N° d'ordre : 01/ 2006-M/ MT

REPUBLIQUE ALGERIENNE DEMOCRATIQUE ET POPULAIRE

MINISTERE D'ENSEIGNEMENT SUPERIEURE ET DE LA RECHERCHE SCIENTIFIQUE

UNIVERSITE DES SCIENCES ET DE LA TECHNOLOGIE
« HOUARI BOUMEDIENNE »
FACULTE DES MATHEMATIQUES

MEMOIRE

Présenté pour l'obtention du diplôme de MAGISTER

EN : MATHEMATIQUES

Spécialité : Recherche Opérationnelle : Mathématiques de Gestion

Par :

Mr Hakim HARIK

Sujet

ELEMENTS DE LA THEORIE DES GRAPHES ET APPLICATIONS DANS LE DOMAINE DE L'INFORMATION

Soutenu publiquement le : 06 Mars 2006, devant le jury composé de :

Mr BOUROUBI Sadek	Maître de Conférences USTHB	Président
Mr AIT HADDADENE Hacène	Maître de Conférences USTHB	Directeur de Thèse
Mme BOUCHEMAKH Isma	Maître de Conférences USTHB	Examinateur
Mr MOULAI Mustapha	Maître de Conférences USTHB	Examinateur
Mr YALAOUI Bilal	Chargé de Recherche CERIST	Invité

A la mémoire de l’être qui m’est très cher, à mon père

A la mémoire de l’être qui m’est très cher, à mon père

A la mémoire de l’être qui m’est très cher, à mon père

REMERCIEMENTS

Je voudrais remercier profondément mon directeur de thèse Monsieur Hacène Ait Haddadene (maître de conférences à la faculté des mathématiques, USTHB) de m'avoir proposé ce sujet et de m'avoir encadré pendant tout ce temps. Je le remercie vivement pour son aide, ses conseils, ses orientations fondamentales tout au long de l'avancement de cette thèse et pour la patience et la générosité avec lesquelles il a su guider mes recherches. Mes reconnaissances pour lui de m'avoir le premier permis de découvrir le monde de la recherche et de m'avoir fait confiance dans ce domaine.

Je remercie Monsieur Sadek Bouroubi (maître de conférences à la faculté des mathématiques, USTHB) qui me fait l'honneur de présider ce jury.

Je remercie également Madame Isma Bouchemakh (maître de conférences à la faculté des mathématiques, USTHB) et Monsieur Mustapha Moulai (maître de conférences à la faculté des mathématiques, USTHB) qui ont bien accepté de faire partie du jury de cette thèse, je les remercie également pour l'intérêt qu'ils ont porté à ce travail.

Je remercie Monsieur Bilal Yalaoui, attaché de recherche au sein du CERIST, qui a été un interlocuteur privilégié tout au long de cette période et qui m'a lancé sur les premières pistes de recherche dans un domaine passionnant dont est issu ce travail. Les discussions que nous avons eu m'ont guidé dans mes recherches et m'ont permis de préciser bon nombre d'idées encore floues. Son expérience et sa disponibilité m'ont été d'une aide précieuse. Je le remercie encore une fois pour sa sympathie et sa patience.

Je tiens aussi à remercier Monsieur Madjid Dahmane, Directeur de la division Recherche et Développement en Science de l'Information au CERIST, qui m'a gentiment accueilli durant cette période dans son laboratoire. Il trouve ici toute ma gratitude pour son enseignement, son aide qu'il n'a pas hésité à m'apporter et pour ses précieux conseils.

Mes remerciements s'adressent également à tous ceux qui, au sein du CERIST, ont participé d'une façon ou d'une autre à la concrétisation de ce travail et je pense surtout à Mr Mounir Bouter, Mr Bachir Idri et Mr Mourad Guezou.

A tous mes amis pour leur disponibilité et leur soutien indéfectible, je pense plus particulièrement à Mr Yahia Tighilt, Mr Mourad Ben Maouche et Mr Djamel Sator.

Je dédie ce travail à ma mère, mes frères et mes sœurs que je trouve toujours dans les moments difficiles et à tous mes oncles et tantes pour leur encouragement pendant tout ce temps.

TABLE DES MATIERES

“L’information demeure essentielle à la bonne marche de la société. Et il n’y a pas de démocratie possible sans un bon réseau de communication et sans le maximum d’information. C’est grâce à l’information que l’homme vit comme un homme libre”

IGNACIO RAMONET

“Medias, sociétés et démocratie : l’ère du soupçon”
Dans “le monde diplomatique”- n°446 (mai 1991)

Introduction générale

INTRODUCTION GENERALE

Avec l'essor de la nouvelle technologie de l'information, le nombre de documents disponible dans des bases de données documentaires croit d'une manière exponentielle surtout avec le développement des communications électroniques, ce qui entraîne une surinformation de l'utilisateur qui se trouve submergé par la quantité d'information, il n'a plus ni le temps ni les ressources cognitives pour faire face à un tel volume d'informations, il se trouve dans une situation de saturation. L'utilisateur n'est donc pas capable d'analyser ou d'appréhender ces informations dans leur globalité, ceci lui implique une démotivation et un abandon de sa consultation dans la plupart des cas. Il devient alors indispensable de proposer de nouvelles approches pour remédier à cette situation et de surmonter ainsi à la problématique de surinformation, ceci par l'analyse des masses importantes d'informations qui produit une représentation de son contenu, en faisant apparaître les informations principales des contenues et de concevoir des outils capables de comprendre des corpus textuels par la représentation simplifiée du contenu de ce corpus qui offre notamment une aide à une consultation [Rajman & al, 1997], il ne s'agit pas de comprendre tout le texte mais d'extraire un certain nombre d'informations pertinentes par rapport à ce texte.

Ainsi, au-delà de la recherche d'information qui se contente de sélectionner des documents dans une base documentaire à partir d'une requête donnée en pointant sur les textes eux-mêmes et non sur le contenu des textes, c'est-à-dire sur ce dont il est question et laissant le traitement du contenu à l'utilisateur, l'un des enjeux majeurs aujourd'hui consiste à développer des outils permettant l'exploration du contenu ceci par l'analyse de contenu textuel qui s'intéresse aux significations du texte, aux indications qu'il apporte sur le sujet traité.

Notre travail s'inscrit dans cette optique. Sachant que la représentation graphique est un outil puissant et un excellent vecteur d'analyse des données complexes [Tufte, 1983], [Tufte, 1990], [Tufte, 1997] alors la méthode que nous proposons se base sur l'utilisation d'éléments de la théorie des graphes pour représenter le contenu d'un corpus textuel sous forme d'une cartographie d'information. L'idée de base est de permettre à l'utilisateur de prendre en compte les liens qui existent entre les différentes notions représentées. Ce qui permet de synthétiser des informations sous une forme facile à interpréter et à exploiter et qui conduit à une compréhension simple de la structure du corpus étudié sous forme d'une carte de liens sémantiques entre les termes représentatifs du contenu de ce corpus. Une des caractéristiques de cette représentation est que le support graphique permet d'aider une communauté d'utilisateurs, travaillant sur une thématique donnée, dans ses consultations d'appréhender et de visualiser globalement l'ensemble des termes représentatifs du corpus avant de porter l'attention sur la partie intéressante dans le graphe.

Ce mémoire est développé en trois chapitres :

Le premier chapitre est consacré aux concepts fondamentaux utilisés dans ce manuscrit. On présente une vue globale des éléments fondamentaux et aux concepts de base de la théorie des graphes et de la science de l'information. Ainsi que l'impact de la théorie des graphes dans le domaine de la science de l'information par les applications divers et variées de cette théorie dans ce domaine pour le traitement de la production de l'information.

Le chapitre deux est consacré à étudier le problème du clustering par le principe du partitionnement d'un graphe. Après avoir présenté le modèle du graphe d'association de termes, on propose trois approches du partitionnement de ce graphe à partir de la densité d'un sommet, de l'importance d'une arête et la triangulation du graphe d'association de termes afin de regrouper les termes en ensembles homogènes.

Le chapitre trois est consacré à la cartographie d'un corpus textuel basé sur le réseau de termes associé qu'on a proposé. Ce réseau permet de montrer les relations d'influences entre les termes. On propose un algorithme de détermination d'un ensemble appelé générique représentatif qui permet de donner une vue globale sur le contenu du corpus textuel. Nous terminons par proposer une cartographie d'un corpus textuel basé sur les résultats qu'on a obtenus.

Enfin, nous donnons une conclusion et quelques perspectives.

Chapitre 1

Généralités et notions de base

1.1. Notations et quelques concepts de base de la théorie des graphes

La théorie des graphes est le domaine des mathématiques initialement développé par Léonard Euler (1707-1783). Le problème des ponts de Koenisberg en est l'application la plus célèbre et est le premier résultat formel de la théorie des graphes. Le problème posé était le suivant. Deux îles A et D sur la rivière Pregel à Koenisberg (alors capitale de la Prusse de l'Est, aujourd'hui rebaptisée Kaliningrad) étaient reliées entre elles ainsi qu'aux rivages B et C à l'aide de sept ponts (désignés par des lettres minuscules) comme le montre la figure 1.1.

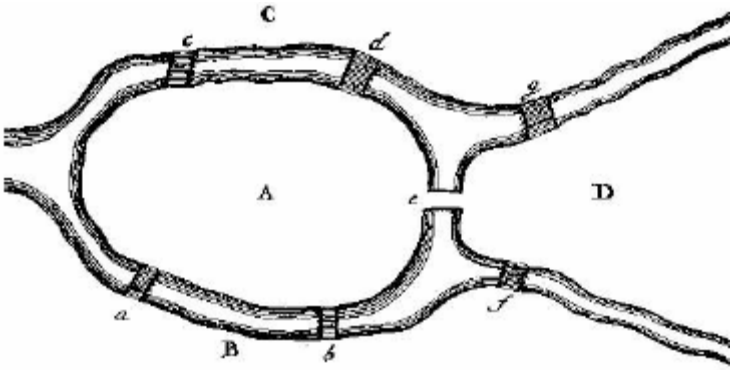


Figure1.1

Le problème posé consistait, à partir d'une terre quelconque A, B, C, ou D, à traverser chacun des ponts une fois et une seule et à revenir à son point de départ (sans traverser la rivière à la nage !). Le problème des ponts de Koenisberg est identique à celui consistant à tracer une figure géométrique sans lever le crayon et sans repasser plusieurs fois sur un même trait. L'histoire veut que Léonard Euler, en visite dans cette ville, ait eu à résoudre le problème qui préoccupait fortement ses habitants. Il modélisa les quartiers de cette ville sous la forme d'un graphe (figure 1.2) et il démontra que quel que soit le quartier de départ, on ne pouvait revenir à ce quartier en n'empruntant qu'une seule fois le même pont. Ainsi, Euler démontra que ce problème n'a pas de solution et il généralisa ce problème à l'étude des graphes, cherchant notamment à répondre à la question suivante : Peut-on circuler sur le graphe à partir d'un sommet en empruntant une fois et une seule chaque arête ?

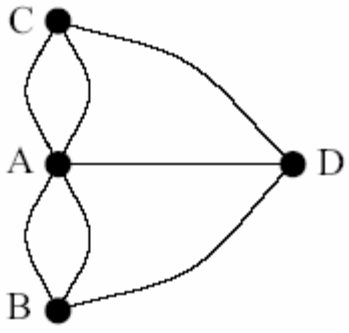


Figure 1.2

La théorie des graphes s'est alors développée dans diverses disciplines telles l'analyse de circuits électriques, la chimie, la biologie, les sciences sociales, ceci sous l'impulsion de chercheurs motivés par la résolution de problèmes concrets et elle connaît un essor depuis le début du XX^{eme} siècle où elle constitue une branche à part entière des mathématiques, grâce aux travaux de König, Menger, Cayley puis de Berge et d'Erdös ce qui marque sans doute l'avènement de l'ère moderne de la théorie des graphes par l'introduction d'une théorie unifiée et abstraite rassemblant de nombreux résultats épars dans la littérature. Depuis, cette théorie a pris sa place, en subissant de très nombreux développement au sein d'un ensemble plus vaste d'outils et de méthodes sous l'appellation « mathématiques discrètes ».

De manière générale, un graphe permet de représenter la structure, les connexions d'un ensemble complexe en exprimant les relations entre ses éléments : réseau de communication, réseaux routiers, interaction de diverses espèces animales, circuits électriques… il constitue donc une méthode de pensée qui permet de modéliser une grande variété de problèmes en se ramenant à l'étude de sommets et d'arêtes.

Les derniers travaux en théorie des graphes sont souvent effectués par des informaticiens, du fait de l'importance qu'y revêt l'aspect algorithmique.

1.1.1. Définitions et notation principales

Un graphe non orienté (ou plus simplement un graphe) est la donnée d'un couple (V, E), où V est un ensemble fini d'éléments distincts appelés sommets et E (ou bien E (V)) un ensemble de paires d'éléments de V dit ensemble d'arêtes, on le notera $G = (V, E)$. L'ordre d'un graphe est le nombre de sommets, on notera n l'ordre du graphe et m le nombre de ses arêtes. Il peut y avoir plusieurs arêtes reliant deux sommets u et v de G, leur nombre représente la multiplicité de l'arête $\{u, v\}$. **Une boucle** est une arête dont les extrémités sont confondues.

Un graphe simple est un graphe sans boucle et tout couple de sommets est relié par au plus une arête.

Dans le plan, on représente un graphe G par une figure géométrique où les sommets sont représentés par des petits cercles (ou des points) et les arêtes par des traits joignant les points qui représentent leurs extrémités (Figure1.3).

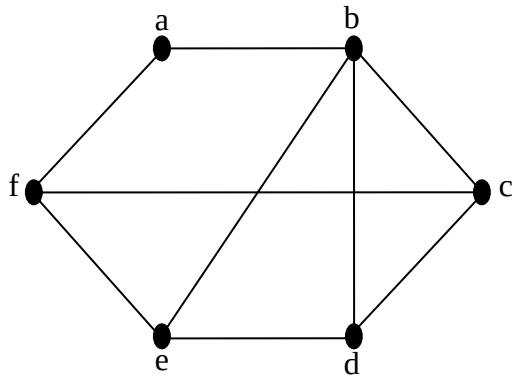


Figure1.3- *Un graphe simple* $G = (V, E)$.

Avec:

V={a, b, c, d, e, f}.

E={{a, b}, {a, f}, {b, c}, {b, d}, {b, e}, {c, d}, {c, f}, {d, e}, {e, f}}.

Dans tout ce qui suit nous ne considérons que des graphes simples.

Pour une arête {*u, v*} de *G*, qu'on note *uv*, on dit que :
- *uv* est **incidente** à *u* (resp. *v*),
- *u* et *v* sont **les extrémités** de *uv*,
- *u* et *v* sont **adjacents**,
- *u* (resp. *v*) est **un voisin** de *v* (resp. *u*).

Deux arêtes sont **adjacentes** si elles ont une extrémité commune.

L'ensemble des voisins d'un sommet *u* donné, est *N* (*u*) :

$$N(u) = \{v \in V / uv \in E\}.$$

Le degré d'un sommet *v* d'un graphe *G* est le nombre $d_G(v)$ (*d(v)* s'il n'y a pas de confusion), d'arêtes incidentes à *v* dans *G*, en d'autres termes, $d_G(v) = |N(v)|$.

Pour tout graphe *G*, nous avons :

$$\sum_{u \in V(G)} d_G(u) = 2|E(G)|.$$

L'ensemble des voisins d'un ensemble $A \subset V$ donné, est $N(A)$:

$$N(A) = \{v \in V\text{-}A / \exists\, u \in A : uv \in E\}.$$

Un sous graphe $G' = (V', E')$ de $G = (V, E)$ est un graphe dont l'ensemble des sommets V' est un sous-ensemble de *V*, et dont l'ensemble des arêtes E' est un sous-ensemble de *E* tel que toute arête de E' joint deux sommets de V' c'est-à-dire $E' \subset E \cap (V' \times V')$. Dans le cas où *E'* contiendrait toutes les arêtes qui relient les sommets de V' dans *G*, on dira que G' est **induit** par V' et on le notera $G_{V'}$. Si $V' = V$, on dit que G' est **un graphe partiel** de *G*.

La figure1.4, montre un sous graphe induit $G' = (V', E')$ du graphe $G = (V, E)$ de la figure 1. 3, avec $V' = \{a, b, c, d\}$ et $E' = \{\{a, b\}, \{b, c\}, \{b, d\}, \{c, d\}\}$

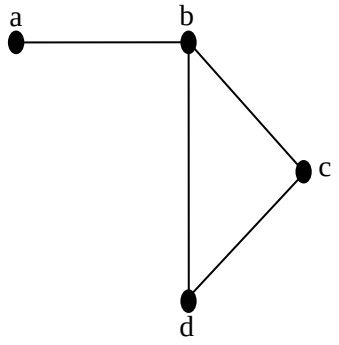


Figure1.4- *Sous graphe induit G' de G*

La figure1.5 montre un graphe partiel $G'' = (V'', E'')$ du graphe $G = (V, E)$ de la figure 1. 3, avec $V'' = \{a, b, c, d, e, f\}$ et $E'' = \{\{a, b\}, \{b, c\}, \{b, d\}, [b, e]\{c, d\}, \{c, f\}\}$

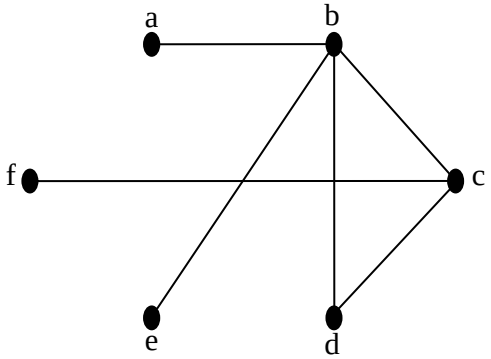


Figure1.5- *Sous graphe partiel G'' de G*

Une chaîne $[u_0, e_1, u_1, e_2, ..., u_{k-1}, e_k, u_k]$ dans un graphe est une séquence alternée de sommets et d'arêtes distincts, tels que $\forall i = 0,...,k-1 : u_i$ est adjacent à u_{i+1} par l'arête e_{i+1} la longueur de cette chaîne est *k*, c'est le nombre d'arêtes qui la constitue ; v_0 et v_k sont appelés les extrémités de cette chaîne. Elle est notée, par abréviation, $(u_0, u_k)-$chaîne. Pour $k \geq 2$, les sommets $u_1, \ldots, u_{k-1}$ sont les sommets internes de la $(u_0, u_k)-$chaîne. Si les sommets utilisés sont tous distincts elle sera dite élémentaire.

Un cycle est une chaîne dont les deux extrémités sont confondues. Un cycle qui ne rencontre pas deux fois le même sommet est dit **élémentaire**. Un cycle qui n'utilise pas deux fois la même arête est dit **simple**.

Une corde dans un cycle (respectivement dans une chaîne) est une arête reliant deux sommets non consécutifs du cycle (respectivement de la chaîne).

Un trou est un cycle sans corde de longueur au moins cinq, **un anti-trou** est le complémentaire d'un trou.

Un graphe est **connexe** si l'on peut atteindre n'importe quel sommet à partir d'un sommet quelconque en parcourant différentes arêtes. C'est-à-dire si pour toute paire de sommets *u* et *v*, il existe une (*u, v*)-chaîne les reliant. Dans le cas contraire, *G* sera dit non connexe et pourra s'écrire comme l'union disjointe de sous graphes induits connexes $G_1, G_2, \ldots, G_p$ appelés les composantes connexes du graphe *G,* engendrés par les sous-ensembles $V_1, V_2, \ldots, V_P$.

Le nombre *p* est appelé nombre de connexité du graphe.

1.1.2. Quelques paramètres et invariants d'un graphe

Une clique dans un graphe simple *G* est un sous-ensemble de sommets deux à deux adjacents. Une clique de *n* sommets possède $\frac{n(n-1)}{2}$ arêtes. Si *V* (*G*) est une clique, le graphe *G* est dit **complet** et est noté K_n (Figure1.6).

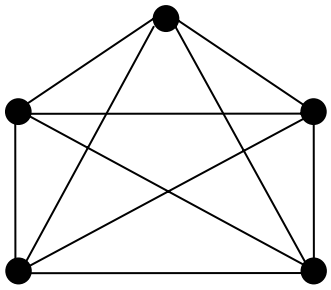


Figure1.6- *Graphe simple complet d'ordre* $5(K_5)$.

Nous désignons par $\vartheta(G)$: la taille minimale d'une partition de *V* en cliques.

Un arbre est un graphe connexe sans cycles. Il est également définit comme un graphe connexe dans lequel il existe une chaîne et une seule entre toute paire de sommets quelconques (Figure1.7).

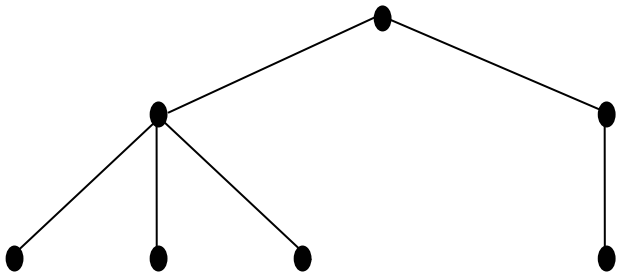


Figure1.7- *Un arbre*.

La densité d'un sommet *u* calcule pour chaque sommet, elle mesure le nombre d'arêtes entre les voisins d'un sommet par rapport au nombre d'arêtes qui pourraient exister. Ainsi, pour un sommet *u*, on peut définir sa densité notée *De* (*u*) par :

$$De(u) = \frac{|E(N(u))|}{\binom{d(u)}{2}} = \frac{2 \times |E(N(u))|}{d(u) \times (d(u)-1)}$$

Dans l'exemple suivant, on remarque que dans le voisinage du sommet *u*, il existe quatre arêtes, parmi les six arêtes possibles ainsi la densité du sommet *u* est $De(u) = \frac{2}{3}$ (Figure1.8).

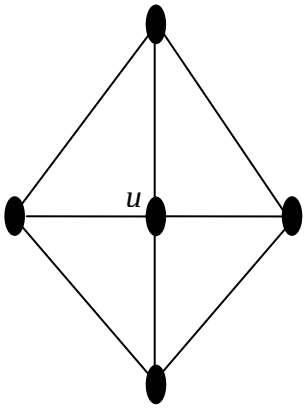


Figure1.8- *Indice de clustering du sommet u,* $De(u)=\frac{2}{3}$

On définit la densité *De* d'un graphe par la valeur moyenne de la densité de tous ses sommets, c'est-à-dire :

$$De=\frac{\sum_{u\in V} De(u)}{|V|}$$

Reprenons l'exemple précèdent, on remarque bien que la densité pour chaque sommet est égal à $\frac{2}{3}$, ainsi la densité de graphe est $De=\frac{2}{3}$ (Figure1.9).

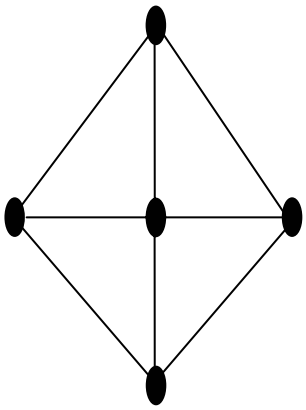


Figure1.9- *Indice de clustering du graphe* $De=\frac{2}{3}$

La valeur prise par la densité variera entre zéro (cas d'un arbre) et un (cas d'un un graphe complet).

1.1.3. Graphe triangulé

La notion de graphe triangulé a été introduite par Hajnal et Suranyi [Hajnal & al, 1958].

Définition

Un graphe $G = (V, E)$ est dit triangulé si tout cycle dans *G* de longueur supérieur ou égale à 4 admet une corde.

Dans la littérature anglo-saxonne, on trouve les dénominations de triangulated graphs, chordal graphs.

La figure1.5 montre un exemple d'un graphe triangulé

Propriété

Tous sous-graphe d'un graphe triangulé est triangulé

Plusieurs caractérisations des graphes triangulés ont été proposées. Avant de citer les plus intéressantes, rappelons les définitions suivantes :

Définitions :

Soit un graphe $G = (V, E)$.

- Un sommet **simplicial** de *G* est un sommet dont le voisinage est une clique.
- **Un ordre d'élimination simplicial** de *G* est un ordre [v_1,…., v_n] de ses sommets, tel que tout v_i est un sommet simplicial dans le sous-graphe *G* [v_i,….,v_n].

Dans le graphe de la figure1.5 , [f, e, a, c, b, d] est un ordre d'élimination simplicial.

Théorème :

Soit $G = (V, E)$ un graphe, les propriétés suivantes sont équivalentes :

(i) *G* est triangulé ;

(ii) Tout sous-graphe induit de *G* contient un sommet simplicial (Dirac [Dirac, 1961])

(iii) Tout sous-graphe induit de *G* est soit une clique, soit il contient deux sommets simpliciaux non adjacents (Dirac [Dirac, 1981]).

(iv) *G* admet un ordre d'élimination simplicial de ses sommets (Fulkerson et Gross [Fulkerson & al, 1965]).

Rose, Tarjan et Lucker [Rose & al, 1976] ont présenté un algorithme polynomial de reconnaissance des graphes triangulés en utilisant l'ordre d'élimination simplicial, de même F. Gavril [Gavril, 1972] a présenté des algorithmes polynomiaux pour les problèmes d'optimisation.

Définition

Le graphe $G' = (V', E')$ est une triangulation de *G* si G' est un graphe triangulé, tel que $V = V'$ et $E \subset E'$.

Ainsi, la forme d'une triangulation d'un graphe quelconque $G = (V, E)$ est $G' = (V, E \cup F)$ avec *F* est l'ensemble d'arêtes qui sont ajoutés au graphe *G* pour le triangularisé.

1.1.4. Graphe orienté

Un graphe orienté $G = (V, U)$ est un graphe dont toutes les arêtes sont orientées. Les éléments de *U* sont appelés **arcs** de *G* (Figure1.10).

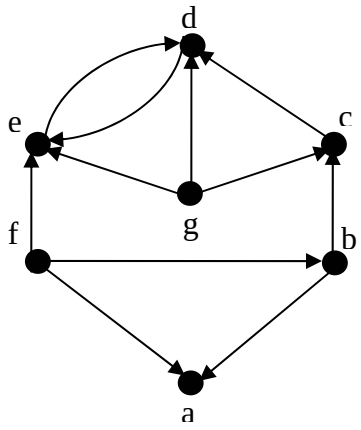


Figure1.10- *Un graphe orienté*

Avec:

V={a,b ,c ,d, e, f,g}, *E*={(f, a), (b, a), (f, b), (b, c), (c, d), (g, d), (g, c), (d, e), (e, d), (f, e), (g, e)}.

Pour un arc $e = (u, v)$, l'élément *u* est son **extrémité initiale**, et *v* est son **extrémité terminale**, les sommets *u* et *v* sont dits alors **voisins**. Deux arcs sont **adjacents** s'ils ont une extrémité commune. De plus, ils sont consécutifs si l'extrémité initiale de l'un est l'extrémité terminale de l'autre.
Soit *u* un sommet du graphe *G*. Le sommet *v* est **un successeur** de *u*. s'il existe un arc ayant son extrémité initiale en *u* et son extrémité terminale en *v*. L'ensemble des successeurs de *u* se note : $\Gamma^+(u)$ et pour un ensemble $A \subset V$ on pose $\Gamma^+(A) = \{v \in V - A / \exists u \in A, uv \in U\}$

De même, le sommet *v* est **un prédécesseur** de *u*, s'il existe un arc ayant son extrémité initiale en *v* et son extrémité terminale en *u*. L'ensemble des prédécesseurs de *u* se note $\Gamma^-(u)$ et pour un ensemble $A \subset V$ on pose $\Gamma^-(A) = \{v \in V - A / \exists u \in A, vu \in U\}$

L'ensemble des sommets voisins du sommet *u* se note $\Gamma(u) = \Gamma^+(u) \cup \Gamma^-(u)$
Dans la figure1.10 on a :

$$\left.\begin{array}{l}\Gamma^+(d) = \{e\} \\ \Gamma^-(d) = \{c, e, g\}\end{array}\right\} \Rightarrow \Gamma(d) = \Gamma^+(d) \cup \Gamma^-(d) = \{c, e, g\}.$$

Si un sommet *u* est l'extrémité initiale d'un arc *e* = (*u, v*), on dit que l'arc est incident à *u* vers l'extérieur. Dans un graphe *G,* le nombre d'arc de la forme (*u, v*) se note $d^+(u)$, et s'appelle le demi degré extérieur de *u.* De même le nombre d'arc de la forme (*v, u*) se note $d^-(u)$, et s'appelle le demi degré intérieur de *u*. $d(u) = d^+(u) + d^-(u)$ est le nombre d'arcs ayant une extrémité en *u* (chaque boucle étant comptée deux fois), et s'appelle le degré de *u.*

Un chemin d'un graphe orienté *G* est une chaîne où deux arcs consécutifs sont dans le même sens**.**

Un circuit est un chemin tel que les deux sommets aux extrémités coïncident

Un ensemble $F \subset V$ est **une composante fortement connexe** dans un graphe orienté $G = (V,U)$ si pour n'importe quel deux sommets *u* et *v* dans *F,* il existe un chemin de *u* vers et *v* et un chemin de *v* vers *u*.

1.1.5. Complexité algorithmique

Un algorithme de résolution d'un problème (*P*) donné, est une procédure décomposable en opérations élémentaires, transformant une chaîne de caractères représentant les données de n'importe quel exemple du problème (*P*) en une chaîne de caractères représentant les résultats de (*P*).

Ainsi, un algorithme ne consiste pas à résoudre seulement un problème unique mais toute une classe de problèmes qui ne différent que par les données mais gouvernées par les mêmes prescriptions.

Un algorithme est dit polynomial ou efficace si le nombre d'opérations nécessaires pour résoudre un problème est borné par une fonction polynomial d'un paramètre caractérisant la taille du problème.

Un problème est dit polynomial ou appartenant à la classe (P) s'il existe un algorithme polynomial pour le résoudre. On dit que les problèmes de la classe (P) sont faciles.

Un problème de décision est un problème qui peut se formuler comme une question à laquelle la réponse est soit oui, soit non.

Un problème de décision est dit NP[non déterministe polynomial] si dans le cas où la réponse est affirmative, on peut produire un certificat qui permet de vérifier en temps polynomial la réponse donnée.

Un problème dans NP est dit NP- complet si tout problème dans NP peut se transformer en ce problème en temps polynomial. Ainsi, l'existence d'un algorithme polynomial pour résoudre un problème NP-Complet entraînerait immédiatement l'existence d'un algorithme polynomial pour résoudre tout problème dans NP.

La classe des problèmes NP peut donc se partitionner en trois en sous – classes, les problèmes de la classe P, les problèmes NP-complets et les autres [Sakarovitch, 1983] (Figure1.11).

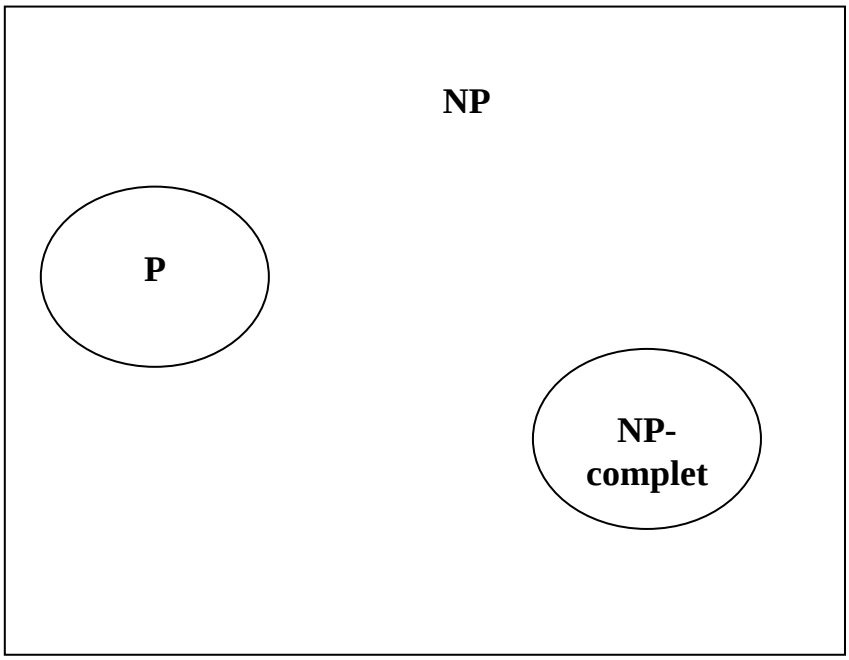


Figure1.11

1.2. Information et science de l'information

L'information est une notion d'actualité qui joue un rôle capital, voire vitale grâce à son importance pour toute communauté scientifique, elle a fait l'objet de nombreuses études par des disciplines très diverses, telles que les sciences de la communication, la linguistique, et les sciences cognitives, pour n'en citer que quelques-unes [Paradis, 1996].

La définition de ce concept est présente dans un grand nombre de travaux mais dans ce mémoire, nous nous sommes basés notamment sur les points de vue adoptés par Davenport et Prusak [Davenport & al, 1998] pour illustrer ce que l'information.

1.2.1. L'information

Avant de définir le concept « information », il est bon de voir ce que peut être la notion de donnée qui est sa matière première et la notion de base pour l'information.

Une donnée, ce n'est rien d'autre qu'un signe ou qu'un symbole. Il s'agit d'un élément brut, qui n'a pas encore été interprété, mis en contexte. Une donnée est un fait discret, un élément fondamental et objectif, qualitatif ou quantitatif, servant de base à un raisonnement.

Par exemple, lorsqu'un client se rend en voiture à une station d'essence pour y faire le plein, cette transaction peut être partiellement décrite par des données : l'heure à laquelle il a rempli son réservoir d'essence, la quantité exprimée en litres de carburant consommé, le montant à payer, etc. Ces données restent toutefois des chiffres bruts, des symboles qui ne nous disent aucunement pourquoi le client s'est rendu dans cette station service particulière, ni s'il reviendra un jour. Ces données sont totalement incapables d'évaluer le service offert ou encore si la station est en train de perdre de l'argent ou de croître….

Toutes les organisations ont besoin de données. Les banques, les compagnies d'assurances…etc, en sont des exemples évidents. L'enregistrement ainsi que la gestion de ces millions de symboles et caractères représente un véritable challenge. En résumé, les données décrivent seulement une partie de ce qui se passe lors d'une transaction. Elles ne fournissent pas un jugement ou une interprétation de la situation et ne peuvent par conséquent constituer la base d'une prise de décision orientée vers une action à entreprendre.

L'information est un ensemble de données organisées pour donner forme à un message résultant d'un contexte donné, c'est une donnée interprétée. En d'autres termes, la mise en contexte d'une donnée crée de la valeur ajoutée pour constituer une information et lui donner une signification.

Considérons la donnée suivante: "35°C". Il s'agit bien d'un élément brut en dehors de tout contexte. Nous savons que cette donnée représente une température mais elle ne possède pas de réelle valeur. Est-ce la température de l'air ambiant? De l'eau du robinet? Ou d'un quelconque objet? Si nous remettons cette donnée dans son contexte, par exemple le bulletin des prévisions météorologiques pour la ville d'Alger, nous créons de la valeur. Nous savons désormais à quoi correspond cette donnée initiale: la température sera de 35°C demain à Alger. Nous sommes passés d'un élément brut à un fait, d'une donnée à une information.

Nous pouvons donc décrire une information comme une association significative et subjective d'un ensemble ou d'une collection de données organisées, représentée par des signes et symboles qui sont des éléments du langage (signe alphabétique, mot, signe de ponctuation), inscris sur un support et visant à transmettre un message d'un émetteur à un récepteur. Elle est supposée changer la façon dont le récepteur perçoit quelque chose, avoir un impact sur son jugement et son comportement. Le but est de l'informer; il s'agit de données qui font une différence. Le mot "inform" signifie originellement "donner forme à" [Dahmane, 1990]. L'information va donc modifier la "forme" de la personne qui reçoit l'information. Du moins elle va bousculer les choses dans sa façon de voir les choses sur un sujet particulier. Il s'ensuit logiquement que le récepteur (et non l'émetteur) décidera si le message reçu constitue véritablement une information - est ce qu'il l'informe réellement.

On remarque bien que cette définition diffère bien de celle introduite dans les travaux de Claude E. Shannon et de Warren Weaver sur l'échange d'information dans un processus de communication [Shannon & al, 1949] qui furent parmi les premiers essais modernes tentant de cristalliser la notion d'information, dont les applications ont permis les progrès actuels des télécommunications.

En effet, Shannon décrit l'information comme un flux physique circulant entre un émetteur et un récepteur lors d'un processus de communication. Le processus de communication (Figure1.12 tirée de [Shannon & al, 1949]) est composé d'une source d'information qui choisit d'abord un message parmi un ensemble de messages possibles à partir d'une source d'information. Ce message est ensuite encodé par l'émetteur en un signal qui sera transmis au récepteur par le biais d'un canal de communication. Le récepteur à son tour change le signal reçu en message et l'achemine à sa destination.

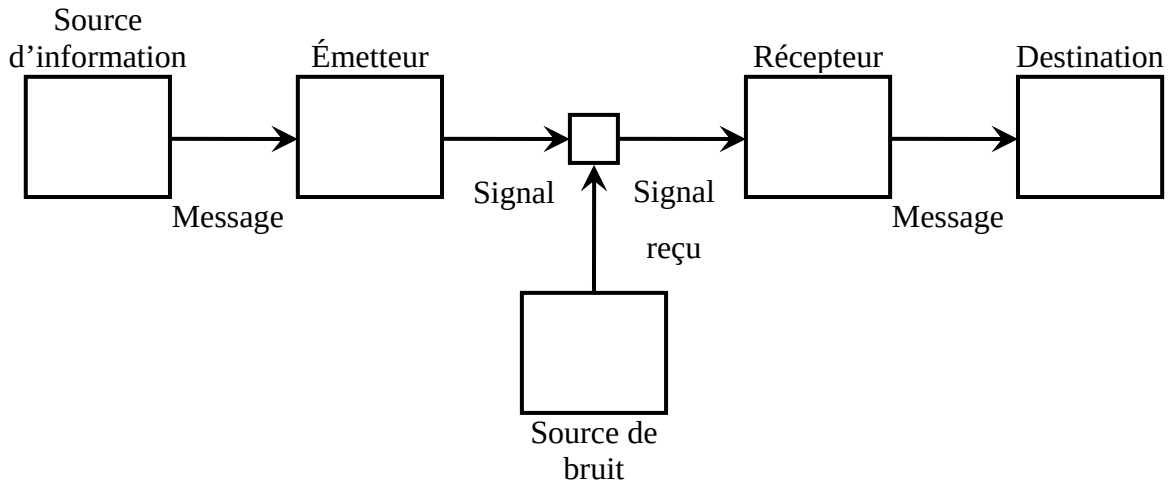


Figure1.12- *Un système de communication*

Ces travaux négligent délibérément tous les aspects sémantiques de l'information, c'est-à-dire relevant de la signification des messages, ils ne couvrent en fait qu'un aspect, celui de la transmission de messages entre deux agents. Delà, au sens de Shannon, l'information est fondée sur une description physique, elle est plus une mesure qu'une entité. A ce titre, les chercheurs sur l'information ont toujours convergé vers une dissociation de la forme et du sens, autonomisant par là la forme qui devient objet de savoir technique et mathématique et qui constitue la phénoménologie de la théorie de l'information [Dahmane, 1990].

1.2.1.1. Typologie de l'information :

L'information peut être classée et différenciée suivant ces différentes typologies, à savoir :

Nature de l'information:

- **Textuelle :** L'information textuelle est une information sous forme écrite, qu'elle soit autographiée, dactylographiée, imprimé ou produite par d'autres moyens lisibles à l'oeil nu comme par exemple : Un livre.
- **Sonore (phonétique) :** Se sont des enregistrements magnétiques ou optiques de la voix et du son, disque, cassette audio.
- **Iconographique** : Représenté par des images, photographie, diapositive

Nature du support :

- **Imprimé** (thèse et mémoire, périodique, rapport).
- **Photographique** (photographie,…)
- **Magnétique** (disque, bande…).
- **Optique** (CD-ROM…).

Son rôle :

- **Information primaire :** véhiculée par des documents primaires c'est à dire des documents qui présentent une information à caractère originale (telle qu'elle a été écrite par son auteur). Exemples : les thèses, les livres.
- **Information secondaire :** c'est une information qui cite et qui signale l'information primaire telles que les bibliographies, les index.
- **Information de référence :** c'est une information véhiculée par les documents résultant des documents primaires et secondaires et faisant l'état de l'état tels que annuaire, encyclopédie.

Sa diffusion :

- Information à diffusion interne.
- Information à grande Diffusion.
- Information à diffusion restreinte.

1.2.1.2. L'information scientifique et technique

Information scientifique et technique, désigné habituellement par le sigle IST, est une information spécialisée et utile, produite par des spécialistes pour un public spécialisé, aux fins de recherche, de gestion et / ou de décision.

Elle est issue du monde de la recherche. C'est le résultat de l'activité scientifique de la communauté scientifique qui concerne l'information contenue dans des revues scientifiques, les thèses, les rapports internes.

1.2.2. Document

Un document est un objet porteur d'information, il est à la fois le support et l'information qu'il renferme. Selon le petit ROBERT : " le document est un écrit qui sert de preuve ou de renseignement". Cette définition renferme trois concepts: l'écrit, la preuve et le renseignement. Le premier concept traduit l'existence du texte en tant que moyen de communication, le second la valeur probante du document qui découle de son caractère tangible et le dernier l'information que renferme le document.

La définition du petit Robert n'a pas souligné l'existence d'un support d'une manière explicite, mais il serait aisé de déduire implicitement à partir de la signification des trois concepts (écrit, preuve, renseignement) l'existence de ce support. Un document est donc un ensemble formé par un support et une information, généralement enregistré de façon permanente et tel qu'il puisse être lu par l'homme ou la machine.

Aujourd'hui, avec la progression rapide des nouvelles technologies de l'information, le concept document est caractérisé par une importante évolution, il ne s'agit plus seulement de textes mais d'images, de son, de la vidéo, etc. qui pouvant être intégrés à son contenu, ce qui permet son interactivité qui facilite sa diffusion.

1.2.3. Corpus

Un corpus est une collection et un ensemble organisé de données constituées dans un but spécifique d'études langagières. Il est défini comme un échantillon fini représentatif de la langue ou des aspects de la langue qu'on veut étudier. Cet échantillon de textes doit être sélectionné selon des critères précis. Il est souvent associé à des domaines, des sous-domaines voire à des micro-domaines spécialisés.

Pour créer un corpus, deux problèmes sont à considérer : l'homogénéité et la taille. La taille est caractérisée par le nombre de mots. A l'heure actuelle des gros corpus comprennent plusieurs centaines de millions de mots. Un corpus homogène couvre un domaine spécifique dans toute sa diversité [Turenne, 2000].

Nous nous imposons certaines hypothèses que doit vérifie le corpus pour la suite de notre travail:

- Le corpus est de nature textuel.
- Le corpus est homogène c'est-à-dire il traite une certaine thématique donnée.
- Il est exhaustif pour le traitement d'une thématique.
- Le corpus est de taille fini c'est-à-dire qu'il contient un nombre fini de termes.
- Le corpus est stable dans le temps donc notre corpus n'est pas dynamique c'est-à-dire qu'on n'ajoute pas d'autres informations dans le corpus au fil du temps.

1.2.4. La science de l'information

D'origine anglo-saxonne, la science de l'information est issue de la science des bibliothèques et donc pris comme objet l'étude de l'information délivrée par ces organismes, qu'ils soient bibliothèques publiques, bibliothèques universitaires ou centres de documentation. C'est une science rigoureuse (Figure1.13 tirée de [Dragulanescu, 2003]) en pleine expansion, qui prend appui sur une technologie toute aussi rigoureuse et en interconnexion avec d'autres disciplines comme :
-psychologie : comportement de communication, représentation des connaissances.
-linguistiques : sémiotique, reformulation.
-informatique : base de données, logiciel.

-mathématiques : algorithmiques, logiques booléenne.

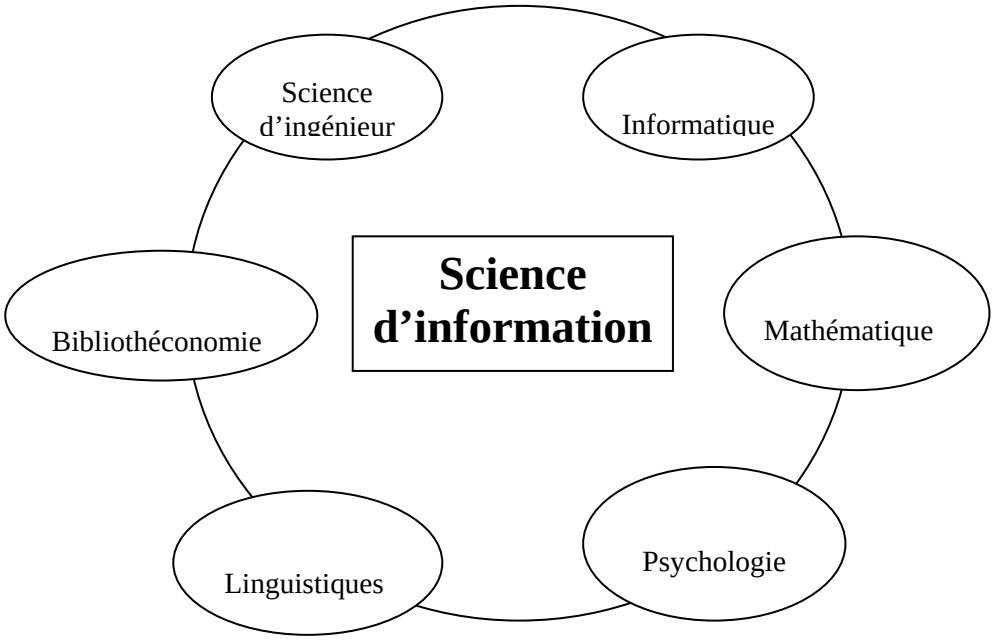


Figure1.13- *L'interdisciplinarité de la science de l'information*

La science de l'information doit accumuler des informations afin de pouvoir les analyser puis les diffuser pour prendre une décision ou communiquer l'information relative à un besoin d'un utilisateur. Elle se repose donc sur trois processus fondamentaux (Figure1.14) qui se succèdent cycliquement et indéfiniment pour tout étude des propriétés de l'information et de sa communication:

(a) Collection de l'information

Cette phase consiste à recueillir des informations en conformité avec les besoins et attentes existantes de ses usagers afin de générer l'information à partir des événements de l'environnement.

(b) Traitement de l'information

Cette phase est l'une des plus complexe du point de vue technologique [Molino, 1982]. Elle consiste à analyser des informations collectées lors de la première phase par le biais de sous processus de transformation, organisation, représentation et de mise en forme afin d'obtenir davantage la valeur ajoutée de l'information traitée pour l'usager.

(c) Diffusion de l'information

C'est la phase finale au quelle on diffuse et on communique l'information rassemblée et traitée lors des précédentes phases pour l'exploiter le plus avantageusement possible.

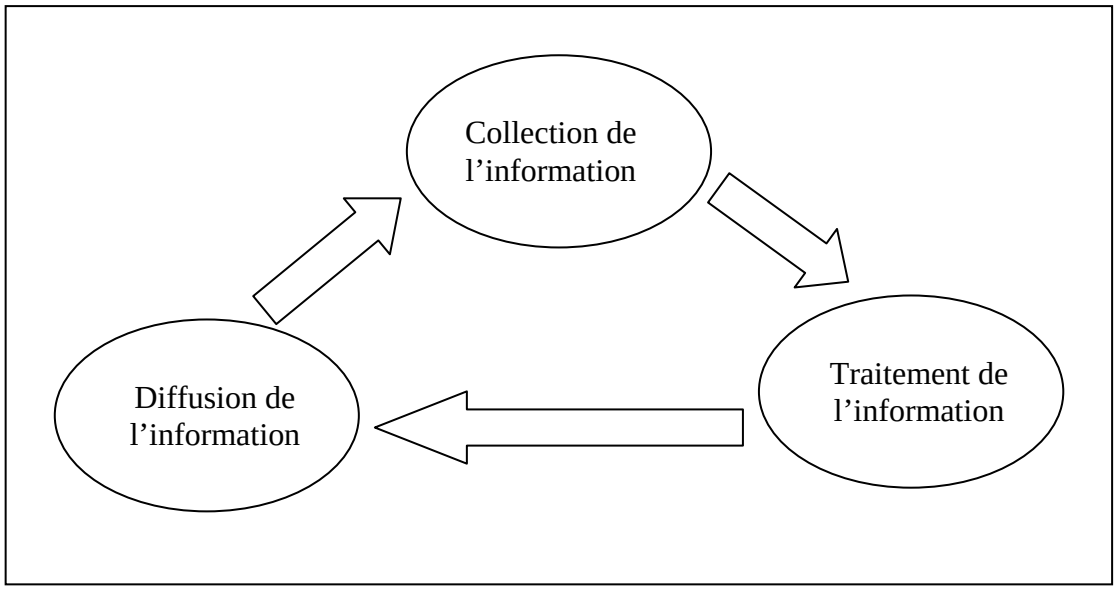


Figure1.14- *Différents cycles en science de l'information*

Aujourd'hui, les recherches en science de l'information sont très variées, compte tenu de la diversité des sujets et couvrent un champ assez étendu. Les besoins d'informations, la façon dont l'information circule et est utilisée, les comportements individuels et collectifs de communication, les relations entre les hommes et les machines au moyen desquelles les informations sont accessibles sont un groupe de sujets tout à fait essentiel et qui tiendra une place grandissante. Ces recherches portent essentiellement à l'amélioration, la gestion et la performance des systèmes d'information, cette dynamique se manifeste par énormes travaux portant sur l'information en analyse sémantique, traitement automatique des textes [Danlos, 2000], linguistique automatique [Habert, 1997] et représentation des connaissances [Sowa, 1984].

1.3. Structures des graphes et cartographie de l'information

L'entrée des mathématiques en science de l'information date des années 1920. Les premières lois scientifiques sont ainsi apparues comme celle de [Lotka, 1926] relative à la production d'articles, [Bradford, 1934] relative à la répartition des articles scientifiques pour un domaine précis, [Zipf, 1949] relative à l'étude de la distribution des mots. Ces lois sont au sens de relations quantitatives et exprimables sous la forme de fonctions mathématiques qui

établissent des relations universelles et nécessaires entre l'apparition d'un phénomène et les conditions qui le font apparaître, permettant de faire des prévisions. C'est ce qui a entraîné la naissance en science de l'information d'un nouvel axe de recherche et de développement appelé l'infométrie.

Les techniques mathématiques issues de la théorie des graphes ont des applications diveres et variées dans le domaine de la science de l'information pour le traitement de la production de l'information. En effet, les récentes avancées dans ce domaine ont fait ressortir le rôle central que joue la cartographie d'information dans la dynamique de nombreux phénomènes informationnels. Ces cartographies peuvent être modélisés par des graphes dont les sommets représentent les acteurs du phénomène et les liens représentent les interactions entre eux. Ce qui permet de synthétiser des informations sous une forme facile à interpréter et à exploiter et qui conduit à une compréhension simple de la structure sous forme d'une carte des liens entre les acteurs du phénomène. Parmi ces cartographies informationnel, on peut citer

1.3.1. Graphe du web

Parmi les applications de la théorie des graphes dans le domaine de l'information, on peut citer le développement de l'Internet où les graphes sont particulièrement utiles pour analyser la structure de la toile [Broder & al, 2000], [Flake, 2002] puisque la représentation graphique du web montre que ce dernier peut être considéré comme un graphe orienté où les pages web sont les sommets et les liens hypertextuels reliant les pages web sont les arcs du graphe et il est de type dynamique (auxquels on ajoute ou on supprime des sommets ou des arcs au cours du temps) ceci à cause des caractéristiques d'Internet à savoir : la masse importante d'information qui y circule et qui ne cesse de s'accroître exponentiellement du jour en jour et l'évolution permanente des données suite aux opérations d'ajout, de modification et de suppression. Ainsi en analysant la topologie du graphe ainsi obtenu, on pourra améliorer le fonctionnement des moteurs de recherche et localiser rapidement les documents où l'information dont on a besoin, estimer la quantité d'information contenue dans Internet.

1.3.2. Graphe de termes

Un terme est une unité signifiante constituée d'un mot (terme simple) ou de plusieurs mots (terme complexe), qui désigne une notion de façon univoque à l'intérieur d'un domaine [Dubois & al, 1994].

Un graphe de termes d'un corpus est un graphe dans lequel il montre toutes les relations qui existent entre les termes d'un ensemble $\mathcal{T}$, qui est extrait à partir d'un logiciel d'extraction de termes (voir Annexe B), sous la forme d'un graphe.

La détermination d'un contexte est un point important pour l'élaboration du graphe de termes. De manière générale, dans les différentes approches existantes, un contexte peut s'identifier à une entité textuelle. Une entité textuelle peut être une phrase, un paragraphe, ou un document [Mokrane & al, 2004], [Haddad, 2002].

Ainsi, pour construire un graphe de termes, on doit segmenter et découper notre corpus initial en entités textuelles selon le choix qui est porté sur le contexte.

Nous posons:

$$H = \{c \,/\, c \text{ est une entité textuelle du corpus } \Gamma \text{ qui correspond au contexte choisi}\},$$

L'ensemble H va correspondre à l'annotation des différentes phrases ou paragraphes ou documents du corpus suivant le contexte choisi.

Soit $\mathcal{T}$ un ensemble de termes du corpus textuel Γ:

$$\mathcal{T} = \{i \,/\, i \text{ est un terme dans le corpus } \Gamma\}$$

Pour chaque terme $i \in \mathcal{T}$, on pose :

$$D(i) = \{ c \in H \,/\, i \in c \}$$

$D(i)$ est une liste d'entités textuelles relative au terme i, elle représente l'ensemble de tous les entités textuelles contenants le terme i et donc le nombre d'occurrences t_i d'un terme i dans un corpus Γ suivant un contexte donnée correspond au nombre d'entités textuelles contenants le terme i dans ce corpus alors que le nombre de cooccurrences t_{ij} représente le nombre d'apparition commune d'un terme i avec un autre terme j dans un contexte c'est-à-dire le comptage du nombre de fois de la présence simultanée de deux termes dans le corpus, ce qui permet de rechercher les associations de termes fréquents.

La nombre d'occurrences et le nombre cooccurrences peuvent être exprimés en fonction de $D(i)$, ainsi :

- $t_i = |D(i)|$.
- $t_{ij} = |D(i) \cap D(j)|$.

Ceci peut être schématisé par la figure suivante :

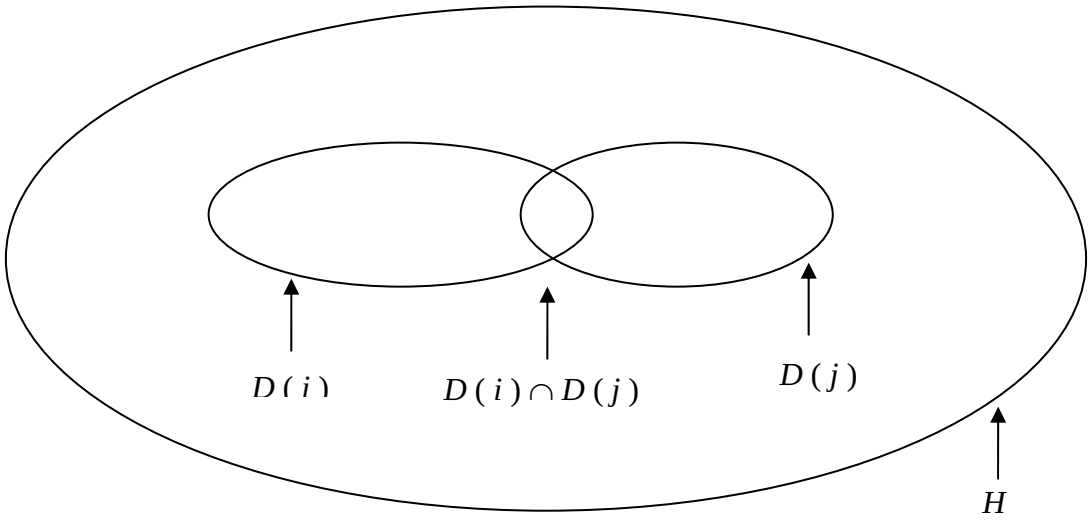


Figure1.15- *Ensemble de toutes les entités textuelles du corpus Γ.*

Ces outils sont développés dans le domaine de la recherche d'information pour rapprocher des termes qui apparaissent fréquemment dans le même contexte d'un corpus, et qui possèdent donc sans doute une certaine proximité sémantique.

Définition :

Soit $\mathcal{T}$ un ensemble de termes significatifs d'un corpus textuel Γ extrait à partir d'un logiciel d'extraction de termes (voir annexe B).

Un graphe de termes $G = (V\ E)$, relative au corpus Γ dans un contexte donné, est un graphe simple non orienté, tel que :

- $V = \mathcal{T}$.
- $\forall\ i \in V, \forall\ j \in V : ij \in E \Leftrightarrow t_{ij} > 0 \Leftrightarrow D(i) \cap D(j) \neq \phi$.

Un graphe de termes est donc un graphe non orienté. Les sommets sont formés par les unités lexicales. Il existe une relation entre deux termes *i* et *j* si l'intersection des deux listes relatives au termes *i* et *j* n'est pas vide.

1.3.3. Graphes petits mondes

Une population vérifie la propriété dite du petit monde s'il ne faut que peu d'intermédiaires pour faire rencontrer deux individus sélectionnés au hasard dans cette population.

En 1967 Milgram [Milgram, 1967] proposa la notion de graphes petits mondes. Il mena une expérience originale pour donner une interprétation scientifique et comprendre ce qui nous fait dire "Le monde est petit". Il demanda à des volontaires d'envoyer une lettre à des personnes qu'ils ne connaissaient pas personnellement. Ils devaient envoyer la lettre à des personnes qui la rapprocheraient de son destinataire. Toutes les lettres arrivèrent avec en moyenne cinq intermédiaires (les valeurs variant entre 2 et 10) pour que deux personnes dans le monde soient mises en contact. Cette propriété est souvent appelée six degrés de séparation.

Watts et Strogatz [Watts & al, 1998], [Watts, 1999] définissent les graphes petits mondes par deux propriétés.

- La longueur moyenne des chemins dans le graphe.
- La densité moyenne de graphe.

Ainsi, deux sommets ayant un voisin commun ont plus de chance d'être connectés que deux sommets pris au hasard. En d'autres termes, deux sommets adjacents d'un graphe « petit monde » ont davantage de chances de partager des voisins communs. Cette caractéristique, correspondant à l'adage populaire « les amis de mes amis sont mes amis ».

La classe des graphes petits mondes n'est pas établit précisément. Elle n'est pas définie par une propriété structurelle qui peut être identifiée à coup sûr. Cette classe correspond aux graphes qui ont une longueur moyenne d'arête faible et aux graphes qui ont une densité moyenne élevée.

Aujourd'hui, à la suite de l'article de [Watts & al, 1998], les petits mondes ont fait l'objet de nombreux travaux, et cette structure a été découverte dans de très nombreux réseaux réels : Dans les graphes sociaux, mais aussi dans les réseaux de neurones par Watts [Watts, 1999], dans les voies métaboliques [Fell & al, 2000], dans les réseaux d'interactions protéine-protéine [Wagner, 2001], dans des graphes du web par Adamic [Adamic, 1999] et dans des graphes de termes [Matsuo & al, 2001].

Chapitre 2

Partitionnement d'un graphe d'association de termes

2.1. Introduction

Un utilisateur à la première observation d'une représentation des données ne s'attarde pas sur les détails. Il regarde la donnée dans son ensemble et cherche à comprendre comment sont organisés ses éléments (comment ils sont regroupés dans l'espace de représentation). Pour aider l'utilisateur dans cette étape, il faut lui proposer une vue où les éléments sont déjà regroupés. Le regroupement est réalisé en utilisant une méthode dite de **clustering**.

Un clustering, appelé également classification non-supervisée [Jain & al, 1988], est un processus qui permet d'organiser un ensemble de données en classes cohérentes ou homogènes, appelées **clusters**, pour simplifier la représentation des données initiales et qui s'applique, a priori, sur n'importe quel type de données.

Ainsi, classifier un ensemble X signifie allouer chacun de ses éléments dans un ou plusieurs des sous-ensembles de X à définir donc on appelle cluster (ou bien groupe, ou classe) un sous-ensemble de l'ensemble X. Le groupage des éléments de l'ensemble X en classes est fait avec un certain but.

Il existe plusieurs modalités pour exprimer le but d'un clustering : intuitivement, ce but correspond au groupage des éléments de X doit être fait de telle sorte qu'à l'intérieur d'un cluster les éléments soient aussi « homogènes » que possible, tandis que deux clusters doivent être aussi « différents » que possible. Autrement dit, les clusters entre eux doivent être aussi « hétérogènes » que possible, ces termes d'homogène et hétérogène faisant référence aux propriétés communes des éléments de l'ensemble X, propriétés en rapport avec lesquelles est faite le clustering.

Pour cette formulation intuitive, Saporta et Stefanescu, dans [Saporta & al, 2003], font correspondre une formulation mathématique d'un problème de clustering, à savoir : classifier l'ensemble X signifie construire une décomposition (qui peut être un partitionnement) de X formée de sous-ensembles qui doivent avoir la propriété de similarité-interne et dissimilarité-externe. Une autre formulation mathématique du but d'un clustering peut être, par exemple, le partitionnement de l'ensemble X tel que le partitionnement soit optimal par rapport à une certaine fonction objectif [Arthanary & al, 1980].

La formulation qui est donnée par Arthanary et Dodge [Arthanary & al, 1980] peut être résumé de la manière suivante :

Soit X un ensemble de cardinalité n qu'on souhaite classifier en m classes, avec $m \leq n$. Soit τ un indice de dispersion (fonction qui à chaque sous ensemble de X associe un nombre réel plus grand que zéro, et (souvent) zéro pour les singletons). Soit $\pi_m = (A_1, A_2, ..., A_m)$ une partition de l'ensemble X. Notons Π_m l'ensemble de toutes les partitions m classes de l'ensemble X, alors $|\Pi_m|$ représente le nombre de Stirling de 2eme espèce ($|\Pi_m| = S(n, m)$). Soit enfin une fonction $C : R^m \rightarrow R$ appelée fonction objective, $C(\pi_m) = C(\tau(A_1), \tau(A_2), ..., \tau(A_m))$. Alors le problème de classification, tel qu'il est formulé par Arthanary et Dodge ([Arthanary & al, 1980]) est le problème d'optimisation suivant :

$$(PC) : \ opt_{\pi_m \in \Pi_m} C(\pi_m)$$

Les méthodes de clustering ont un objectif précis : former des classes cohérentes (ou homogènes) et bien isolées [Govaert, 2003]. L'adjectif cohérent veut dire que les éléments appartenant à une classe partagent de nombreuses caractéristiques communes et donc se ressemblent fortement selon un but donné. Par isolée, on veut dire que deux classes ne se ressemblent pas, c'est-à-dire qu'elles ne partagent pas du tout les mêmes caractéristiques, ce qui fait que le problème de clustering est un problème mal définit puisque le critère de partage des données est subjectif, elle dépend selon le domaine d'application et selon un point de vue adopté par l'utilisateur.

Le clustering est très utile pour la navigation car il permet de proposer une vue simple des données [Shneiderman, 1996]. Par exemple si un cluster est symbolisé par un seul objet le nombre d'objets représentés va diminuer. Ainsi la représentation sera plus claire (Figure2.1).

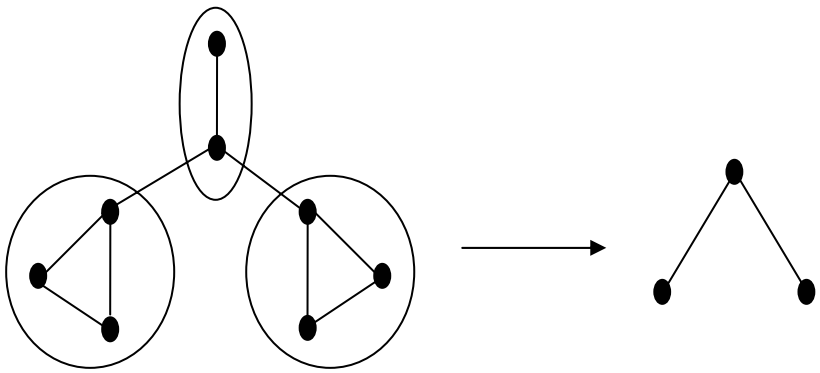


Figure2.1- *Réduction de la visualisation par un clustering*

2.2. Présentation du problème

Le problème du clustering se ramène, pour l'essentiel, à un problème de partitionnement du graphe [Manolis & al, 1991], [Manolis & al, 1992]. Un partitionnement est la division des données ou objets en groupes. Chaque groupe, appelé communauté (cluster), est un ensemble d'objets partageant un intérêt commun qui sont similaires entre eux et dissimilaires aux objets des autres groupes. En terme de graphe, elle consiste à regrouper les sommets en communautés selon un critère prédéfinit.

Une communauté dans un graphe correspond intuitivement à l'existence de groupes de sommets plus fortement connectés entre eux que vers les autres sommets. Les communautés peuvent avoir des interprétations différentes suivant le type de réseau à considérer. Le terme de communauté dans un réseau d'information a souvent été synonyme de « thème ». Ainsi, dans ce réseau la problématique revient à organiser un ensemble de données en différentes thématiques.

Le problème de partitionnement d'un graphe peut s'énoncer synthétiquement de la manière suivante:

Soit un graphe $G = (V, E)$ alors un partitionnement $C = (c_1, c_2, \ldots, c_p)$ de taille p du graphe G est un ensemble de sous ensembles de V qui vérifie les propriétés suivantes :

- $\forall\, i \in \{1, 2,.., p\} : c_i \neq \phi$, c_i est appelé communauté ou cluster.
- $\forall\, i \in \{1, 2,.., p\}, \forall\, j \in \{1, 2,.., p\},\ i \neq j : c_i \cap c_j = \phi.$
- $\bigcup_{i=1}^{i=p} c_i = V.$

Soit $H : L(V) \rightarrow R$ où $L(V)$ est l'ensemble de toutes les partitions possibles de l'ensemble de sommets V alors $|L(V)|$ représente le nombre de Bell ($|L(V)| = B_n$). Ainsi, *opt H (C)* ($C \in L(V)$) représente le but que l'utilisateur veut assigner à son clustering.

Malheureusement, cet objectif est difficile à atteindre, car les problèmes posés en terme d'optimisation d'un critère sont NP-complet [Garey & al, 1979], [Day, 1996] et les solutions optimales de ce type de problème restent inaccessibles en pratique. Diverses démarches ont donc été développées pour se rapprocher le plus possible de ces optima [Hansen & al, 1997].

Par la suite, nous allons donner quelques approches pour la résolution de ce problème pour le cas d'un graphe d'association de termes définie ci-dessous.

2.2.1. Etat de l'art

Il existe de nombreuses méthodes de partitionnement dans la littérature [Berkhin, 2002] et qui peuvent être divisées en deux classes [Shahookar & al, 1991]: le partitionnement constructif et l'amélioration itérative. La première classe construit directement un partitionnement entièrement nouveau. La seconde démarre d'un partitionnement initial et tente d'améliorer celui-ci itérativement en minimisant une fonction de coût.

Parmi ces méthodes de partitionnement en peut citer :

Le regroupement en profondeur (*Depth First Search* en anglais) et en largeur (*Breadth First Search* en anglais) qui sont présentés par Stamos [Stamos, 1984]. Le principe de base repose sur le parcours du graphe soit en profondeur, soit en largeur. Les sommets sont regroupés dans l'ordre où ils sont visités.

Le partitionnement sous forme de décomposition hiérarchique (*Hierarchical Decomposition* en anglais) qui a été proposé par Wilson et al. [Wilson & al, 1991] afin de combiner les avantages de *DFS* et de *BFS*.

Partitionnement optimal d'arbres (*OPT*) qui est proposé par Lukes [Lukes, 1974].

L'algorithme *KL* de Kernigan et Lin [Kernigan & al, 1970] qui est une heuristique d'optimisation itérative destinée à améliorer un partitionnement initial d'un graphe. Étant donné un partitionnement initial, *KL* minimise le coût total du partitionnement en échangeant la position des sommets dans les clusters deux à deux.

L'algorithme *FM* de Fiduccia-Mattheyses [Fiduccia & al, 1982], comme *KL*, est une heuristique d'optimisation itérative destinée à améliorer un partitionnement initial d'un graphe. Il s'inspire largement de *KL*.

L'algorithme de partitionnement hiérarchique (*Hierarchical Partitioning* en anglais) qui est présenté par Gerlhof et al. [Gerlhof & al, 1993] d'après une idée de Breuer [Breuer, 1977]. Il combine un mécanisme hiérarchique de placement initial avec une heuristique d'optimisation qui peut être *KL* ou *FM*.

2.2.2. Mesures de qualité d'un partitionnement

Pour un graphe donné on peut trouver plusieurs partitionnements possibles. Afin de comparer entre ces partitionnements, il existe dans la littérature des indices qui peuvent nous indiquer celui qui correspond au bon partitionnement [Boutin & al, 2004]. Parmi ces indices, on trouve la mesure de qualité. Cette mesure est la plus recommandée pour la détermination d'un bon partitionnement [Boutin & al, 2004].

Un partitionnement $C = (c_1, c_2, \ldots, c_p)$ aura une bonne mesure de qualité [Mancoridis & al, 1998], [Mancoridis & al, 1999] s'il minimise le nombre d'arêtes entre les clusters (arêtes inter-clusters) et maximise leur nombre à l'intérieur des clusters (arêtes intra-clusters).

Soient :

$|E(c)|$ est le nombre d'arêtes dans la classe *c* (arêtes intra-clusters).

$|E(c,c')|$ est le nombre d'arêtes entre la classe *c* et la classe *c'* (arêtes inter-clusters).

$|V(c)|$ est le nombre de sommets dans la classe *c*.

La mesure qui correspond à la densité d'arêtes dans un cluster c_i (arêtes intra-cluster) avec $i \in \{1,2,...,p\}$ se calcule de la manière suivante :

$$\text{int}ra(c_i) = \frac{|E(c_i)|}{\frac{|V(c_i)| \times (|V(c_i)| - 1)}{2}}$$

La densité d'arêtes dans les clusters (arêtes intra-clusters) est :

$$\overline{\text{int}ra} = \frac{\sum_{i=1}^{i=p} \text{int}ra(c_i)}{p} = \frac{\sum_{i=1}^{i=p} \frac{2 \times |E(c_i)|}{|V(c_i)| \times (|V(c_i)| - 1)}}{p}$$

$\overline{\text{int}ra}$ est la moyenne de toutes les densités d'arêtes intra-clusters.

La densité qui correspond à l'arête entre deux clusters c_i et la classe c_j (arêtes inter-cluster) se calcule de la manière suivante :

$$\text{int}er\left(c_i, c_j\right) = \frac{\left|E\left(c_i, c_j\right)\right|}{\left|V\left(c_i\right)\right| \times \left|V\left(c_j\right)\right|}$$

La densité d'arêtes entre les clusters (arêtes inter-clusters) est :

$$\overline{\text{int}er} = \frac{\sum_{i,j\ i \prec j}^{p} ineter\left(c_i, c_j\right)}{p \times (p-1)/2} = \frac{\sum_{i,j\ i \prec j}^{p} \frac{\left|E\left(c_i, c_j\right)\right|}{\left|V\left(c_i\right)\right| \times \left|V\left(c_j\right)\right|}}{p \times (p-1)/2}$$

De même, $\overline{\text{int}er}$ est la moyenne de toutes les densités d'arêtes inter-clusters.

Une mesure de qualité d'un partitionnement $C = \left(c_1, c_2, \ldots, c_p\right)$ de taille p d'un graphe $G = (V, E)$ est la valeur MQ tel que :

$$MQ(C) = \overline{\text{int}ra} - \overline{\text{int}er} = \frac{\sum_{i=1}^{i=p} \frac{\left|E\left(c_i\right)\right|}{\left|V\left(c_i\right)\right| \times \left(\left|V\left(c_i\right)\right| - 1\right)/2}}{p} - \frac{\sum_{i,j\ i \prec j}^{p} \frac{\left|E\left(c_i, c_j\right)\right|}{\left|V\left(c_i\right)\right| \times \left|V\left(c_j\right)\right|}}{p \times (p-1)/2}$$

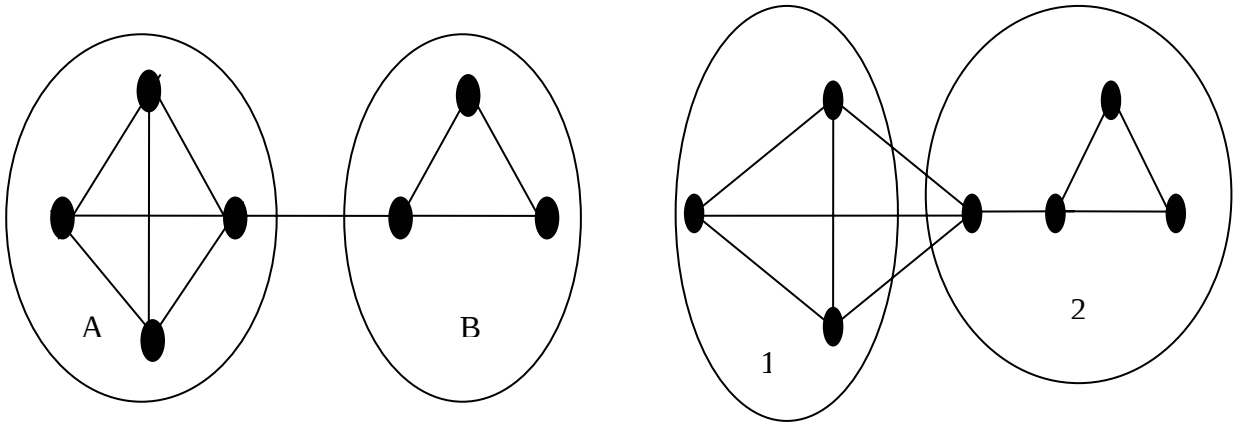
Un bon partitionnement sera celui qui aura la meilleure mesure de qualité, c'est à dire celui qui maximise la valeur de cette mesure.

Dans [Boutin & al, 2004], les auteurs ont proposés une amélioration de cette mesure, ils justifient cette amélioration par le fait que la mesure MQ ne prend pas en compte la taille des clusters, ainsi une nouvelle mesure notée MQ^* est donné et qui prend en considération ce paramètre, cette mesure se calcule de la manière suivante:

$$MQ^*(C) = \frac{\sum_i \left|E\left(c_i\right)\right|}{\sum_i \frac{\left|V\left(c_i\right)\right| \times \left(\left|V\left(c_i\right)\right| - 1\right)}{2}} - \frac{\sum_{i,j\ i \prec j}^{p} \left|E\left(c_i, c_j\right)\right|}{\sum_{i,j\ i \prec j}^{p} \left|V\left(c_i\right)\right| \times \left|V\left(c_j\right)\right|}$$

Pour plus de détail sur les indices de comparaison entres les clusters qui existent dans la littérature, il est conseillé de voir l'article [Boutin & al, 2004].

Exemple :



$|V(A)| = 4 \qquad |V(B)| = 3$

$|E(A)| = 6 \qquad |E(B)| = 3$

$|E(A,B)| = 1$

$$MQ^*(G,\{A,B\}) = \frac{6+3}{6+3} - \frac{1}{6\times 3}$$

$$MQ^*(G,\{A,B\}) = \frac{17}{18}$$

$|V(1)| = 3 \qquad |V(2)| = 4$

$|E(1)| = 3 \qquad |E(2)| = 4$

$|E(1,2)| = 3$

$$MQ^*(G,\{1,2\}) = \frac{3+4}{3+6} - \frac{3}{3\times 4}$$

$$MQ^*(G,\{1,2\}) = \frac{19}{36}$$

Figure2.2- *Deux partitionnement du même graphe et leur mesure de qualité*

Nous proposons une mesure de qualité dans le cas d'un graphe valué (les arrêtes pondérés) ceci afin de trouver un bon partitionnement. Nous définissons par analogie la mesure de qualité dans le cas d'un graphe valué à partir de celle d'un graphe non valué ceci en prenant compte le poids des arêtes.

Ainsi, un partitionnement $C = (c_1, c_2, \ldots, c_k)$ d'un graphe valué $G = (V, E)$ aura une bonne mesure de qualité s'il minimise le poids des arêtes entre les clusters (arêtes inter-clusters) et maximise leur poids à l'intérieur des clusters (arêtes intra-clusters) alors nous proposons une nouvelle mesure de qualité ψ qui ce base sur la mesure MQ^* d'un graphe non valué avec quelques variations :

$$\psi(C) = \frac{\sum_{i=1}^{i=p} \sum_{vv' \in E(c_i)} w_{vv'}}{\sum_{i} \frac{|V(c_i)| \times (|V(c_i)| - 1)}{2}} - \frac{\sum_{i,j\ i \prec j}^{p} \sum_{v \in c_i, v' \in c_j, vv' \in E} w_{vv'}}{\sum_{i,j\ i \prec j}^{p} |V(c_i)| \times |V(c_j)|}$$

Avec $w_{vv'}$ est le poids de l'arrête $vv' \in E$.

Remarque 1

Pour un graphe non valué, les poids valent 1 et on retrouve la formule *MQ**

2.3. Graphe d'association de termes

Un graphe d'association de termes $G = (V, E, W)$ est un graphe vérifiant les conditions suivantes

- $V = \mathcal{T}$ (où $\mathcal{T}$ est l'ensemble de termes candidats du corpus textuel).
- W : fonction degré d'association définie de $\mathcal{T} \times \mathcal{T}$ vers [0, 1] telle que $\forall\ i \in V$, $\forall\ j \in V$: $W(i, j) = \max(\Im^j{}_i, \Im^i{}_j)$ notée w_{ij}.
- $\forall\ i \in V, \forall\ j \in V : ij \in E \Leftrightarrow w_{ij} > 0$.

Avec $\Im^j{}_i = \frac{t_{ij}}{t_i}$ et $\Im^i{}_j = \frac{t_{ij}}{t_j}$ (t_i est le nombre d'unités textuels dans lesquels le terme *i* apparaît et t_{ij} est le nombre d'unités textuels dans lesquels le termes *i* et *j* apparaissent simultanément).

$\Im^j{}_i$ est le pourcentage qu'occupent les entités textuelles contenants le terme *j* dans l'ensemble des entités textuelles contenants le terme *i*, et inversement, $\Im^i{}_j$ le pourcentage qu'occupent les entités textuelles contenants le terme *i* dans l'ensemble des entités textuelles contenants le terme *j*. Ces indices sont estimés à partir des fréquences.

Ainsi, nous affectons à chaque arête *ij* de *E* un poids w_{ij} d'autant plus fort que les termes sont fréquemment associés.

On remarque bien que $0 \leq w_{ij} \leq 1$ et que la formule précédente du poids w_{ij} peut être remplacer par:

$$w_{ij} = \frac{t_{ij}}{\min(t_i, t_j)}$$

D'après [Véronis, 2003], ces graphes sont des graphes petits mondes et qu'un graphe petit monde est constitué d'un ensemble de communautés alors ce sont ces communautés que nous cherchons à mettre en évidence à partir des méthode de partitionnement que nous décrivons dans la suite.

2.4. Une approche basée sur la densité d'un sommet

2.4.1. Le principe

L'idée est de proposer une méthode de partitionnement qui repose sur une recherche de zones denses dans un graphe pondéré. Cette méthode se déroule en trois étapes :

- Tout d'abord on crée des noyaux à partir d'une mesure locale de densité d'un sommet et d'un seuil de densité donné.
- Ensuite, on étend les noyaux en respectant certains critères sur la qualité des classes obtenues.
- Réitérer la procédure jusqu'au partitionnement de graphe.

Ce type d'approches été utilisé dans divers travaux de recherche notamment en biologie [Guénoche & al, 2003], [Denoeud & al, 2005]. Néanmoins, les graphes considérés ne sont pas pondérés.

2.4.2. La méthode

D'après ce qui précède, la méthode que nous proposons pour la construction des communautés est basée sur la définition d'une mesure locale de la densité d'un sommet et d'un seuil de densité.

On procède de la manière suivante :

- On cherche les noyaux qui sont des composantes connexes du graphe tel que pour chaque noyau, ses sommets ont une densité supérieure au seuil fixé.
- Etendre chaque noyau en ajoutant d'autres sommets qui ont suffisamment de liens avec lui, i.e. favorisant l'amélioration de la qualité de clustering.

(A) Mesures de densité locale

Pour la construction des noyaux des communautés, on utilise une mesure de densité locale qui permet de donner la densité en arêtes au voisinage des sommets d'un graphe. Dans leur article [Guénoche & al, 2003] proposent et comparent les trois mesures de densité suivantes :

- Le degré $d(v)$ de *v* divisé par le plus grand degré Δ :

$$De_1(v) = \frac{d(v)}{\Delta}$$

- Le degré moyen dans le voisinage de *v* : la somme des degrés du sommet *v* et de ses sommets adjacents rapportée au nombre de sommets considérés:

$$De_2(v) = \frac{d(v) + \sum_{v' \in N(v)} d(v')}{d(v) + 1}$$

- La mesure $De_3(v)$: le nombre d'arêtes entre les voisins du sommet *u* par rapport au nombre d'arêtes qui pourraient exister. Cette mesure est celle qui est utilisé par Watts et Strogatz [Watts & al, 1998] dans le modèle de graphe de petits mondes, elle correspond à la mesure $De(v)$ d'un sommet *u*, ainsi :

$$De_3(v) = De(v) = \frac{|E(N(v))|}{\binom{d(v)}{2}} = \frac{|E(N(v))|}{\frac{d(v) \times (d(v)-1)}{2}}$$

Cette dernière est la plus utilisée et la plus satisfaisante dans les approches de classification [Guénoche & al, 2003].

Nous allons utiliser une mesure densité locale $\overline{De}(v)$ dans le cas d'un graphe pondéré. Cet indice évalue la densité en arêtes pondérées au voisinage d'un sommet *u*. Il se base sur la pondération des arêtes, il est définit comme suit :

$$\overline{De}(v) = \frac{\sum_{uu' \in E(N(v))} w_{uu'}}{\binom{d(v)}{2}} = \frac{\sum_{uu' \in E(N(v))} w_{uu'}}{\frac{d(v) \times (d(v)-1)}{2}}$$

Cette mesure est un peu plus fine que la mesure de densité $De(v)$: au lieu de tenir compte simplement de la présence ou de l'absence d'une arête, elle tient compte également de leurs poids respectifs [Jéronis, 2003].

(B) Noyaux dans un graphe

Définition

Soient un graphe d'association de termes $G = (V, E, W)$et σ un seuil donné.

On dit que $A \subset V$ est un noyau si est seulement si :

- $\forall\ v \in A,\ \overline{De}(v) \geq \sigma$.
- G_A est une composante connexe.

L'algorithme suivant permet de déterminer les noyaux d'un graphe d'association de termes $G = (V, E, W)$

Algorithme Noyau: Méthode recherche des noyaux d'un graphe d'association de termes $G = (V, E, W)$.

Données: *Un graphe d'association de termes* $G = (V, E, W)$*, un seuil* σ .

Résultat: *noyaux du graphe G.*

Début

(1) Pour chaque sommet v dans V

(i) Calculer sa densité $\overline{De}(v)$.

(ii) Si $\overline{De}(v) \geq \sigma$ *alors placé v dans J.*

(2) Construire le graphe G_J *qui est sous graphe de G induit par l'ensemble de sommets J.*

(3) Les noyaux sont les composantes connexes du graphe G_J .

Fin

(C) Critères d'extension

Dans cette étape, il s'agit d'étendre les noyaux par les éléments qui ne sont pas encore classés c'est-à-dire ceux qui n'atteignent pas le seuil de densité voulu.

Le principe de cette extension est d'affecter tous les éléments qui s'y rattachent aux noyaux indexés selon des critères et de considérer les nouveaux noyaux qui sont constitués par les noyaux de départ avec les nouveaux éléments affectés, si les éléments de tous ces noyaux

constituent l'ensemble V alors on arrête sinon on réitère la procédure jusqu'au partitionnement de tous les éléments de V.

Ainsi, il reste juste de déterminer les critères d'affectation des sommets adjacents aux noyaux. L'idée est de traiter les éléments voisins des noyaux séquentiellement. Nous ordonnons les éléments voisins des noyaux selon l'ordre décroissant de la densité et nous ajoutons un élément pris dans l'ordre décroissant dans un noyau qui lui est très connecté car cet élément a plus de possibilité qu'il soit homogène et donc il traite la même chose avec ce noyau que par rapport à d'autres noyaux puisque il a une relation forte avec lui. S'il existe plusieurs noyaux qui sont très connectés avec cet élément alors on prend celui qui a un nombre d'arêtes minimales sinon on l'affecte au noyau d'indice minimum.

Algorithme Ex: Méthode d'extension des noyaux d'un graphe d'association de termes $G = (V, E, W)$.

Données: *Un graphe d'association de termes* $G = (V, E, W)$*, un seuil* σ *.*

Résultat: *un partitionnement du graphe* $G = (V, E, W)$*.*

Début

(1) Poser $V = \{v_1, v_2, \ldots, v_n\}$

(2) Déterminer un ensemble de noyaux $(A_i)_{i=1\cdots p}$ *d'un graphe d'association de termes et qui est déterminé par l'algorithme Noyau par rapport au seuil* σ *.*

(3) Trouver l'ensemble $L = \bigcup_{i=1}^{i=p} N(A_i) - \bigcup_{i=1}^{i=p} A_i$ *des éléments adjacents aux différents noyaux.*

(4) Classer les éléments de L selon l'ordre décroissant de la densité, s'il existe plusieurs sommets qui ont même densité alors on les classe selon l'ordre décroissant de leurs indices.

(5) Pour chaque sommet u pris dans l'ordre décroissant de densité:

(i) Pour chaque noyau A_i

(a) Calculer les nombres h_i *et* k_i *avec* $h_i = \sum_{v \in A_i, uv \in E} w_{uv}$, $k_i = \sum_{v \in A_i, v' \in A_i \; uv \in E} w_{vv'}$

(ii) Le sommet v est affecté au noyau A_j *tel que* h_j *est maximum, en cas d'égalité entre plusieurs noyaux, celui qui a pour* k_i *minimum sinon on affecte v au noyau d'indice maximum.*

(6) Si $\bigcup_{i=1}^{i=p} A_i = V$ *. Terminer, on obtient un partitionnement* $(A_i)_{i=1\cdots p}$ *relatif au seuil* σ *.*

(7) Sinon, on revient à (2).

Fin

Remarque 2

Pour un ensemble de noyaux initial, on trouve qu'un seul partitionnement par l'algorithme Ex.

2.4.3. L'algorithme

Le principe de l'algorithme de partitionnement d'un graphe d'association de termes par la méthode de densité d'un sommet se base sur le choix d'un seuil qui maximise la mesure de qualité à partir de l'algorithme EX. Dans le cas général, choix du seuil pour trouver un bon clustering est un passage critique. Il est impossible de savoir, a priori, quelle valeur du seuil choisir pour faire un bon partitionnement du graphe.

Le choix du seuil par cette approche sera itératif, c'est-à-dire que nous devons appliquer le processus en calculant une série de partitionnement correspondant à des différentes valeurs. Le seuil choisi sera celui qui correspond à la meilleure mesure de qualité du clustering et dans notre cas cette valeur est atteinte puisque la fonction qui met en correspondance la mesure de qualité par rapport au seuil est une fonction en escalier, en effet :

Soit g : [0, 1]$\longrightarrow$ [0, 1]

$x \longmapsto g(x) = \Psi(C_x)$

Avec $\Psi(C_x)$ est la mesure de qualité du partitionnement C_x trouvé par l'algorithme *Ex* relatif au seuil x

Choisir un seuil qui correspond au bon partitionnement revient à trouver un seuil $x' \in$ [0, 1] tel que :

$$\forall x \in [0, 1] : g(x') \geq g(x)$$

Si on pose $V = \{v_1, v_2, ..., v_n\}$ avec $0 \leq \overline{De}(v_1) \leq \overline{De}(v_2) \leq \overline{De}(v_3) \leq ... \leq \overline{De}(v_n) \leq 1$, alors pour deux seuils σ, σ' tel que $\overline{De}(v_i) \leq \sigma \leq \sigma' \prec \overline{De}(v_{i+1})$, les noyaux relatif au seuil σ sont les mêmes qu'aux noyaux qui sont relatif au seuil σ' ceci à partir de l'algorithme noyaux et d'après la remarque 2 ils vont constitué un même partitionnement par l'algorithme Ex , ce qui montre que :

$$\psi(C_\sigma) = \psi(C_{\sigma'})$$

La fonction *g* est donc une fonction en escalier sur l'intervalle [0, 1] et la borne supérieur de cette fonction est l'élément $x' \in$ [0, 1] : $g(x') = \psi(C_{x'}) = Max\{\psi(C_x) / x \in F\}$ où $F = \{\overline{De}(v) / v \in V\}$

A partir des observations précédentes, on en déduit l'algorithme de partitionnement d'un graphe d'association de termes par la méthode de densité d'un sommet.

Algorithme: Approche de partitionnement basée sur la densité d'un sommet.

Données: *Un graphe d'association de termes* $G = (V, E, W)$.

Résultat: *Un partitionnement C du graphe G.*

Début

(1) Poser $F = \{\overline{De}(v) / v \in V\}$

(2) Pour chaque valeur du seuil $\sigma \in F$.

(i) Appliquer l'algorithme Noyau.

(ii) Déterminer un partitionnement $C_\sigma = (c_1^\sigma, c_2^\sigma, \ldots, c_p^\sigma)$ *à partir de l'algorithme Ex*

(iii) Calculer la mesure de qualité $\psi(C_\sigma)$ *du partitionnement* C_σ.

(3)Trouver un seuil $\sigma' \in F$ *tel* $\psi(C_{\sigma'}) = Max\{\psi(C_\sigma) / \sigma \in F\}$.

(4) $C_{\sigma'}$ *relatif au seuil* σ' *est le partitionnement basé sur l'approche de la densité d'un sommet.*

Fin

2.5. Une approche basée sur l'importance d'une arête

2.5.1. Principe

Dans la première approche, on s'est intéressé au partitionnement d'un graphe d'association de termes à partir de la densité d'un sommet. L'idée qui a conduit à cette deuxième approche est qu'en se basant sur la question pourquoi ne pas faire la même chose mais au lieu de mesurer la densité d'un sommet, on mesure le densité d'une arête. Comme la mesure de densité d'un sommet qui évalue la densité en arêtes pondérées au voisinage de ce sommet, la densité d'une arête évalue la densité du voisinage de ses extrémités.

Dans ce cas, le principe de partitionnement du graphe est basé sur la connectivité des voisinages d'une arête pondérée. En d'autre terme, elle se base sur la densité de liens qui existe entre les sommets proches qui est représenté dans notre cas par la densité de liens dans le voisinage des extrémités d'une arête.

En effet, une communauté dans un graphe correspond intuitivement à l'existence des sommets fortement connectés entre eux. Alors si les sommets du voisinage appartiennent à la même communauté les liens entre ces sommets deviennent **importants** donc les deux voisinages des extrémités d'une arête sont très connectés ce qui implique que l'arête est importante. Par contre, si les voisinages sont très peu connectés cela signifie que les deux extrémités sont très probablement dans deux communautés disjointes donc le lien sera faible. Par conséquent, cette métrique a été définie pour que les arrêtes ayant une valeur forte sont à l'intérieur des communautés alors que les arêtes ayant une valeur faible relient les communautés.

Le premier qui s'est intéressé à cette mesure est Jourdan [Jourdan, 2004], il l'appela force d'une arête. Néanmoins, la méthode proposée ne peut être appliqué dans notre cas. Nous allons développer notre approche à partir de ses travaux.

2.5.2. La méthode

Soient un graphe d'association de termes $G = (V, E, W)$ et $uv \in E$.

Pour faire le partitionnement dans un graphe $G = (V, E, W)$ basé sur l'importance des arêtes, nous allons diviser les voisins de *u* et *v* en trois ensembles distincts (voir figure2.3) :
- *L* (*u*) : L'ensemble des voisins de *u* qui ne sont pas voisins de *v*.
- *L* (*v*) : L'ensemble des voisins de *v* qui ne sont pas voisins de *u*.
- *L (u,v)* : L'ensemble des voisins communs à *u* et *v*.

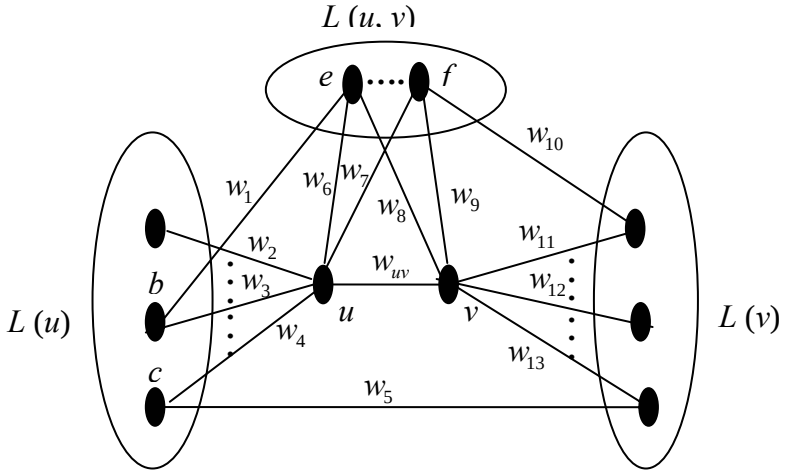


Figure2.3- *Division des voisins des extrémités u et v en trois ensembles L (u), L (v) et L (u, v)*

L'importance d'une arête est une mesure en fonction de liens entre $L(u,v)$, $L(v)$ et les liens entre les ensembles $L(u,v)$, $L(u)$ de même entre les ensembles $L(v)$, $L(u)$ et les liens dans l'ensemble $L(u,v)$ ce qui permet de mesurer tous les liens possible entre $N(u)-\{v\}$ et $N(v)-\{u\}$ et donc mesurer la connectivité des voisinages de l'arête *uv.*

Soit *r* la fonction définie comme suit :

Si *A* et *B* sont deux ensembles de l'ensemble *V*:

$$r(A,B)=\sum_{i\in A,\ j\in B,\ ij\in E} w_{ij}\ .$$

et *s* la fonction définie par :

$$s(A)=\sum_{ij\in E(A)} w_{ij}\ .$$

Où A est un ensemble de l'ensemble V.

A) Définition de la mesure de l'importance d'une arête

On définit la mesure d'importance d'une arête *uv* par:

$$IP(uv)=\frac{r(L(u),L(v))+r(L(u),L(u,v))+r(L(v),L(u,v))+s(L(u,v))}{|L(u)|\,|L(v)|+|L(u)|\,|L(u,v)|+|L(v)|\,|L(u,v)|+\frac{|L(u,v)|\,(|L(u,v)-1|)}{2}}.$$

On remarque bien que $0\le IP(uv)\le 1$.

B) Identifier les communautés

La notion d'importance d'une arête a été définie pour que les arêtes ayant une valeur faible relient les communautés et que les arêtes qui sont à l'intérieur de ces communautés ont une valeur élevée. Mais la notion de faible ou de forte importance d'une arête est relative à un seuil noté α.

Alors, pour une arête *e,* on dit qu'elle est de forte importance (respectivement de faible importance) si $IP(e)\ge\alpha$ (respectivement $IP(e)\prec\alpha$).

Ainsi, pour déterminer les communautés dans un graphe d'association de termes à partir de la notion de l'importance d'une arête, on supprime toutes les arêtes qui ont une importance inférieure au seuil α (i.e. les arêtes de faible importance). Les communautés seront les composantes connexes engendrées par les arêtes sélectionnées

S'il existe des sommets non classés (i.e. des sommets isolés dans le graphe construit par les arêtes sélectionnées), alors, il paraît plus probable d'affecter ces sommets vers d'autre clusters que de constituer des clusters à un seul sommet. Le principe d'affectation de ces sommets se fait comme le principe d'extension dans la méthode précédente.

Algorithme: PARTITIONNEMENT

Données: *Un graphe d'association de termes* $G = (V, E, W)$, *un seuil* α.

Résultat: *Une partitionnement* C_α *du graphe* $G = (V, E, W)$.

Début

(1) Poser $E = \{e_1, e_2, \ldots, e_m\}$

(1) Pour chaque arête e dans E, on calcule la mesure d'importance $IP(e)$

(2) On détermine les ensembles $H_\alpha = \{e \in E / IP(e) \prec \alpha\}$ *et* $H'_\alpha = E - H_\alpha$

(3) $G_\alpha = (V, H'_\alpha)$ *un graphe obtenu à partir de G après la suppression des arêtes de l'ensemble* H_α.

(4) Soit $C_\alpha = (c_1^\alpha, c_2^\alpha, \ldots, c_p^\alpha)$ *où* c_i^α *est une composante connexe dans le graphe* G_α.

(5) Poser $I = \{i \in \{1,\ldots,p\} / |c_i^\alpha| = 1\}$ *et* $J = \{i \in \{1,\ldots,p\} / |c_i^\alpha| \neq 1\}$

(6) poser $K = \bigcup_{i \in I} c_i^\alpha$

(7) Trouver l'ensemble $L = E(\bigcup_{i \in I} c_i^\alpha, \bigcup_{i \in J} c_i^\alpha)$.

(8) Classer les éléments de L selon l'ordre décroissant de l'importance d'arête s'il existe plusieurs arêtes qui ont même importance alors on les classe selon l'ordre décroissant de leurs indices.

(9) Pour chaque arête uv $\left(u \in \bigcup_{i \in I} c_i^\alpha, v \in c_s^\alpha \text{ avec } s \in J\right)$ *de L pris dans l'ordre décroissant d'importance d'arête.*

(i) Affecter u à la classe c_s^α.

(ii) Eléminer toutes les arêtes de L qui ont u comme sommet incident. .

(10) Si $\bigcup_{i \in J} c_i^\alpha = V$ *. Terminer, on obtient un partitionnement* $(c_i^\alpha)_{i \in J}$ *relatif au seuil* α.

(11) Sinon, on revient à (4).

Fin

Remarque 3

Pour un seuil donné, on trouve qu'un seul partitionnement par l'algorithme PARTITIONNEMENT

2.5.3. Algorithme

De même que le principe de l'algorithme précèdent (l'algorithme de partitionnement basé sur la densité d'un sommet), le principe de l'algorithme de partitionnement d'un graphe d'association de termes par l'approche d'importance d'une arête se base sur le choix d'un seuil qui maximise la mesure de qualité à partir de l'algorithme PARTITIONNEMENT.

Le choix du seuil par cette approche sera itératif, c'est à dire que nous devons appliquer le processus en calculant une série de partitionnement correspondant à des différentes valeurs du seuil. La valeur choisie sera celle qui correspond à la meilleure mesure de qualité du clustering, en effet :

Soit g' : [0, 1] $\longmapsto$ [0, 1]

$x \longmapsto g(x) = \Psi(C_x)$

Avec $\Psi(C_x)$ est la mesure de qualité du partitionnement C_x trouvé par l'algorithme PARTITIONNEMENT relatif au seuil x

Choisir un seuil qui correspond au bon partitionnement revient à trouver un seuil $x' \in$ [0, 1] tel que :

$$\forall x \in [0, 1] : g'(x') \geq g'(x)$$

Si on pose $E = \{e_1, e_2, ..., e_m\}$ avec $0 \leq IP(e_1) \leq IP(e_2) \leq IP(e_3) \leq ... \leq IP(e_m) \leq 1$, alors pour deux seuils σ, σ' tel que $IP(e_i) \leq \sigma \leq \sigma' \prec IP(e_{i+1})$, le partitionnement relatif au seuil σ est le mêmes que celui relatif au seuil σ' ceci à partir de l'algorithme PARTITIONNEMENT, ce qui montre que :

$$\psi(C_\sigma) = \psi(C_{\sigma'})$$

La fonction g' est donc une fonction en escalier sur l'intervalle [0, 1] et la borne supérieur de cette fonction est l'élément $x' \in$ [0, 1] : $g'(x') = \psi(C_{x'}) = Max\{\psi(C_x) / x \in F'\}$où $F' = \{IP(e) / e \in E\}$

Maintenant on peut écrire l'algorithme pour trouver un partitionnement d'un graphe d'association de termes à partir de la méthode de l'importance d'une arête.

Algorithme: Méthode de partitionnement par la mesure d'importance d'arête.

Données: *Un graphe d'association de termes* $G = (V, E, W)$.

Résultat: *Un partitionnement C du graphe G.*

Début

(1) Poser $F' = \{IP(e) / e \in E\}$

(2) Pour chaque valeur du seuil $\sigma \in F'$.

(i) Trouver un partitionnement C_α *à partir de l'algorithme PARTITIONNEMENT.*

(ii) Calculer la mesure de qualité $\psi(C_\alpha)$ *du partitionnement* C_α.

(3)Trouver un seuil $\sigma' \in F'$ *tel* $\psi(C_{\sigma'}) = Max\{\psi(C_\sigma) / \sigma \in F'\}$.

(4) $C_{\sigma'}$ *relatif au seuil* σ' *est le partitionnement basé sur l'approche d'importance d'une arête.*

Fin

2.6. Méthode basée sur la triangulation du graphe d'association de termes

2.6.1. Le principe

La notion de clique est très importante dans le cas d'un graphe d'association de termes, puisqu'elle revient à avoir un ensemble de termes où chaque deux termes se cooccurent mutuellement, ce qui peut représenter une sous thématique donnée. Alors il est intéressant de trouver un partitionnement de graphe d'association de termes en cliques mais ce qui est plus intéressant est de trouver un partitionnement minimum en cliques.

Il est connu que le problème de partitionnement de graphe en cliques minimales est NP–complet [Karp, 1972]. Il existe quelques classes connues des graphes où le problème de partitionnement en clique minimum est résolu par un algorithme polynomial. Pour notre cas, on s'est intéressé des graphes triangulés pour deux raisons :

- On dispose d'un algorithme polynomial pour la triangulation du graphe.
- L'opération d'ajout d'arêtes dans la procédure de triangulation des graphes non triangulés convient bien à la nature du problème qu'on a abordé.

2.6.2. La méthode

Avant d'aborder la méthode du partitionnement qui se base sur la triangulation d'un graphe d'association de termes, il est intéressant de parler de la notion de triangulation minimale d'un graphe.

A) Triangulation minimale

Soit un graphe $G = (V, E)$.

Définition

$G' = (V, E \cup F)$ est une triangulation minimale de G si et seulement si $G' = (V, E \cup F)$ est un graphe triangulé et que pour tout F' tel que $F' \subset F$, le graphe $H = (V, F')$ n'est pas une triangulation de G.

Dans la figure2.4, le graphe qui est à gauche est une triangulation minimale de celui qui est à droite.

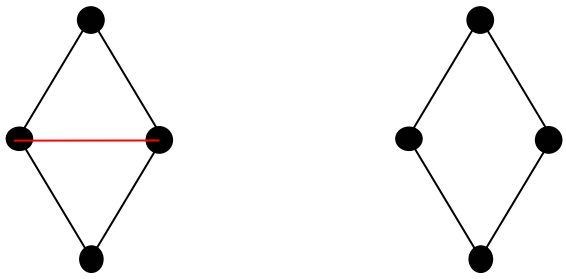


Figure2.4

Théorème [Rose & al, 1976]

Pour tout graphe G, il est possible de construire en temps $O(nm)$ une triangulation minimale de G.

L'algorithme (appelé Lex-M) qui est proposé dans [Rose & al, 1976] fournit pour tout graphe G et en temps $O(nm)$, une triangulation minimale H de G. Il utilise un ordre lexicographique [Dourisboure, 2003] basé sur un parcours en largeur du graphe. Pendant le parcours, les sommets sont numérotés par des entiers entre 1 et n.

Par la suite, $\alpha(i)$ représentera le sommet numéroté *i*. Tout sommet u possède également une étiquette, notée label(u), qui est un ensemble de valeurs prises parmi $\{1,...,n\}$ et classées par ordre décroissant.

Algorithme Lex-M

Données : *Un graphe* $G = (V, E)$

Résultat : *Une triangulation minimale de G :* $G' = (V, E \cup F)$

Début

(1) Mettre toutes les étiquettes ainsi que l'ensemble F à ϕ*;*

(2) Pour i allant de n à 1 ***faire***

(i) Sélection :

Choisir un sommet u non numéroté d'étiquette maximum;

Affecter à u le numéro i : $\alpha(i) = u$ *;*

(ii) Ajout :

Pour tout sommet non numéroté v

S'il existe une chaîne $u, x_1, \ldots x_p v$ *dans G avec* x_j *non numéroté et*

$label(x_j) \prec label(v)$ *pour tout* $j \in \{1, \ldots, p\}$ ***faire***

Ajouter i dans l'étiquette de v;

Ajouter l'arête uv dans F;

Fin

Remarque 4

Cet algorithme est aussi un algorithme de reconnaissance de graphe triangulé car si *G* est triangulé alors la triangulation minimale de *G* est lui-même.

B) Partitionnement minimale en cliques d'un graphe d'association de termes

Comme mentionné ci-dessus, la détermination d'une partition minimum en cliques est une notion très importante pour le cas d'un graphe d'association de termes puisque elle détermine des communautés où les termes de chaque communauté se cooccurent mutuellement.

Pour faire un tel partitionnement, nous allons faire une triangulation du graphe d'association de termes $G = (V, E, W)$ par l'algorithme Lex-M et à chaque arête *uv* ajouté entre les sommets *u* et *v*, on lui attribue un poids $\hat{w}_{uv} \in [0,1]$. Nous obtenons ainsi un graphe triangulé pondéré *G'* à partir du graphe d'association de termes. Ensuite, on partitionne le graphe *G'* en θ-clique

par un algorithme de partitionnement en cliques [Gavril, 1972]. Soit $C = (c_1, c_2, \ldots, c_{\theta(G')})$ un tel partitionnement où c_i est une clique dans le graphe *G'*, C représente aussi une partition des sommets du graphe *G*.

Si le graphe obtenu après une triangulation de *G* est le graphe lui-même alors *G* est triangulé et on obtient un partitionnement directe de *G* en cliques. Mais si G n'est pas triangulé alors trouver le sens des arêtes à ajoutés ainsi que les valeurs de ses poids restent des problèmes qui ne sont pas faciles à interpréter.

Les arêtes qui peuvent être ajoutés pendant la triangulation du graphe initial peuvent être interprétées comme étant le rapprochement thématique entre les termes qui ne cooccurrent pas Dans les même unités textuelles du corpus et ceci à la base du thème global abordé et de transitivité des liens entre termes (i.e. sommets dans le graphe).

Néanmoins, la valeur à donner à ce lien non apparent reste critique et difficile à spécifier, on propose ainsi de faire une étude sur cette question.

Par la suite, on considère que la valeur à proposer est la moyenne des poids des plus courtes chaînes entre les sommets à reliés ; tel que le poids d'une plus courte chaîne est le produit des poids de ses arêtes.

2.6.3 L'algorithme

Algorithme: partitionnement basée sur la triangulation du graphe d'association de termes.

Données: *Un graphe d'association de termes* $G = (V, E, W)$.

Résultat: *Un partitionnement C du graphe G.*

Début

(1) Triangularisé le graphe G en un graphe G'.

(2) Affecter un poids $\hat{w}_{uv} \in [0,1]$ *pour chaque arête uv ajouté dans la triangulation,*

(3) Partitionner le graphe G' en cliques minimum et soit C *un tel partitionnement.*

Fin

2.7. Conclusion

Dans ce chapitre, nous avons traité le problème du clustering relatif au problème de partitionnement d'un graphe.

Nous avons au début parlé du modèle de graphe d'association de termes qui correspond au graphe de termes mais avec une pondération sur ses arêtes ce qui permet de mesurer la liaison entre les termes

Le graphe d'association de termes est un graphe de petit monde donc il est constitué de communautés ainsi nous avons présenté quelques approches du partitionnement de ce graphe à partir de la triangulation d'un graphe, de la densité d'un sommet et de l'importance d'une arête.

Chapitre 3

Sur la cartographie d'un corpus textuel

3.1. Introduction

La représentation du contenu d'un corpus permet de montrer les relations entre l'ensemble de termes $\mathcal{T}$ d'un corpus textuel. Dans ce contexte, les termes de cet ensemble ne jouent pas le même rôle dans le sens où il y' a:

- des termes porteurs plus d'information que d'autres.
- des termes plus précis que d'autres.

L'un des problèmes qui se pose est de proposer une méthode de réduction de nombre de termes à considérer afin de fournir une liste de termes plus courte mais porteuse suffisamment d'information pour donner une idée générale sur le contenu et les thématiques traiter dans ce corpus. Il s'agit donc de proposer un ensemble $TR \subseteq \mathcal{T}$ qui est une synthèse qui devrait représenter les relations entre thèmes abordés. Cela revient à réduire un vaste espace d'expression à un espace plus petit et donc plus compréhensible et interprétable.

Avant de continuer cette problématique, il est indispensable de parler de la distribution zipfienne pour montrer le rôle que joue les termes dans un corpus donné.

En effet, en 1949, G. K. Zipf [Zipf, 1949] a constaté, en étudiant des corpus de données textuelles, des régularités sur des fréquences d'apparition des mots. Ainsi, après rangement des mots en fonction de leurs fréquences décroissantes, il a pu définir une distribution entre le rang et la fréquence. Selon la loi de Zipf, la courbe de cette distribution (Figure3.1) est de type $\frac{1}{x}$, elle possède une longue queue et est découpée traditionnellement en trois zones:

Zone1 : Information trivial : ensemble de mots clés triviaux et principaux qui définissent le sujet en fonction de la thématique de base.

Exemple : Dans une base de données spécialisée en « théorie de graphes », le mot « théorie des graphes » appartient à cette zone.

Zone2 : Information intéressante : ensemble de mots représentatifs du contenu de la base qui permet de construire les structures des relations des différentes approches du thème étudié.

Exemple : toujours dans la même base spécialisée en « théorie des graphes », le mot « graphe parfait » appartient, logiquement, à cette base.

Zone3 : Information marginale ou bruit : ensemble des mots non pertinents par rapport au thématiques de la base

Exemple : un mot par exemple «Auteur» peut être considéré comme un bruit dans la précédente base.

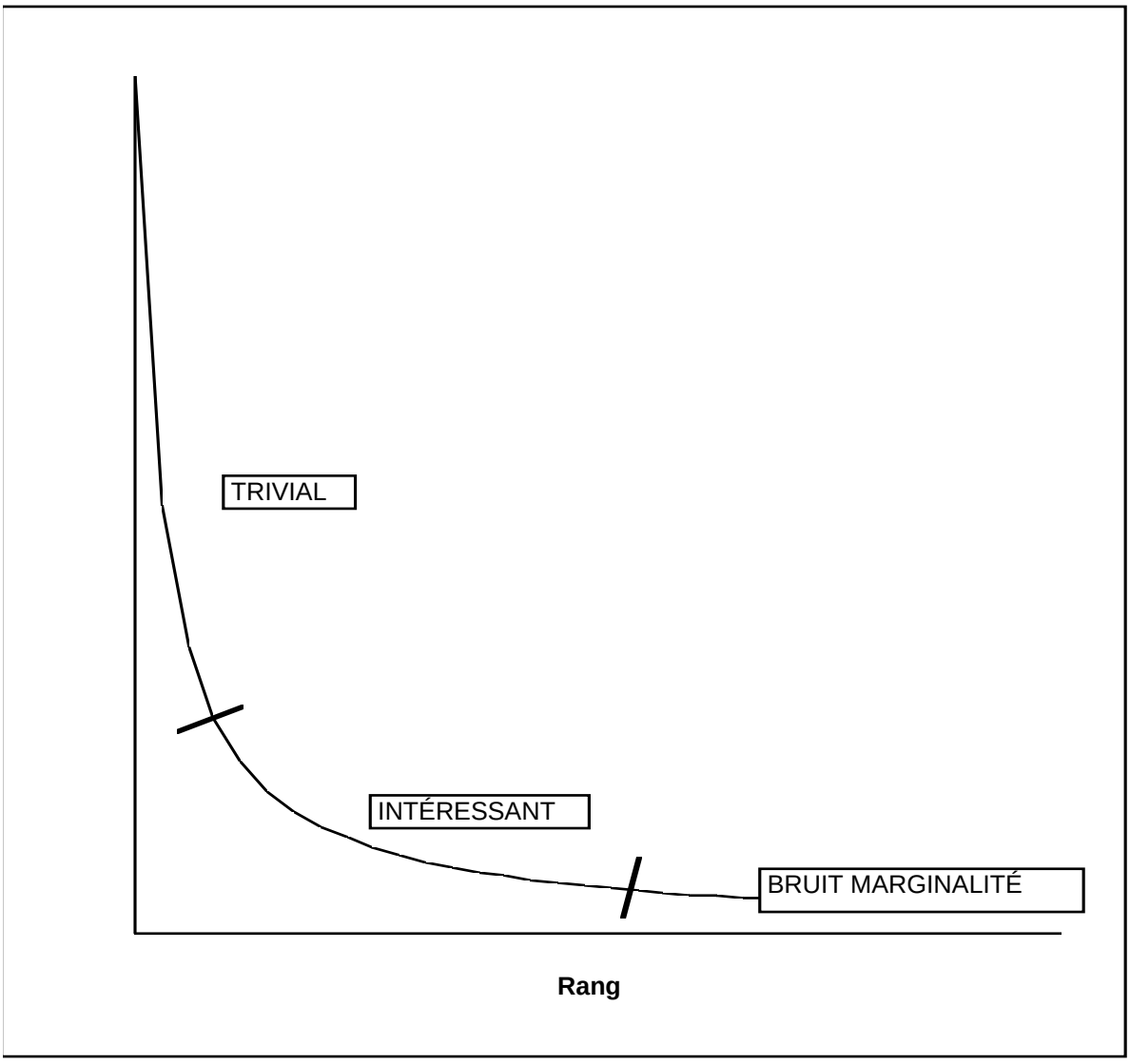


Figure3.1

Il existe dans la littérature des méthodes qui sont basées sur la statistique comme dans les travaux de [Turenne, 2000], [Han & al, 2000], [He & al, 2001], [Chen & al, 2001], [Besançon, 2001] mais l'applications de ces méthodes est limitée dans la mesure où elles sont basées principalement sur les fréquences d'occurrences [Salton & al, 1975] et ne prennent pas en considération les associations des termes ce qui pénalise une partie importante de termes. Un des derniers travaux est l'étude faite par Mokrane et autre [Mokrane & al, 2004], leur

approche est purement statistique, elle est basée sur l'occurrence de termes et les associations qu'entretiennent ces termes.

Dans ce chapitre, nous proposerons une méthode qui permet de représenter les termes qui sont centraux et qui donnent une vue globale sur le contenu du corpus à partir de la structure du réseau de termes associé (voir le paragraphe 3.2). L'objectif consiste à proposer une méthode à des fins de visualisation et d'analyse thématique de textes.

3.2. Réseau de termes associé

Dans le cas d'un graphe de termes, les relations d'influence entre les termes ne sont pas très bien explicitées. Ainsi, s'il existe une arête entre deux termes *i* et *j* dans un graphe de termes alors rien ne peut nous indiquer qui a plus d'influence sur l'autre.

Pour bien montrer ces relations entre les termes, nous proposons un réseau de termes associé qui est un graphe orienté et pondéré ceci en remplaçant chaque arête *ij* de *E* dans un graphe de termes par un arc pondéré d'un terme de plus petite valeur d'occurrence vers un terme de plus grande occurrence (Figure3.2).

Figure3.2

Définition

Un réseau de termes associé $R = (V, U, W')$, relative au corpus Γ dans un contexte donné, est un graphe orienté et pondéré tel que :

- $V = \mathcal{T}$.

- *W'* : fonction degré d'association définie de $\mathcal{T} \times \mathcal{T}$ vers [0, 1] telle que $\forall\ i \in V, \forall\ j \in V$:

$W'(i, j) = \Im^{j}_{i} = \dfrac{t_{ij}}{t_i}$ notée w_{ij}'.

- $\forall\ i \in V, \forall\ j \in V : ij \in U \Leftrightarrow w_{ij}' = max(w_{ij}', w_{ji}') > 0 \Leftrightarrow t_j \geq t_i$ et $t_{ij} > 0$.

Le poids w_{ij}' de l'arc *ij* reflète donc le rapprochement sémantique de terme *i* vers le terme *j*: lorsqu'il vaut 1, le terme *i* est toujours associé au terme *j*, lorsqu'il vaut 0, le terme *i* n'est pas associé au terme *j*.

Ce graphe permet de montrer et d'expliciter l'influence entre un terme et son cooccurrent.

3.3. Notion d'ensemble générique

A partir de la définition du réseau de termes associé, les hypothèses de bases qui sous-tend notre méthode part de deux observations : La première est que si on a un arc de poids important d'un terme *i* vers un terme *j* alors si on parle de terme *i* donc on parle de terme *j* et on peut dire que le terme *j* est plus générique que le terme *i* ou bien que le terme *j* est un terme référent du terme *i*. La seconde part de l'observation qui est mentionnée dans l'article [Véronis, 1996] où Jean Véronis a observé que les degrés et fréquences des termes sont très fortement corrélés, de façon presque linéaire c'est-à-dire que plus le degré d'un terme dans un graphe de termes est élevé plus sa fréquence dans le corpus est grande.

Considérons alors le problème qui revient à déterminer un ensemble de termes $\mathcal{T}$ qu'on le nomme « ensemble générique représentatif» tel que pour n'importe quel terme on peut toujours trouver un chemin qui mène de ce terme vers un terme de cet ensemble générique.

Ainsi, l'idée est qu'un ensemble générique représentatif du réseau de termes associé permettra d'appréhender le contenu du corpus en question. En terme de graphe, ça revient à déterminer un ensemble qui vérifie un certain nombre de conditions (voir la définition ci-dessous).

Définition

Soit $G = (V,U)$ un graphe orienté, un ensemble générique $S \subset V$ est un ensemble qui vérifie les conditions suivantes :

(C1) $\forall v \in V/S$, $\exists\, v' \in S$ tel qu'il existe un chemin de *v* vers *v'*.
(C2) *S* est minimal par la condition (C1).

Un ensemble générique de termes *S** est dit représentatif de $G = (V,U)$ si et seulement si pour n'importe quel ensemble générique *S*: $\sum_{v \in S} d^-(v) \leq \sum_{v \in S^*} d^-(v)$.

Exemple

Dans le graphe $G = (V,U)$ de la figure1.10, l'ensemble $S^* = \{a, e\}$ est un ensemble générique représentatif.

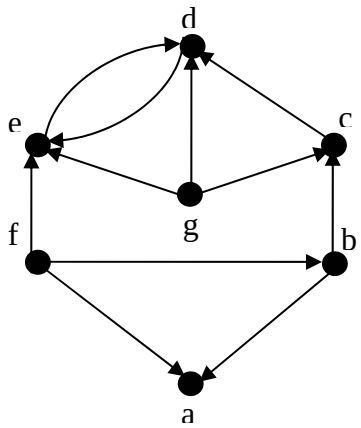


Figure1.10- *Un graphe orienté*

Remarque :

L'ensemble $S' = \{a, d\}$ est aussi un ensemble générique représentatif, alors dans un graphe orienté, l'ensemble générique représentatif n'est pas unique.

Dans ce qui suit, on s'intéresse à la recherche de cet ensemble dans un graphe orienté.

Posant :

$B = \{ C \subset V \ / \ C$ est une composante fortement connexe et $\Gamma^{+}(C) = \phi \}$, alors B représente l'ensemble de toutes les composantes fortement connexes terminales.

A partir de la définition précédente on peut montrer le lemme suivant :

Lemme 1:

Si S un sous ensemble de V d'un graphe orienté $G = (V,U)$ qui vérifie la condition (C1) alors $\forall C \in B, \exists v \in C : v \in S$.

Preuve:

Supposons le contraire, alors $\exists\ C' \in B$ tel que : $\forall\ v \in C', v \notin S$ comme $\Gamma^{+}(C') = \phi$ $(C' \in B)$ donc si on prend un sommet v dans C' il n'existe aucun chemin du ce sommet v vers un sommet de S, contradiction ■

A partir de ce lemme on peut montrer la propriété suivante

Propriété1:

Soient $G = (V,U)$ un graphe orienté, S un sous ensemble de V alors:

S est un ensemble générique $\Leftrightarrow \forall C \in B : \left| S \cap C \right| = 1$ et $S \cap \left(V - \bigcup_{C_i \in B} C_i \right) = \phi$.

Preuve:

a) Montrons la condition nécessaire

Soit S est un ensemble générique donc S vérifie les conditions (C1) et (C2), montrons que $\forall C \in B : \left| S \cap C \right| = 1$ et que $S \cap \left(V - \bigcup_{C_i \in B} C_i \right) = \phi$.

Supposons le contraire alors ou bien $\exists C \in B$ tel que $\left| S \cap C \right| \neq 1$ ou bien $S \cap \left(V - \bigcup_{C_i \in B} C_i \right) \neq \phi$.

a_1) Si $\exists C \in B$ tel que $|S \cap C| \neq 1$

Dans ce cas, $|S \cap C| \geq 2$ ou $|S \cap C| = 0$

a_{11}) Si $|S \cap C| \geq 2$

Alors, $\exists \{v_1, v_2\} \in V$ tel que $\{v_1, v_2\} \in S \cap C$ or C est une composante fortement connexe terminale donc il existe un chemin de v_1 vers v_2 et vice versa.

Posant $S'=S - \{v_1\}$. S' est un ensemble vérifiant la condition (C1), en effet soit un sommet v dans V / S' alors il existe un chemin de v vers un sommet v' de S (car S ensemble générique)

- Si $v' = v_1$: alors il existe un chemin de v vers v_1 mais comme il existe un chemin de v_1 vers v_2 donc il existe un chemin de v vers un sommet v_2 de S'.

- Sinon $v' \neq v_1$: alors il existe un chemin de v vers un sommet v' de S' (car $S'=S - \{v_1\}$).

Ce qui montre que l'ensemble S' vérifie la condition (C1), ainsi l'ensemble S n'est pas minimale par la propriété (C1) donc il n'est pas générique, contradiction.

a_{12}) Si $|S \cap C| = 0$

C'est-à-dire que les ensembles S et C sont disjoints, or S est un ensemble générique et C est une composante connexe terminale, ce qui contredit le lemme1.

Ainsi, si S est un ensemble générique alors $\forall C \in B : |S \cap C| = 1$.

a_2) Si $S \cap \left(V - \bigcup_{C_i \in B} C_i \right) \neq \phi$

Dans ce cas, il existe un sommet w_1 dans V tel que $w_1 \in S \cap \left(V - \bigcup_{C_i \in B} C_i \right)$ ce qui implique que $\forall C \in B, w_1 \notin C$, donc il existe une composante fortement connexe Q tel que $w_1 \in Q$ et $\Gamma^+(Q) \neq \phi$, ainsi $\exists C' \in B, \forall v \in C'$, il existe un chemin de w_1 vers v et en particulier pour $w_2 = S \cap C$ ($w_1 \neq w_2$).

Posant $S'=S - \{ w_1 \}$. S' est un ensemble vérifiant la condition (C1), en effet soit un sommet $w \in V / S'$ alors il existe un chemin de w vers un sommet $v' \in S$ (car S ensemble générique)

- Si $v' = w_1$: alors il existe un chemin de w vers w_1 mais comme il existe un chemin de w_1 vers w_2 donc il existe un chemin de w vers un sommet w_2 de S'.

- Sinon $v' \neq w_1$: alors il existe un chemin de v vers un sommet v' de S' (car $S'=S - \{ w_1 \}$).

Ce qui montre que l'ensemble S' vérifie la condition (C1), ainsi l'ensemble S n'est pas minimale par la propriété (C1) donc il n'est pas générique, contradiction.

Donc, $S \cap \left(V - \bigcup_{C_i \in B} C_i \right) = \phi$.

Ainsi, S est un ensemble générique $\Rightarrow \forall C \in B : \left| S \cap C \right| = 1$ et $S \cap \left(V - \bigcup_{C_i \in B} C_i \right) = \phi$.

b) Montrons la condition suffisante.

Soit S un sous ensemble de sommets de V vérifiant $\forall C \in B : \left| S \cap C \right| = 1$ et $S \cap \left(V - \bigcup_{C_i \in B} C_i \right) = \phi$, nous montrons que S est un ensemble générique (i.e S vérifie les conditions (C1) et (C2)).

b_1) Montrons que S vérifie la condition (C1)

Soit $v \in V / S$ alors on distingue deux cas.

1er Cas : $\exists\, C \in B : v \in C$.

Dans ce cas il existe un chemin de v vers l'élément $v' = S \cap C$ puisque C est une composante fortement connexe, ainsi il existe un chemin de v vers $v' \in S$.

2ème Cas : $\forall\, C \in B : v \notin C$.

Dans ce cas, soit Q une composante fortement connexe tel que $v \in Q$ et $Q \notin B$, donc $\Gamma^+(Q) \neq \phi$ et $\exists C' \in B$ tel qu'il existe un chemin de $v \in Q$ vers n'importe quel sommet $w \in C'$

et en particulier il existe un chemin de *v* vers un sommet *v'* $=S \cap C_i$ ce qui implique l'existence d'un chemin de *v* vers un sommet de *S*.

Ainsi, dans tous les cas il existe un chemin de *v* vers un sommet de *S* ce qui montre que *S* vérifie la condition (C1).

b_2) Montrons que *S* vérifie la condition (C2)

Supposons le contraire c'est-à-dire que *S* n'est pas minimal par la condition (C1) alors il existe un sommet v_i dans *S* tel que $S' = S - \{v_1\}$ vérifie la condition (C1).

Le sommet v_i de *S* appartenant à une composante fortement connexe *C*, mais comme $S \cap \left(V - \bigcup_{C_i \in B} C_i \right) = \phi$ alors $C \in B$,

Puisque *S'* vérifie la condition (C1) alors il existe un chemin d'un sommet $v \in V / S'$ vers un sommet $v' \in S'$, en particulier il existe un chemin de $v_i \in V / S'$ vers un sommet $v_j \in S'$ qui appartient à une autre composante fortement connexe $C' \in B$ ($C \neq C'$) donc $\Gamma^+(C) \neq \phi$ donc *S* est minimal, contradiction

Ce qui montre que : *S* est un ensemble générique $\Leftrightarrow \forall C \in B : |S \cap C| = 1$ et $S \cap \left(V - \bigcup_{C_i \in B} C_i \right) = \phi$ ■

Corollaire 1:

Soient $G = (V, U)$ un graphe orienté, *S* un sous ensemble de *V*, alors :
S est un ensemble générique $\Rightarrow$ *S* est un stable.

Preuve:

Soient $G = (V, U)$ un graphe orienté *S* est un ensemble générique, on distingue deux cas de figures :

a) $|S| = 1$

Dans ce cas il est évident que *S* est un stable.

b) $|S| \geq 2$:

D'après la propriété1 : $\forall C \in B : |S \cap C| = 1$ et $S \cap \left(V - \bigcup_{C_i \in B} C_i \right) = \phi$. Soient *v* et *v'* deux sommets dans *S* alors $\exists C \in B$ et $\exists C' \in B$ tel que $v \in C$ et $v' \in C'$. Si on suppose que $vv' \in U$ donc $\Gamma^{+}(C) \neq \phi$ contradiction car $C \in B$

Ainsi, *S* est un ensemble générique $\Rightarrow$ *S* est un stable ■

Corollaire 2:

Soient $G = (V, U)$ un graphe orienté, *S* et S' deux sous ensembles de *V* alors :
Si *S* et *S'* sont deux ensembles génériques $\Rightarrow |S| = |S'|$.

Preuve:

La preuve est évidente compte tenu de la propriété1 ■

Remarque :

D'après ce corollaire, la condition (C2) (ie *S* est minimal par la condition (C1)) peut être remplacée par la détermination d'un ensemble *S* qui soit minimum par la condition (C1).

Corollaire 3:

Soient $G = (V, U)$ un graphe orienté, *S** un sous ensemble de *V* alors
*S** est un ensemble générique représentatif $\Leftrightarrow$
1. $\forall C \in B : |S^* \cap C| = 1$.
2. $S^* \cap \left(V - \bigcup_{C_i \in B} C_i \right) = \phi$
3. $\forall v \in S^*, \exists C \in B : v \in C \ et \ d^{-}(v) = \max_{v' \in C} \{ d^{-}(v') \}$

Preuve:

a) Montrons la condition nécessaire

Soit *S** est un ensemble générique représentatif donc *S**, montrons que :

1. $\forall C \in B : \left| S^* \cap C \right| = 1.$
2. $S^* \cap \left(V - \bigcup_{C_i \in B} C_i \right) = \phi$
3. $\forall v \in S^*, \exists C \in B : v \in C \ \ et \ \ d^-(v) = \max_{v' \in C} \left\{ d^-(v') \right\}$

a_1) Montrons $\forall C \in B : \left| S^* \cap C \right| = 1 \text{et} S^* \cap \left(V - \bigcup_{C_i \in B} C_i \right) = \phi$

Soit *S** un ensemble générique représentatif alors *S** est générique et donc d'après la propriété1 : $\forall C \in B : \left| S^* \cap C \right| = 1 \text{et} S^* \cap \left(V - \bigcup_{C_i \in B} C_i \right) = \phi$.

a_2) Montrons $\forall v \in S^*, \exists C \in B : v \in C \ \ et \ \ d^-(v) = \max_{v' \in C} \left\{ d^-(v') \right\}$

Soit *S** un ensemble générique représentatif alors pour n'importe quel ensemble générique *S*: $\sum_{v \in S} d^-(v) \leq \sum_{v \in S^*} d^-(v)$ or d'après la propritée1 : $\forall C \in B : \left| S^* \cap C \right| = 1$ et $S^* \cap \left(V - \bigcup_{C_i \in B} C_i \right) = \phi$ donc on doit choisir un et un seul sommet pour chaque ensemble *C* de *B* et pour maximiser $\sum_{v \in S^*} d^-(v)$ il suffit de choisir un sommet *v* dans chaque composante connexe *C* de *B* qui vérifie $d^-(v) = \max_{v' \in C} \left\{ d^-(v') \right\}$.

b) Montrons la condition suffisante.

*S** est un ensemble de sommet dans *V* tel que :

1. $\forall C \in B : \left| S^* \cap C \right| = 1.$
2. $S^* \cap \left(V - \bigcup_{C_i \in B} C_i \right) = \phi$
3. $\forall v \in S^*, \exists C \in B : v \in C \ \ et \ \ d^-(v) = \max_{v' \in C} \left\{ d^-(v') \right\}$

Montrons que *S** est ensemble générique représentatif

Puisque $\forall C \in B : \left| S^* \cap C \right| = 1$et $S^* \cap \left(V - \bigcup_{C_i \in B} C_i \right) = \phi$ alors d'après la propriété1 *S** est un ensemble générique et comme $\forall v \in S^*, \exists C \in B : v \in C$ *et* $d^-(v) = \max_{v' \in C} \left\{ d^-(v') \right\}$ alors pour n'importe quel ensemble générique *S*: $\sum_{v \in S} d^-(v) \leq \sum_{v \in S^*} d^-(v)$ ce qui montre que *S** est ensemble générique représentatif ■

Ce dernier corollaire nous permet de proposer l'algorithme suivant:

Algorithme de détermination d'un ensemble *S**

Données : *Un graphe orienté* $G = (V, U)$

Résultat *: Un ensemble générique représentatif S*.*

Début

(1) *Partitionnement de graphe* $G = (V, U)$ *en composantes fortement connexes maximales, soit* $C_1, C_2, ..., C_p$..

(2) *Trouver l'ensemble K tel que K={i / i∈{1,..p} et* $\Gamma^+(C_i) = \phi$*}, avec* $\Gamma^+(C_i)$ *ensemble de successeurs des sommets de* C_i *dans* $\mathcal{T} / C_i$

(3)

- *Mettre* $S^* = \phi$
- *Pour chaque* $k \in K$*, choisir un sommet* $v \in C_k$ *tel que* $d^-(v) = \max_{v' \in C_k} \left\{ d^-(v') \right\}$, *mettre* $S^* = S^* \cup \{v\}$.

Fin

L'algorithme de détermination des composantes fortement connexes d'un graphe orienté se fait en $O(|V| + |U|)$ [Tarjan, 1972].

3.4. Cartographie de termes

La cartographie d'information joue un rôle central dans la dynamique de nombreux phénomènes informationnels qui peuvent être modélisés par des graphes dont les sommets représentent les acteurs du phénomène et les liens représentent les interactions entre eux sous forme de cartes de navigation tel que dans Kartoo [Chung & al, 2003] ou Mapstan [Spinat, 2002]. Ce qui permet de synthétiser des informations sous une forme facile à interpréter et exploiter.

Nous allons proposer une méthode d'élaboration d'une cartographie qui est basée sur l'idée que l'utilisateur ait une vue globale du réseau et qu'il puisse passer d'une vue à une autre afin de mieux expliciter et comprendre le contenu d'un corpus textuel donné à partir de la notion d'un ensemble générique représentatif d'un réseau de termes associé d'un corpus textuel.

Nous proposerons ainsi une cartographie d'information du contenu à trois niveaux de visualisation :

a) Première vue

La première vue consiste à proposer une vue globale du réseau. Cette vue globale doit être une information générale à l'utilisateur et elle lui permet d'élaborer une stratégie d'exploration.

Le processus général dans cette étape est la suivante:

- Construction d'un réseau de termes associé $R = (V, U, W')$ à partir d'un ensemble de termes et des cooccurrences entre ces termes du corpus.
- Trouver un ensemble génériques représentatif S^* du réseau $R = (V, U, W')$.
- Elaboration d'une cartographie à partir de l'ensemble S^* qui donne une vue synthétique du contenu de ce corpus à partir de l'ensemble des termes génériques représentatifs ou référents du corpus.

b) Deuxième vue

Cette étape est particulièrement adaptée quand l'utilisateur découvre le réseau et qu'il ne sait pas, a priori, comment se regroupent les termes. En effet, les clusters permettent de dégager des pôles et de distinguer les grands sous thèmes du domaine d'analyse du corpus. Ainsi, dans cette étape l'utilisateur identifie des groupements de termes. Pour chaque groupement, on doit le représenter par un ensemble qui lui donne une vue globale et synthétique de son contenu et

dans notre cas ces sommets sont des termes génériques représentatifs et on construit une cartographie qui dont les termes sont la réunion de tous ces termes génériques représentatifs des clusters.

Le processus général dans cette étape est le suivant:

- Construction d'un réseau de termes associé $R = (V, U, W')$ à partir d'un ensemble de termes et des cooccurrences entre ces termes.
- Construire un graphe de termes pondéré $G = (V, E, W)$ à partir de $R = (V, U, W')$ ceci par le remplacement d'un arc qui peut exister entre les termes *i* et *j* par une arête de même poids.
- Trouver un partitionnement $C = (c_1, c_2, \ldots, c_k)$ du graphe $G = (V, E, W)$, à partir des méthodes de partitionnement proposés dans le chapitre 2.
- Pour chaque communauté c_i on construit un réseau de termes associé R_{c_i} induit par c_i
- Sélectionner à partir de R_{c_i} un ensemble générique S_i *.
- Elaboration d'une cartographie de termes $R_{\bigcup_{i=1}^{i=p} S_i *}$ qui est un sous graphe induit par $\bigcup_{i=1}^{i=p} S_i *$ du réseau de termes associé *R*.

c) Troisième vue

Cette étape consiste à donner tous les termes $\mathcal{T}$ du corpus et de montrer les interactions et les relations qui peuvent exister entre eux. En terme de graphe ceci revient à construire un réseau de termes associé $R = (V, U, W')$ à partir d'un ensemble de termes et des cooccurrences entre ces termes du corpus.

3.5. Conclusion

Dans ce chapitre on s'est intéressé à la notion d'ensemble générique représentatif d'un corpus textuel. Après avoir modéliser le problème sous forme d'un réseau pondéré appelé réseau de termes associé. Nous avons montré que cet ensemble peut être trouvé de manière efficace en un temps polynomial. En combinant ces résultats avec ceux du chapitre précédent, nous avons proposé un modèle de cartographie d'un corpus textuel et ce à trois niveaux :

- Une vue globale d'un réseau de termes associé.
- Une vue locale qui permet de dégager des pôles représentés par l'ensemble générique représentatif de chaque cluster.
- Une vue qui permet de montrer les interactions de tous les termes de l'ensemble $\mathcal{T}$.

Ces représentations permettent à l'utilisateur de consulter facilement le contenu du corpus textuel en question et de comprendre les thématiques abordées.

L'approche proposée peut être facilement mise sous forme d'une application informatique, la seule difficulté est de disposer d'un bon outil d'extraction de termes candidats, i.e. l'ensemble $\mathcal{T}$ (voir Annexe B).

Conclusion générale

CONCLUSION GENERALE

Suite à l'essor de la nouvelle technologie de l'information, le nombre de documents disponibles dans des bases de données documentaires croit d'une manière exponentielle. L'un des enjeux majeurs aujourd'hui consiste donc à développer des outils permettant l'exploration du contenu de cette masse d'information ceci par l'analyse de contenu d'un corpus textuel qui s'intéresse aux significations du texte, aux indications qu'il apporte sur le sujet étudié.

Dans notre contribution, nous nous sommes intéressés à l'étude d'un modèle de graphe dans le domaine de la science de l'information pour l'élaboration d'une cartographie d'information du contenu d'un corpus textuel :

- Nous avons proposé des méthodes du partitionnement des graphes d'associations de termes ceci afin d'extraire les communautés homogènes de termes.

- Après, on a défini un modèle du réseau de termes associé qui montre les relations d'influences entre les termes. Nous avons défini la notion d'ensemble générique représentatif d'un réseau de terme associé pour permettre la visualisation globale du contenu d'un corpus textuel, ainsi qu'un algorithme de recherche de cet ensemble.

Les résultats ainsi obtenus nous permettent de proposer des méthodes de visualisation du contenu d'un corpus textuel.

Comme perspectives, nous proposons de combiner nos travaux avec les cartes causales [Yalaoui & al, 2005a], [Yalaoui & al, 2005b] pour construire un moteur de recherche avancé sur les textes, permettant également de faire une analyse profonde.

Références bibliographies

[Adamic, 1999] Adamic. “The small world web”. Rapport technique, Xerox Palo Alto Research Center, Los Angeles. 1999

[Arthanary & al, 1980]. Arthanary. T.S et Dodge. Y. “Mathematical Programming in Statistics”. John Wiley & Sons. 1980

[Berkhin, 2002]. Pavel Berkhin. “Survey of clustering data mining techniques”. Rapport technique, Accrue Software, San Jose, CA. 2002

[Besançon, 2001]. R. Besançon. “Intégration de connaissances syntaxiques et sémantiques dans les représentations vectorielles des textes”. Thèse en Informatique, Ecole polytechnique Fédérale de Lausanne. 2001

[Boutin & al, 2004]. François. Boutin et Mountaz. Hascoët. “Cluster validity indices for graph partitioning”. Proceedings of the conference on Information Visualisation IV2004. 2004

[Bradford, 1934]. S. C. Bradford. “Sources of information on specific subjects”. Engineering, vol 137, pp 85-86. January 1934

[Breuer, 1977]. Melvin A. Breuer. “A class of min-cut placement algorithm”. In Proc. of the Design Automation Conference, pages 284–290. IEEE Press. 1977

[Broder, 2000]. A. Broder, R. Kumer, F. Maghoul, S. Rajagopala. et all. “Graph structure in the web”. Computer Networks 33. pp 309-320. 2000

[Chen & al, 2001]. H. Chen, H. Fan, M. Chau et D. Zeng. “MetaSpider: Meta-searching and categorization on the Web”. Journal of the American Society for Information Science and Technology, vol. (52) : pp 1134-1147. 2001

[Chung & al, 2003]. Chung W., Chen H et Nunamaker. J. “Business intelligence explorer : A knowledge map framework for discovering business intelligence on the Web ”. Proceedings of the 36 Hawaii International Conference on System Sciences (HICSS’03), Hawaii. 2003

[Dahmane, 1990]. Madjid. Dahmane. “Contribution à l’étude des systèmes d’information scientifique et technique : approche théorique et étude de cas de l’Algérie”. Thèse de Doctorat, l’université de Bordeaux III, Institut des sciences de l’information et de la communication. 1990

[Danlos, 2000]. Danlos. L. “Génération automatique de textes”. In. J.M. Pierrel (ed.), Ingénierie des Langues. Hermes Science, Paris. 2000

[Davenport & al, 1998]. Davenport T.H., Pruzak. L. “Working Knowledge: How organizations manage what they know”. Boston, Massachusetts: Harvard Business School Press. 1998

[Day, 1996]. W. H. E. Day. “Complexity theory : An introduction for partitioners of classification, in clustering and classification ”. P. Arabie, L. Hubert, G. DE. Soete (éds). World Scientifique, Singapore, New Jersey, London, Hong Kong, pp 199-233. 1996

[Denoeud & al, 2005]. L. Denoeud, I. Haron, A. Guénoche, O. Hudry. “Classes empiétantes aux interactions entre protéines ”. 6ème conférence de la Société Française de Recherche Opérationnelle et d’Aide à la Décision (ROADEF’05). Ecole Polytechnique de l’Université de Tours, France. 2005

[Dirac, 1961]. G. A. Dirac. “On rigid circuit graphs ”. Abh.Math.Sem.Uni.Humburg 25, pp71-76. 1961

[Dourisboure, 2003]. Par Yon. Dourisboure. “Routage compact et longueur arborescente”. Doctorat, Spécialité Informatique. Université Bordeaux I. Décembre 2003

[Dragulanescu, 2003]. Nicolae. George. Dragulanescu. “De nouveaux modèles pour les sciences de l’information? ”. Communication, CIFSIC Bucarest. 2003

[Dubois & al, 1994]. Dubois J, Guespin L, Giacomo M, Marcellesi C, Marcellesi J.-B, Mével J.-P. “Dictionnaire de linguistique et des sciences du langage”. Collection Trésors du Français, Larousse, Paris. 1994

[Fell & al, 2000]. Fell et Wagner. “The small world of metabolism”. Nature Biotechnology, pp 1121-1122. 2000

[Fiduccia & al, 1982]. C. M. Fiduccia et R. M. Mattheyses. “A linear-time heuristic for improving network partitions”. In Proceedings of the 19th IEEE Design Automation Conference, pages 175–181, IEEE Press. Las Vegas, Nevada, (USA). June 1982

[Flake, 2002]. Gary William Flake, Steve Lawrence, C. Lee Giles, and Frans M. Coetzee. “Self-organization and identification of web communities”. Computer, 35(3), pp 66-71. 2002

[Fulkerson & al, 1965]. D.R Fulkerson, O.A.Gross . “Incidence matrices and interval graphs”. Pacific Journal Of Mathematics, 15, 835-855. 1965

[Garey & al, 1979]. M. Garey et D. Johnson. “Computers and Intractability : A Guide to the Theory of NP-Completeness”. W.H. Freeman and Company. ISBN : 0716710455. June 1979

[Gavril, 1972]. F. Gavril. “Algorithms for minimum coloring, maximum clique, minimum covering by cliques, and maximum independant set of a chordal graph”. SIAM Journal Comb. 1 : 180-187. 1972

[Gerlhof & al, 1993]. Carsten A. Gerlhof, Alfons Kemper, Christoph Kilger et Guido Moerkotte. “Partition-based clustering in object bases : From theory to practice”. In Proc. of the Intl. Conf. on Foundations of Data Organization and Algorithms, volume 730 of Lecture Notes in Computer Science, pages 301–316, Springer-Verlag. Chicago, IL, (USA). October 1993

[Govaert, 2003]. Govaert. G. “Analyse des données”. Hermès. 2003

[Guénoche & al, 2003]. Alain. Guénoche, Tristan. Colombo et Yves. Quentin. “Recherche de zones denses dans un graphe: application aux gènes orthologue”. Colloque Knowledge Discovery and Discrete Mathematics, Actes des Journées Informatiques de Metz, INRIA, pp 203-212. 2003

[Habert, 1997]. Benoit. Habert. “Les linguistiques du Corpus”. Armand Colin, Paris. 1997

[Haddad, 2002]. Mohamed. Hatem. Haddad. “ Extraction et impact des connaissances sur les performances des systèmes de recherche d'information”. Thèse de doctorat. Specialité Informatique. Université de Joseph- Fourier, Grenoble 1. 2002

[Hajnal & al, 1958]. A. Hajnal, J. Suranyl. “Uber die Auflosung von Graphen in vollstandige Teilgraphen”. Annals of University of Sciences Budapest, Eötvös sect. Math.1 :113-121. 1958

[Han & al, 2000]. J. Han et M. Kamber. “ Data Mining : Concepts and Techniques”. Morgan Kaufmann Publishers. 2000

[Hansen & al, 1997]. P. Hansen, B. Jaumard. “ Cluster analysis and mathematical programming”. Mathematical programming 79, pp191-215. 1997

[He & al, 2001]. X. He, C. Ding, H. Zha et H. Simon. “ Automatic topic identification using Webpage clustering”. In Proceedings of 2001 IEEE International Conference on Data Mining, Los Alamitos, CA. 2001

[Jain & al, 1988]. A.-K. Jain, R.-C. Dubes. “Algorithms for Clustering Data”. Prentice Hall Advanced Reference Series, 1988.

[Jourdan, 2004]. Fabien. Jourdan. “Visualisation d'information : dessin, indices structuraux et navigation. Applications aux réseaux biologiques et aux réseaux sociaux”. Thèse de Doctorat en Informatique, Université de Montpellier II. 2004

[Karp, 1972]. R. M. Karp. “Reducibility among combinatorial problems”. Dans R.E Miller et J.W Tacker eds, complexity of compter computations (plenum press, New-york), pp 85-104. 1972

[Kernigan & al, 1970]. Brian W. Kernighan et S. Lin. “An efficient heuristic procedure for partitioning graphs”. The Bell System Technical Journal, 49(2) :291–307. February 1970

[Lotka, 1926]. A. J. Lotka. “The frequency distribution of scientific productivity”. Journal of the washington academy of sciences, vol 16, N° 12, p 317-323. june 1926

[Lukes, 1974]. Joseph A. Lukes. “Efficient algorithm for the partitioning of trees”. IBM Journal of Research and Development, 18(3) :217–224. May 1974

[Mancoridis & al, 1998]. Spiros Mancoridis, Brian S. Mitchell, Rorres. C, Yih-Farn Chen, et Emden R. Gansner. “ Using automatic clustering to produce high-level system organizations of source code”. IWPC. 1998

[Mancoridis & al, 1999]. Spiros Mancoridis, Brian S. Mitchell, Yih-Farn Chen, et Emden R. Gansner. “Bunch : A clustering tool for the recovery and maintenance of software system structures”. Dans ICSM. 1999

[Manolis & al, 1991]. Manolis M. Tsangaris and Jeffrey F. Naughton. “A stochastic approach for clustering in object bases”. In Proc. of the 1991 ACM SIGMOD International Conference on Management of Data, pages 12–21, Denver, CO, (USA). ACM Press. May 1991

[Manolis & al, 1992]. Manolis M. Tsangaris and Jeffrey F. Naughton. “On the performance of object clustering techniques”. In Proc. of the 1992 ACM SIGMOD International Conference on Management of Data, pages 144–153, San Diego, CA, (USA). ACM Press. June 1992

[Matsuo & al, 2001]. Y. Matsuo, Y. Ohsawa et M. Ishizuka. “KeyWord: Extracting keywords from a document as a small world”. In Proceedings the Fourth International Conference on Discovery Science. 2001

[Milgram, 1967]. S. Milgram. “ The small world problem”. Psychology today, 2, pp 60-67. 1967

[Mokrane & al, 2004]. A. Mokrane, P. Poncelet, R. Arezki, G. Dray. “ Cartographie automatique du contenu d’un corpus de documents textuels”. In proceeding of the 7 th International Conférence on the Statistical Analysis of Textual Data JADT, vol 2, pp 816-823, Louvain-la-Neuve : Presses Universitaires de Louvain. mars 2004

[Molino, 1982]. Enzo. Molino. “Bases de données: considérations intéressant les pays en développement ”. RUSIBA, vol IV, n 4.1982

[Paradis, 1996]. François. Paradis. “ Un modèle d'indexation pour les documents textuels structurés ”. Thèse de Doctorat. Spécialité : Informatique. L'université Joseph Fourier - Grenoble 1. 1996

[Rajman & al, 1997]. M. Rajman, R. Besançon. “Text Mining: Natural Language techniques and Text Mining applications”. Proceedings of the 7th IFIP 2.6 Working Conference on Database Semantics (DS-7). 1997

[Rose & al, 1976]. Rose. D, Tarjan. E et Lueker. G. S. “ Algorithmic aspects of vertex elimnation on graphs”. SIAM Journal on Computing, 5, pp 266-283. 1976

[Sakarovitch, 1983]. M. Sakarovitch. “Techniques Mathématiques de la recherche Opérationnelle”. Optimisation combinatoire. Mai 1983

[Salton & al, 1975]. G. Salton, A. Wong, C. S. Yang. “ A vector space model for automatic indexing”. Communications of the ACM, 18(11):613–620. 1975

[Saporta & al, 2003]. Saporta. J et Stefanescu. V. “Analiza datelor in informatica”. Ed. Economica. Bucarest. 1996

[Shahookar & al, 1991]. Shahookar K et Mazumder. P. “VLSI cell placement techniques”. ACM Computing Surveys, 23(2) :143–220. June 1991

[Shannon & al, 1949]. Claude E. Shannon, Warren Weaver. “The Mathematical Theory of Communication”. The University of Illinois Press, Urbana, Illinois, 1949

[Shannon, 1956]. Claude E. Shannon. “The zero error capacity of noisy channel”. Institute of Radio Engineers, Transactions on Information Theory, IT-2, pp8-19. 1956

[Shneiderman, 1996]. Ben Shneiderman. “The eyes have it : A task by data type taxonomy for information visualizations”. Rapport technique UMCP-CSD CS-TR-3665, College Park, Maryland 20742, U.S.A. 1996

[Spinat, 2002]. Spinat. E. “Pourquoi intégrer des outils de cartographie au sein des systèmes d’information de l’entreprise ? ”. Colloque Cartographie de l’information : De la visualisation à la prise de décision dans la veille et le management de la connaissance, Paris. 2002

[Stamos & al, 1984]. James W. Stamos. “Static grouping of small objects to enhance performance of a paged virtual memory”. ACM Transactions on Computer Systems, 2(2) :155–180. May 1984

[Sowa, 1984]. J. Sowa. “Conceptual structures: information processing in mind and machine”. Addison Wesley. 1984

[Tarjan, 1972]. R. E. Tarjan. “Depth first search and linear graph algorithms”. SIAM Journal on Computing, 1(2):146-160. June 1972.

[Tufte, 1983]. Edward Tufte. “The Visual Display of Quantitative Information”. Graphics Press. 1983

[Tufte, 1990]. Edward Tufte. “Envisioning Information”. Graphics Press. 1990

[Tufte, 1997]. Edward Tufte. “Visual Explanations”. Graphics Press. 1997

[Turenne, 2000]. Nicolas. Turenne. “Apprentissage statistique pour l’extraction de concepts à partir de textes. Application au filtrage d’information textuelle”. Thèse de Doctorat, spécialité informatique l’université Louis Pasteur, Strasbourg. .2000

[Véronis, 1996]. J. Véronis. “HyperLex: Cartographie lexicale pour la recherche d’informations”. Actes de la Conférence Traitement Automatique des Langues (TALN). Betz-sur-mer, France. ATALA. 1996

[Wagner, 2001] Wagner. “The yeast protein interaction network evolves rapidly and contains few redundant duplicate genes”. Mol Biol Evol., 18(7) : pp1283-1292. 2001

[Watts & al, 1998] D. J. Watts, S. H. Strogatz. “Collective dynamics of "small-world" networks”. Nature, 393 : 440-442. 1998

[Watts, 1999]. Duncan. J. Watts. “Small Worlds”. Princeton University Press. 1999

[Wilson & al, 1991]. Paul R. Wilson, Michael. S. Lam et Thomas. G. Moher. “Effeective “static-graph” reorganization to improve locality in garbage-collected systems”. In Proc. of the SIGPLAN’91 Conf. on Prog. Lang. Design and Implementation, pages 177–191, Toronto (Canada). June 1991

[Yalaoui & al, 2005a]. B. Yalaoui, H. Aït Haddadene et H. Harik. “Le modèle des cartes cognitives dans la théorie de l'argumentation”. 6ème conférence de la Société Française de Recherche Opérationnelle et d’Aide à la Décision, Ecole Polytechnique de l’Université de Tours, France. 2005

[Yalaoui & al, 2005b]. B. Yalaoui, H. Aït Haddadene et H. Harik. “ Cognitive Map Model in Argumentation Theory”. International Conference on Research Trends in Science and Technology, Lebanese American University, Lebanon. 2005

[Zipf, 1949]. G. K. Zipf. “Human behaviour and the principale of least effort: An introduction to Human ecology Reading”, Mass: Adisson Wesley. 1949

Annexes

Annexes A

Travaux de valorisation

« Le modèle des cartes cognitives dans la théorie de l'argumentation »

Communication avec comité de lecture à :

ROADEF'05 :

6ème conférence de la Société Française de Recherche Opérationnelle et d'Aide à la Décision

14,15, 16 Février 2005

Ecole Polytechnique de l'Université de Tours
France

Le modèle des cartes cognitives dans la théorie de l'argumentation

B. Yalaoui[1], H. Aït Haddadene[2] et H. Harik[3]

[1] Centre de recherche sur l'information scientifique et technique
Laboratoire de recherche et développement information scientifique
Rue des trois frères Aissou, Ben-Aknoun, Alger, Algérie
yalaoui@wissal.dz
[2] Université des sciences et de la technologie Houari Boumediene
Bp 32 El Alia, Alger, Algérie,
hacene.ait_haddadene@caramail.com
[3]Université des sciences et de la technologie Houari Boumediene
Bp 32 El Alia, Alger, Algérie
harik_hakim@yahoo.fr

Une représentation graphique est souvent plus simple et plus lucide qu'une approche quantitative qui est généralement inadéquat pour traiter des interrelations causales, cette représentation permettra de donner plus d'initiatives à l'utilisateur et fournira ainsi un outil intéressant pour l'aide à la prise de décision. Parmi ces representations graphiques on a le modèle de cartes cognitives. De même, la théorie de l'argumentation est une conception de la compréhension des arguments, c'est une activité cognitive qui permet d'étudier la complexité et l'incertitude de la connaissance à un problème donné en augmentant (ou en diminuant) l'acceptabilité d'un point de vu controversé pour justifier une prise de décision. Ce qui est intéressant c'est de voir comment le modèle des cartes cognitives peut être utilisé comme outil d'argumentation pour l'aide à la prise de décision. L'objet de ce papier est d'un coté présenter brièvement la théorie de l'argumentation et le modèle des cartes cognitives avec ces applications dans les problèmes décisionnels, et d'un autre coté montrer la relation qui existe entre les deux concepts dans le contexte d'un processus d'aide à la prise de décision.

L'argumentation est le processus de construction d'arguments pour certaines propositions, elle nous aide à la compréhension et l'interprétation des buts normatives et descriptives pour trouver des conclusions d'une manière rationnelle à des attitudes et des croyances de l'individu [3] et permet d'identifier les prétentions appropriées et les conclusions à faire pour l'analyse d'un problème donné. Pour VAN EAMEREN [7], l'argumentation est une activité social du raisonnement qui vise à augmenter (ou diminuer) l'acceptabilité d'un point de vu controversé qui est pris avant un jugement raisonnable pour le justifier. Elle peut être vue comme un sous type du raisonnement puisque son principe est de déclencher l'opération à partir d'une proposition (point de départ) pour se retrouver dans une conclusion (point final) à partir d'un ensemble de raisonnement logique [9].
Pour expliquer comment les gens arguent dans la vie quotidienne, STEPHAN TOULMIN dans son travail « The use of argumentation » [6], développa sa célèbre théorie de l'argumentation pour montrer et expliquer comment l'argumentation se produit dans le processus naturel de tout les jours, il donne une excellente introduction à la notion de l'argumentation où il nous offre un modèle qui nous permet d'analyser et d'évaluer nos arguments. Aujourd'hui, l'argumentation est une approche qui a montré son utilité et a attiré l'intérêt de plusieurs chercheurs et dans divers domaines comme : résonnement légal, système d'agent [2]. Pour plus de détails sur le formalisme d'argumentation, le lecteur peut voir et consulter [4], [8].

Le concepts de carte cognitive est composée de deux termes; La carte est une représentation de l'espace sur un support physique sur lequel s'appuie un individu pour s'orienter, estimer des distances et donc un outil de synthèse et d'analyse alors que le terme cognitive vient de latin cognitus (connu), qui concerne la connaissance puisque qu'elle qualifie les processus mentaux où un organisme acquiert des informations sur son environnement et les traite pour ajuster son comportement ce qui pourrait nous amener à dire qu'une carte cognitive est une modélisation des représentation mentales. Cette notion est attribuée à TOLMAN [5]. Dans ce contexte, une carte cognitive est une représentation mentale de l'environnement auquel nous appartenons suite à une série de transformations psychologique où l'individu acquiert, code, stocke, rappelle, décode les informations sur l'environnement spatiale. En 1976, ROBERT AXELROD [1] décrit les cartes cognitives qu'il a appelé « cartes causales » comme modèle graphique de causalité appliqué à la modélisation des croyances des décideurs politiques dans un contexte décisionnel. De façon générale, une carte cognitive est une représentation graphique de la représentation mentale issue d'un ensemble de représentations discursives énoncées par un sujet à partir de ces propres représentations cognitives à propos d'un objet particulier.

L'homme utilise souvent l'argumentation pour qu'il justifie ses propres décisions, souvent il a besoin des outils pour supporter ses arguments. Dans cette optique, la carte cognitive peut être utile pour formaliser des arguments et donc fournir une forme d'argumentation, cette faisabilité se résulte par le type d'inférence qui est la base de tout système d'aide à la décision et qui peut être traduit à l'utilisateur en terme d'argument. Dans le modèles des cartes cognitives, l'inférence causale est une caractéristique importante qui concerne la détermination qu'elle est l'effet qu'un concept cause donné génère sur un concept effet, ceci pour déduire l'influence d'une partie du graphe sur l'ensemble du graphe suivant la technique de propagation de causalité, elle donne la réponse et le diagnostique à la question de type « qu'arrivera-t-il si » qui veut dire qu'arrivera pour d'autres événements si on déclenches une collection d'événements dans le graphe et donc à partir de ce principe on peut prendre une décision appropriée qui est arguée par le mécanisme d'inférence dans la carte cognitive. Ce processus correspond donc au processus de raisonnement dans l'argumentation. L'intérêt de ce modèle est justifié par le fait qu'il contient des influences qualitatives qui sont propagés à travers le réseau [1]. Il détient une forme de raisonnement avec une signification relative des influences directes, ce qui donne un moyen pour l'utilisateur à argumenter et l'aider à justifier ces décisions.

Ainsi, la carte cognitive est une manière d'aborder la complexité et la dynamique qui caractérise le comportement de tout les jours, par la représentation des assertions causales d'une personne sur un domaine limité et qui peut être un outil intéressant dans le processus de l'argumentation par utilisation du mécanisme d'inférence causale, les futures travaux doivent se focaliser sur l'extension du modèle de base de la carte cognitive en introduisant des nouvelles techniques d'inférence.

Références :

[1]. Axelrod, R: Structure of decision: The Causal Maps of political Elites R. Axelrod, ed. Princeton Univ Press (1976).
[2]. Parsons, S. Sierra, C and Jennings, N: Agents that reason and negotiate by arguing. Journal of Logic and Computation, 8, pp 261-292 (1998).
[3]. Pollock, J: Defeasible reasoning. Cognitive Science, 11. pp 481-518 (1987).
[4]. Prakken, H and Vreeswijk, G: Logical systems for defeasible argumentation. In D Gabbay, editor, Handbook of Philosophical Logic. Kluwer (2000).
[5]. Tolman, E: cognitive maps in rats and men. Pyschogical review, vol. 55 (1948).
[6]. Toulmin, S: The Use of Argument. Cambridge University Press (1958).
[7]. Van Eemeren, Frans H. Rob Grootendorst, Francisca Snoeck: Fundamentals of Argumentation Theory: A Handbook of Historical Backgrounds and Contemporary Developments. Mahwah, NJ: Lawrence Erlbaum (1996).
[8]. Vermeir, D. Laenens, E and Geerts, P: Defeasible logics. In Handbook of Defeasible Reasoning and Uncertainty Management, volume2. Kluwer (1998).
[9]. Walton. Douglas N: What is reasoning? What is Argument?. The Journal of Philosophy pp87, 399-419 (1990).

« Une méthode pour l'analyse et partitionnement du graphe de termes d'un corpus textuel»

Exposé Interne *(CERIST)*

Et

Papier soumis à :

Revue d'Information Scientifique et Technique *(RIST)*

Ed. CERIST, ISSN 1111-0015

Une méthode pour l'analyse et le partitionnement du graphe de termes d'un corpus textuel

B. Yalaoui[1], H. Aït Haddadene[2] et H. Harik[3]

[1] Centre de recherche sur l'information scientifique et technique
Laboratoire de recherche et développement information scientifique
Rue des trois frères Aissou, Ben-Aknoun, Alger, Algérie
yalaoui@wissal.dz

[2] Université des sciences et de la technologie Houari Boumediene
Bp 32 El Alia, Alger, Algérie,
hacene.ait_haddadene@caramail.com

[3]Université des sciences et de la technologie Houari Boumediene
Bp 32 El Alia, Alger, Algérie
harik_hakim@yahoo.fr

Résumé

La masse d'information textuelle existant que ce soit sous forme de documents accessibles sur Internet, dans les bases de données bibliographiques des entreprises et des institutions ne cesse d'augmenter régulièrement. L'usager, qui n'a plus ni le temps ni les ressources cognitives pour faire face à un tel volume d'informations, se trouve dans une situation de saturation. Il lui faut donc pouvoir prendre connaissance du contenu des textes par des moyens rapides et efficaces afin de représenter ce contenu et d'acquérir un ensemble de connaissances utiles, en s'appuyant sur l'extraction de termes et des relations entre ces termes à partir d'un corpus textuel.

Dans cet article, nous nous intéressons plus particulièrement à la fouille de données dans des corpus textuels volumineux par des méthodes issus du text mining ceci par l'utilisation de quelques supports tel que la théorie des graphes pour la modélisation du contenu global d'un corpus d'une thématique, pour une représentation cartographique du contenu d'un corpus textuel. Nous avons proposé trois méthodes pour le partitionnement du graphe d'association de termes afin de rassembler les termes du corpus qui sont homogènes en se basant sur la notion de densité d'un sommet, sur la notion d'importance d'une arête et enfin sur la triangulation du graphe d'association de termes. Nous avons aussi proposé la notion d'ensemble générique représentatif d'un corpus textuel qui permet de donner une vue globale sur le contenu du corpus. Après avoir modéliser le problème sous forme de réseau orienté et pondéré, on a montré que cet ensemble peut être trouver efficacement en un temps polynomial.

Le principe de la cartographie d'un corpus textuel est de permettre à l'utilisateur de prendre en compte les liens qui existent entre les différentes notions représentées. Ce qui permet de synthétiser des informations sous une forme facile à interpréter et à exploiter. Une des caractéristiques de cette représentation est que le support graphique permet d'aider une communauté d'utilisateurs, travaillant sur une thématique donnée, d'appréhender et de visualiser globalement l'ensemble des termes représentatifs de corpus volumineux avant de porter l'attention sur la partie intéressante dans le graphe. Elle peut également être exploitée, dans le cadre d'un système de recherche documentaire pour aider à la recherche d'information et à la navigation dans le contenu d'un corpus textuel.

« Un algorithme pour la recherche d'un ensemble générique de termes dans un réseau de termes associés»

Communication proposée pour :

WEA'06 :

5ème International Workshop on Experimental Algorithme

May 24-27 2006

Menorca Island
Spain

Un algorithme pour la recherche d'un ensemble générique de termes dans un réseau de termes associés

Résumé : Dans cette contribution on s'intéresse à la notion de termes génériques d'un corpus textuel. Après avoir modélisé le problème sous forme de réseau orienté et pondéré, nous montrerons que cet ensemble peut être trouvé efficacement en un temps polynomial. Nous terminerons par illustrer quelques applications possibles pour le domaine text-mining.

Mots-clefs : graphes, réseau de termes associés, ensemble générique.

Introduction

La représentation du contenu d'un corpus [5] permet de montrer les relations entre un ensemble $\mathcal{T}$ de termes issues d'un corpus textuel. Ces termes ne jouent pas le même rôle dans le sens où il y a des termes porteurs de plus d'information que d'autres et des termes plus précis que d'autres. L'un des problèmes qui se pose est de proposer une méthode de réduction de cet ensemble de termes afin de fournir une liste plus courte mais porteuse d'information essentielle [5], [3], i.e. qui peut donner une idée générale et synthétique sur le contenu et les thématiques traitées dans le corpus textuel en question.

Pour montrer le rôle que joue les termes dans un corpus textuel, ZIPF [7] a constaté, en étudiant des corpus de données textuelles, des régularités sur des fréquences d'apparition des mots. Ce qui lui a permis de découper l'ensemble de termes d'un corpus en trois zones:

(2) *Zone d'informations triviales :* ensemble de mots triviaux et principaux qui définissent le sujet en fonction de la thématique de base.

(3) *Zone d'informations intéressantes :* ensemble de mots représentatifs du contenu de base qui permet de construire les structures des relations des différentes approches du thème étudié.

(4) *Zone d'information marginales ou de bruits :* ensemble de mots non pertinents par rapport au thématiques de base.

Dans ce travail, nous proposerons une méthode qui permet de sélectionner un ensemble de termes qui sont centraux et qui donnent une vue globale sur le contenu d'un corpus textuel, et ce à partir d'un réseau de termes associés qu'on a construit.

Présentation du problème

On construit le réseau de termes associés $\mathcal{R}=(\mathcal{T},\mathcal{U},\mathcal{F})$ correspondant à un corpus textuel subdivisé en un ensemble fini d'unités textuelles comme suit :

$\mathcal{T}$: ensemble de termes significatifs du corpus en question

$\mathcal{F}$: la fonction degré d'association définie de $\mathcal{T}x\mathcal{T}$ vers [0,1] telle que $\mathcal{F}(T_i, T_j) = \frac{NO_{i,j}}{NO_i}$ notée $\mathcal{F}_{i,j}$

Où NO_i représente nombre d'unités textuelles [2], [4] dans lesquels le terme $T_i \in \mathcal{T}$ apparaît

$NO_{i,j}$ le nombre d'unités textuelles dans lesquels les deux termes $T_j, T_i \in \mathcal{T}$ apparaissent simultanément.

L'ensemble des arcs est tel que : $\forall T_i \in \mathcal{T}, \forall T_j \in \mathcal{T}$: $(T_i,T_j) \in \mathcal{U} \Leftrightarrow \mathcal{F}_{i,j} > 0$.

Définition

Soit $\mathcal{R}=(\mathcal{T},\mathcal{U},\mathcal{F})$ le réseau de termes associés correspondant à un corpus textuel donné. Un ensemble générique de termes $S \subset \mathcal{T}$ est un ensemble qui vérifie les deux conditions suivantes :

(C1) : $\forall$ t $\in \mathcal{T}$ / S, $\exists$ (t,t')-chemin dans $\mathcal{R}$ avec $t' \in S$

(C2) : *S* est minimal par la condition (C1).

Un ensemble générique de termes S* est dit représentatif de $\mathcal{R}$ si et seulement si $\sum_{t \in S^*} d_R^-(t)$ est maximum.
Dans ce qui suit, on s'intéresse à la recherche de cet ensemble dans un réseau de termes associés.

Les résultats

Soient :

- $C_1, C_2, \ldots, C_p$, la décomposition de $\mathcal{R}$ en p composantes fortement connexes.
- $K=\{i / i \in \{1,..p\}$ et $\Gamma^+(C_i)=\phi\}$, avec $\Gamma^+(C_i)$ ensemble de successeurs des sommets de C_i dans $\mathcal{T}$ / C_i

On montre alors les résultats ci-dessous :

Lemme 1:

Soit *S* un ensemble de sommets de $\mathcal{R}=(\mathcal{T},\mathcal{U},\mathcal{F})$ qui vérifie la condition (C1) alors :

$\forall k \in K, \exists t \in C_k$ tel que $t \in S$.

Propriété:

Si *S* est un ensemble de sommets de $\mathcal{R}=(\mathcal{T},\mathcal{U},\mathcal{F})$ alors :

S est un ensemble de sommets génériques $\Leftrightarrow \forall k \in K : |S \cap C_k| = 1$ et $S \cap \left(V - \bigcup_{C_i \in B} C_i \right) = \phi$.

Corollaire 1:

Un ensemble générique d'un réseau de termes associés est un stable.

Corollaire 2:

Deux ensembles génériques d'un même réseau de termes associés sont de même cardinalité.

Corollaire 3:

L'ensemble de termes génériques représentatif d'un réseau de termes associés $\mathcal{R}=(\mathcal{T},\mathcal{U},\mathcal{F})$ est l'ensemble *S** tel que :

1. $\forall t \in S^*, \exists k \in K : t \in C_k$ et $d_R^-(t) = \max_{v \in C_k} \{d_R^-(v)\}$.
2. $S^* \cap \left(V - \bigcup_{C_i \in B} C_i \right) = \phi$
3. $\forall k \in K : |S^* \cap C_k| = 1$

Ce dernier corollaire nous permet de proposer l'algorithme suivant :

Algorithme de détermination d'un ensemble *S**

Entrée : $\mathcal{R}=(\mathcal{T},\mathcal{U},\mathcal{F})$
Sortie : Un ensemble de sommets génériques représentatif *S**.
(1) Partitionnement de $\mathcal{R}=(\mathcal{T},\mathcal{U},\mathcal{F})$ en composantes fortement connexes, soit : C_1, C_2, …, C_p.
(2) Trouver l'ensemble *K* tel que K={i / i∈{1,..p} et $\Gamma^+(C_i)=\phi$}, avec $\Gamma^+(C_i)$ ensemble de successeurs des sommets de C_i dans $\mathcal{T}$/ C_i
(3)

- Mettre $S^* = \phi$
- Pour chaque $k \in K$, choisir un sommet $v \in C_k$ tel que $d_R^-(v) = \max_{t \in C_k} \{d^-(t)\}$, et mettre $S^* = S^* \cup \{v\}$.

(4) Fin

Applications

La notion d'ensemble générique représentatif d'un réseau de termes associés à un corpus textuel, peut être exploitée dans une perspective d'élaboration d'une cartographie du contenu. En effet, on peut envisager une visualisation du corpus en question à trois niveaux :

- La première consiste à proposer une vue globale du réseau en déterminant l'ensemble de termes générique représentatif du réseau. Cette vue globale est une première information à l'utilisateur et elle lui permet d'élaborer une stratégie d'exploration.
- La deuxième vue peut être obtenue si le réseau est partitionné en communautés [1][7] ; selon un critère donné, on peut trouver les termes génériques représentatifs de chaque communauté
- La troisième vue consiste à naviguer dans le réseau de termes associés

Références

[1] P. Berkhin (2002). Survey of clustering data mining techniques. Rapport technique, Accrue Software, San Jose, CA.

[2] M. H. Haddad (2002). Extraction et impact des connaissances sur les performances des systèmes de recherche d'information. Thèse de doctorat. Spécialité Informatique. Université de Joseph- Fourier, Grenoble 1.

[3] J. Han et M. Kamber (2000). Data Mining : Concepts and Techniques. Morgan Kaufmann Publishers.

[4] P. Poncelet, A. Mokrane, R. Arezki, G. Dray (2004). Cartographie automatique du contenu d'un corpus de documents textuels. In proceeding of the 7 th International Conférence on the Statistical Analysis of Textual Data JADT, vol 2, pp 816-823, Louvain-la-Neuve : Presses Universitaires de Louvain.

[5] N. Turenne (2000). Apprentissage statistique pour l'extraction de concepts à partir de textes. Application au filtrage d'information textuelle. Thèse de Doctorat, spécialité informatique l'université Louis Pasteur, Strasbourg.

[6] B. Yalaoui (2005). On related combinatory problems in information cartography. The 20th british combinatorial conference, Durham from 10 to 15 jully 2005 UK

[7] G. K. Zipf (1949). Human behaviour and the principale of least effort: An introduction to Human ecology Reading. Mass: Adisson Wesley.

Annexes B

Techniques d'extraction de termes à partir d'un corpus textuel

L’extraction de termes à partir de données textuelles est une discipline émergente, construite par l’assemblage de diverses disciplines : « …statistiques, intelligence artificielle, base de données,…», elle peut être définie comme étant la tache qui consiste à identifier de l’information bien précise d’un texte en langue naturelle et à le représenter sous forme structurée dans le but de réduire l’effort intellectuel humain. Elle sert donc à retrouver dans un ou plusieurs documents donnés les termes discriminants et utiles à la compréhension du discours. Elle consiste donc à parcourir les immenses volumes des données textuelles, à la recherche de connaissances.

Jusqu’à tout récemment, deux grandes avenues ont été empruntées pour recenser automatiquement les termes : les modèles linguistiques et les modèles statistiques. De nouvelles recherches entreprises au cours de la dernière décennie tendent à tirer profit de ces deux grandes approches pour proposer des méthodologies qui ne sont ni purement linguistiques, ni purement statistiques (modèles hybrides). Ainsi, nous allons maintenant décrire ces approches d’extraction de termes.

a. Approche statistique

Les méthodes statistiques sont les premières méthodes qui ont été utilisées pour le traitement des informations contenues dans un texte. C’était l’idée avancée dès 1957 par Hans Peter Luhn [Luhn, 1957].

Les méthodes statistiques reposent sur l’idée qu’il existe un rapport entre le contenu véhiculé par un texte et les mots utilisés dans ce texte, que ce rapport est fonction de la fréquence d’usage de mots, et qu’il existe une relation entre la capacité d’un mot à être choisi et sa fréquence d’emploi. Ainsi, l’extraction de termes à partir de critères statistiques part de principe que le sens d’une unité terminologique est étroitement lié avec la distribution de son utilisation dans le contexte en se basant sur hypothèse que l’importance d’un terme pour le traitement du contenu d’une collection de document repose sur le nombre de son apparition dans cette collection et l’emploi de deux termes en cooccurrence est l’expression d’une relation sémantique entre eux. Afin d’exploiter la relation entre les termes, Il existe un grand nombre de méthodes et mesures de similarité à base de calculs statistiques ou probabilistes qui ont été utilisées, il n’est donc pas pensable de les présenter toutes en détail ici, d’autant que certaines sont assez complexes. Les plus connues sont les méthodes de cooccurrence (ou recherche de voisinage) et celles de calcul de la fréquence.

L’avantage de cette approche est qu’elle est facile à mettre en œuvre. De plus, elle est indépendante du corpus. Néanmoins, elle est souvent limitée et pose quelques problèmes :

- L’extraction terminologique basée uniquement sur des critères statistiques se heurte à une difficulté supplémentaire liée aux différentes variantes terminologiques possibles pour exprimer un concept ou une notion puisque un terme peut être présent dans un texte sous différente variante morphologique (masculin, pluriel,…).

- il existe des termes qui sont plus fréquents mais qui ne sont pas intéressants pour la représentation du texte. Cette classe de termes est connue sous le nom de mots vides, se sont des mots ne jouant qu’un rôle syntaxique, apportant peu de sens aux documents et leur suppression ne modifie pas le concept sémantique du texte. Il en existe beaucoup dans les langues (ex : le, la, de, des, alors, lequel...)

b. Approche linguistique

Les méthodes linguistiques se sont développées pour remédier aux insuffisances des méthodes statistiques. A l’inverse de l’approche statistique, l’approche linguistique n’observe pas la régularité des termes dans le corpus. Ce qui est important, ce sont les informations linguistiques exploitées à partir du texte. Autrement dit, cette approche vise à extraire les dépendances ou les relations entre les termes grâce aux phénomènes langagiers. Approche linguistique concerne les combinaisons des éléments textuels au niveau du texte. C’est un niveau qui est très proche de la syntaxe et qui prend en considération les rapports syntagmatiques entre les différentes unités textuelles. Il permet de mettre en évidence les combinaisons linguistiquement correctes donc qui sont susceptibles d’être sémantiquement plus riches, ceci en s’appuyant successivement sur les analyses suivantes :

- L’analyse morphologique : La morphologie concerne l’étude de la formation des mots et leurs variations de forme. Elle vise à ramener tous les mots reconnus dans une phrase à leur forme canonique, en séparant les variations grammaticales (pluriels, conjugaisons) et elle s’appuie sur un dictionnaire qui contiendra tous les éléments de la phrase [Pillet, 2000].

- L’analyse syntaxique : La syntaxe s’intéresse à l’agencement des mots et à leurs relations structurelles dans le texte. Elle s’appuie sur la structure grammaticale de la langue et analyse les phrases ou extraits de phrases pour identifier, lorsque c’est ambigu, le rôle des différents termes : nom, verbe, adverbe, adjectif..., le but de cette analyse est de lever les ambiguïtés en analysant les dépendances syntaxiques entre termes, ainsi dans la phrase : "poses ça sur la table!" le terme Table ne peut être qu’un nom et non une forme conjuguée du verbe Tabler [Mokhtari & al, 2000].

- L’analyse sémantique : C’est l’étape la plus importante car elle met en œuvre tous les principes linguistiques pour analyser les mots et les phrases d’un texte [Mokhtari & al, 2000], afin de déterminer le contenu sémantique (le sens) du texte. Les mots et les structures des phrases identifiées lors des l’analyses morphologique et syntaxique constituent autant d’indice pour le calcul du sens.

Même si cette approche va à un niveau plus élevé de compréhension du texte, et même si l’on obtient donc une meilleure capacité de représentation de contenu du corpus, elle n’est pas facile à mettre en œuvre et il faut beaucoup de connaissances pour résoudre les ambiguïtés de la langue naturelle.

c. Approche hybride

Selon [Fluhr, 1984] l'approche qui semble donner satisfaction est une méthode mixte linguistique et statistique, où un modèle statistique simple s'appuie sur une analyse linguistique plus poussée. Ainsi, les travaux récents se centrent sur la combinaison de l’approche statistique et linguistique, afin de profiter de l’avantage des deux approches, ainsi l’approche hybride est à mi-chemin entre les modèles linguistiques et les modèles statistiques. Cette approche utilise souvent des données statistiques pour filtrer les données linguistiques.

A partir de ces approches, il existe actuellement un grand nombre de logiciels dits de terminologie qui ont été développées afin d’extraire les termes à partir d’un corpus, parmi ces logiciels spécifiques à la terminologie on peut citer, à titre d’exemple :

- Nomino qui compte parmi les systèmes d’acquisition automatique de termes, c’est le doyen des logiciels d’acquisition automatique de termes [Drouin, 2002]. La première version de ce logiciel se nommait Termino [David & al, 1990], il a depuis été remplacé par un nouveau système nommé Nomino [Perron, 1996]. Il est présenté comme un système de dépouillement terminologique [Perron, 1996].
- LEXTER (Logiciel d'EXtraction de Terminologie) qui a été élaboré par Didier Bourigault dans le cadre des approches linguistiques appliquées au repérage automatique des termes [Bourigault, 1992a], [Bourigault, 1992b]. Il reçoit en entrée un corpus de textes portant sur un domaine quelconque, il effectue une analyse grammaticale de ces textes, pour fournir en sortie une liste d'unités terminologiques candidates.
- ANA (Auomatic Natural Acquisition) est un logiciel d’extraction de termes élaboré dans le cadre de travaux sur l’indexation automatique de textes en langue française [Enguehard & al, 1992], il procède au choix des concepts sans recours à l’analyse syntaxique, sémantique ou morphologique.

- XTRACT qui est un logiciel né de la recherche dans le domaine du repérage d'information et de l'indexation automatique, introduit par [Smadja, 1993]. C'est un système purement statistique pour l'extraction de termes complexes (mots composés).
- ACABIT (Automatic Corpus-based Acquisition of BInary Terms) a été développé par Béatrice Daille [Daille, 1994] et qui constitue parmi les premiers pas vers une intégration des statistiques aux techniques linguistiques. C'est un logiciel extraction terminologique basée sur une analyse linguistique et des comptages statistiques [Drouin, 2002].
- FASTER (Filtrage et Acquisition Syntaxique de TERmes) est un logiciel basé sur l'approche linguistique, il a été crée suite au travaux de Jacquemin [Jacquemin, 1997].
- SYNTEX [Bourigault & al, 2000] développer, à partir des principes de base de LEXTER, c'est un nouvel analyseur syntaxique à large couverture pour le français.
- Mantex qui utilise une méthode statistique originale afin d'extraire la terminologie [Frath & al, 2000]. Cette méthode est essentiellement fondée sur le calcul du nombre d'occurrences des segments dans les textes.

Références :

[Bourigault, 1992a]. Didier. Bourigault. "Surface Grammatical Analysis for the Extraction of Terminological Noun Phrases ". Dans Proceedings of the Fourteenth International Conference on Computational Linguistics-COLING 92, Nantes, pp 977-981. 1992

[Bourigault, 1992b]. Didier. Bourigault. "LEXTER, un logiciel d'extraction de Terminologie". Dans Actes de TAMA 92 : 2e symposium international de TermNet, Avignon, mai, pp 229-258. 1992

[Bourigault & al, 2000]. D. Bourigault, C. Fabre. "Approche linguistique pour l'analyse syntaxique de corpus". Cahiers de grammaire, 25, pp.131-151. 2000

[Daille, 1994]. Béatrice. Daille. "Approche mixte pour l'extraction automatique de terminologie : statistiques lexicales et filtres linguistiques". Thèse de Doctorat en Informatique Fondamentale. Université Paris 7. 1994

[David & al, 1990]. Sophie. David, Pierre. Plante. "De la nécessité d'une approche morpho-syntaxique dans l'analyse des textes ". Intelligence artificielle et sciences cognitives au Québec 3(3), pp. 140-154. 1990

[Drouin, 2002]. Patrick. Drouin. "Acquisition automatique des termes : l'utilisation des pivots lexicaux spécialisés". Thèse de Doctorat. Faculté des arts et des sciences, Département de linguistique et de traduction. Université de Montréal. 2002

[Enguehard & al, 1992]. Chantal. Enguehard, Pierre. Malavache, Philippe Trigano. “Indexation de textes : l’apprentissage automatique de concepts”. Dans Actes du XVème colloque international en linguistique informatique, Nantes, pp 1197-1202. 1992

[Fluhr, 1984]. C. Fluhr. “Problèmes d'optimisation de l'accès à l'information dans les bases de données textuelles”. Dans les actes des journées ‘Applications informatiques conversationnelles et le langage naturel', Volume 2. Juin 1984

[Frath & al, 2000]. P. Frath, R. Oueslati et F. Rousselot. “Identification de relations sémantiques par repérage et analyse de cooccurrences de signes linguistiques”. In J. Charelet, M. Zacklad, G. Kassel et D. Bourigault (Eds). Ingénierie des Connaissances: Evolutions récentes et nouveaux défis. Eyrolles. 2000

[Jacquemin, 1997]. Christian. Jacquemin. “Variation terminologique: reconnaissance et acquisition automatique des termes et de leurs variantes en corpus”. Habilitation à diriger des thèses, Nantes, Université de Nantes. 1997

[Luhn, 1957]. Hans. Peter. Luhn. “A statistical approach to mechanized encoding and searching of literary information ”. IBM Journal of Research and Development. Vol 1, n 4. 1957

[Mokhtari & al, 2000]. Amdjed. Mokhtari, Adel. Benaouda. “ Conception et réalisation d’un Moteur d’indexation et de recherche dans un Intranet”. Mémoire de fin d’étude. Institut National de Formation en Informatique, Alger, Algérie. 2000

[Perron, 1996]. Jean. Perron. “ADEPTE-NOMINO : un outil de veille terminologique ”. Dans Terminologies nouvelles, no 15, Bruxelles, RINT, pp 32-47. Juin et décembre 1996

[Pillet, 2000]. Violaine. Pillet. “Méthodologie d’extraction automatique d’information à partir de la littérature scientifique en vue d’alimenter un nouveau système d’information ”. Thèse de doctorat. Discipline : Science de l’Information et de la Communication. Université de droit, d’économie et des sciences d’Aix Marseille III. 2000

[Smadja, 1993]. Frank. Smadja. “Retrieving collocations from text : Xtract”. Computational Linguistics. Vol 19, pp 143-177. 1993

Résumé

Notre travail est consacré à l'utilisation de quelques éléments de la théorie des graphes dans le domaine de la science de l'information et précisément sur la notion de la cartographie d'information. Nous nous sommes intéressés à l'étude de deux approches : le partitionnement d'un graphe pondéré et la notion d'ensemble générique d'un graphe orienté. L'application de ces approches dans le domaine de l'information permet de représenter le contenu d'un corpus textuel sous forme d'une cartographie d'information. Nous avons proposé trois méthodes pour le partitionnement du graphe d'association de termes afin de rassembler les termes du corpus qui sont homogènes en se basant sur la notion de densité d'un sommet, sur la notion d'importance d'une arête et enfin sur la triangulation du graphe d'association de termes. Nous avons aussi proposé la notion d'ensemble générique représentatif d'un corpus textuel qui permet de donner une vue globale sur le contenu du corpus. Après avoir modéliser le problème sous forme de réseau orienté et pondéré, on a montré que cet ensemble peut être trouver efficacement en un temps polynomial. Le principe de la cartographie d'un corpus textuel est de permettre à l'utilisateur de prendre en compte les liens qui existent entre les différentes notions représentées. Elle permet de synthétiser des informations sous une forme facile à interpréter et à exploiter. Elle conduit à une compréhension simple de la structure du corpus étudié et ce, sous forme d'une carte de liens sémantiques entre les termes importants issus de ce corpus.