toutes les paires de sommets pour ne retenir que la plus longue échelle. C'est possible, mais sera très long en temps de calcul, puisque l'on devra appliquer *Dijkstra* $n^2/2$ fois, où n est le nombre de sommets du graphe. On ira beaucoup plus vite en se rappelant que l'algorithme de *Dijkstra* ne trouve pas seulement le plus court chemin d'un sommet s vers un sommet t, mais tous les plus courts chemins entre le sommet s et les sommets atteignables depuis s. Il suffit donc d'appliquer *Dijkstra* n fois.

De ZERO à CENT

Voici un petit exercice amusant. Partir de ZERO pour arriver à CENT en passant par DEUX, CINQ, SEPT, HUIT, NEUF et ONZE, selon les règles des doublets de Carroll.

La solution la plus courte (40 échelons) :

ZERO, HERO, HERE, HELE, FELE, FEUE, FEUX, **DEUX**, FEUX, FEUE, FEDE, CEDE, CENE, CINE, **CINQ**, CINE, CENE, CENT, SENT, **SEPT**, SERT, SERF, NERF, **NEUF**, NERF, NERE, GERE, GORE, GODE, IODE, INDE, ONDE, **ONZE**, ONDE, INDE, IODE, CODE, CEDE, CENE, **CENT**.

Remarquons au passage que l'échelle de ZERO à CENT n'a que 7 échelons :

ZERO, HERO, HERE, GERE, GENE, CENE, **CENT**.

À la recherche de la plus longue échelle

À notre connaissance, c'est la première fois que l'on recherche la plus longue échelle dans la langue française. Une telle étude a été faite en anglais [5], à l'aide du logiciel *Mathematica*. Selon cet article, il semblerait que la plus longue échelle dans la langue de Shakespeare comporte 46 échelons, avec des mots de 7 lettres, en allant de GIMLETS à THEEING.
Qu'en est-il en français ?

Une fois les programmes écrits pour les mots de quatre lettres, il n'y a pas beaucoup de travail à faire en plus pour analyser les graphes des mots plus longs. Je me suis arrêté à onze lettres, car, d'après le tableau obtenu ci-après, il est fort probable que nous ne trouverons pas d'échelles plus longues au-delà.

Après quelques heures de calcul, il ressort que l'une des plus longues échelles en français (il y en a plusieurs) comporte 66 échelons et relie SERVANTE à FRESSURE :

SERVANTE, SERRANTE, SERRANTS, FERRANTS, FERMANTS, FERMENTS, SERMENTS, SERGENTS, SERGENTE, SERGETTE, SERRETTE, SARRETTE, BARRETTE, BARBETTE, BARBOTTE, BARBOTEE, BARBOTES, BARBATES, BARDATES, BORDATES, CORDATES, CORSATES, CORSETES, CORSETAS, CORSERAS, CORDERAS, COUDERAS, COUTERAS, CONTERAS, CONFERAS, CONFIRAS, CONFINAS, CONFINES, CONFIEES, CONVIEES, CONVIENS, CONTIENS, CONTIONS, COITIONS, CUITIONS, CUISIONS, CRISIONS, CRISSONS, CRESSONS, PRESSONS, PRESSENS, PRESSEES, DRESSEES, DROSSEES, CROSSEES, CRASSEES, CLASSEES, CLISSEES, CLIPSEES, CLIPPEES, CLIPPERS, CLIPPERA, CLIPSERA, CLISSERA, CRISSERA, TRISSERA, TRESSERA, PRESSERA, PRESSURA, PRESSURE, **FRESSURE**

Le tableau ci-après résume l'analyse des graphes obtenus en fonction de la longueur des mots.

	Mots (tirés de [3])	Mots isolés	Composantes connexes*	Arêtes	Nombre max d'échelons	Doublets « optimaux »
4 lettres	2'441	94	1 cardinalité max: 2'313	10'226	20	ISBA GNOU
5 lettres	7'483	594	107 cardinalité max: 6'625	23'638	25	ISBAS SMOLT
6 lettres	17'035	1'914	623 cardinalité max: 12'988	39'720	50	UREIDE TCHANS
7 lettres	30'633	4'604	2'015 cardinalité max: 17'248	52'936	49	GOBEURS UREIDES
8 lettres	45'642	9'086	4'557 cardinalité max: 15'096	60'627	66	SERVANTE FRESSURE
9 lettres	56'573	14'828	7'438 cardinalité max: 3'999	57'408	61	EVASASSES SERINAMES**
10 lettres	59'526	19'129	9'405 cardinalité max: 1'699	45'952	39	REPORTIONS SERINERAIS
11 lettres	54'442	20'761	9'591 cardinalité max: 605	31'832	33	EVASASSIONS GIVRASSIONS

* sans compter les mots isolés
** à noter que ces deux mots ne font pas partie de la plus grande composante connexe, mais de la deuxième en taille, qui contient 2684 mots.

Conclusion

Ces graphes créés à partir de mots sont un terrain de jeux idéal pour la théorie des graphes : ils nécessitent d'utiliser l'ordinateur pour résoudre des problèmes qui ne sont ni trop simples, ni trop compliqués.
On pourrait facilement trouver des variantes du jeu inventé par Lewis Carroll : par exemple permettre de modifier la place des lettres, changer deux lettres au lieu d'une, etc. Nul doute que l'on tomberait sur d'autres problèmes intéressants.

Références

[1] Lewis Carroll, *Doublets, a word-puzzle*, (1879), <http://www.archive.org/details/doubletsawordpu00dodggoog >

[2] Ces informations proviennent du premier chapitre du livre de Knuth *The Stanford GraphBase : A Platform for Combinatorial Computing* (Addison Wesley, 1993).

[3] « Mots du Scrabble », <www.motsduscrabble.com>

[4] Nicolas Graner, « Doublets de Carroll », <http://graner.net/nicolas/divers/doublets.php>, mars 2011

[5] Jon McLoone, « The Longest Word Ladder Puzzle Ever » , <http://blog.wolfram.com/2012/01/11/the-longest-word-ladder-puzzle-ever>

[6] « Les doublets de Lewis Carroll », <www.nymphomath.ch/graphes/doublets/>, 2012

[7] Didier Müller, « Introduction à la théorie des graphes », Cahier de la CRM no 6, 2012.
Aussi disponible en ligne sur <www.nymphomath.ch/graphes/>