Résumés de textes par extraction de phrases, algorithmes de graphe et énergie textuelle

Silvia Fernández, Eric SanJuan, Juan-Manuel Torres-Moreno

LIA & IUT STID, Université d'Avignon et des Pays de Vaucluse, France
{silvia.fernandez,eric.sanjuan,juan-manuel.torres}@univ-avignon.fr

RÉSUMÉ. *Lorsqu'il s'agit de résumer automatiquement une large quantité de texte, l'approche actuellement la plus répandue consiste à pondérer les phrases selon leur représentativité. Les calculs sont généralement effectués sur la matrice mots × phrases. Le résumé reprend alors les n phrases les plus lourdes dans l'ordre de leur occurrence. Comme il s'agit de matrices creuses, il est naturel de les représenter par des graphes et cela a conduit à appliquer aux phrases les mêmes algorithmes qu'au Web. Ainsi deux approches LexRank et TextRank ont été dérivées du très populaire algorithme PageRank. Cependant la même matrice peut être interprétée comme un système magnétique, et la pondération comme un calcul d'énergie. Cela conduit à des calculs de pondérations bien plus simples et qui pourtant produisent presque les mêmes classements. Il apparait alors que l'élément déterminant à la production de ces classements sont les chemins d'ordre 2 dans le graphe d'intersection des phrases.*

MOTS-CLÉS : *Résumé automatique par extraction, Page Rank, Algorithmes de graphes, Systèmes magnétiques, Traitement automatique de la langue naturelle écrite*

1. Introduction

Le résumé automatique par extraction de phrases (RAEP) est une des techniques de la fouille de données textuelles [IBE 07]. Il consiste à pondérer les phrases selon leur représentativité dans le texte et à afficher celles de poids le plus fort dans l'ordre de leur apparition dans le texte et dans la limite de la taille du résumé. Généralement les systèmes combinent une large variété de fonctions de pondération (tel que le système CORTEX [TOR 02]) qui assure que le vocabulaire du résumé corresponde bien au contenu informatif du texte initial. Du point de vue de ce seul critère informatif ce type de systèmes ne sont pas de si "mauvais élèves" [FER 08]. Cependant, pour améliorer la lisibilité du résumé produit, et en particulier la résolution des anaphores les plus visibles, des post-traitements s'avèrent nécessaires.

Quoiqu'il en soit, la majorité des méthodes de résumé automatique évaluées lors des conférences DUC/TAC[1] menées par l'agence NIS[2] utilisent une représentation des phrases par sac de mots. Ainsi le texte est assimilé à une matrice S phrases×mots qui code les occurrences des mots dans les phrases sans tenir compte de leur position.

Contrairement au cas des matrices textes×termes utilisées en Recherche d'information (RI) ou l'analyse de données textuelles, le fait de travailler au niveau des phrases et non à celui plus large de textes entiers, fait que les valeurs $S_{i,j}$ sont généralement 0 ou 1. En effet, les mots qui apparaissent plus d'une fois dans une même phrase sont rarement informatifs et correspondent plutôt à des articles, des adjectifs non qualificatifs ou des verbes auxiliaires. Ainsi les matrices S considérées pour le résumé automatique par extraction de phrases sont majoritairement des matrices binaires symétriques creuses qui peuvent être avantageusement représentées par des graphes non orientées.

1. http ://duc.nist.gov/pubs.html et http ://www.nist.gov/tac/
2. http ://www.nist.gov/

2. Représentation du texte sous forme d'un graphe

Les phrases étant ramenées à des ensembles de mots, un texte T de P phrases peut être représenté par un hypergraphe H_T, c'est à dire un famille d'ensembles de mots : $\Phi_T = \{\varphi_1, ..., \varphi_P\}$ où chaque φ_i est l'ensemble de mots $\{w_{i,1}, ..., w_{i,l_i}\}$ de la i° phrase du texte, l_i étant sa longueur. De H_T on dérive le graphe valué G_T d'intersection des phrases. G_T est un triplé (V_T, E_T, L_T) tel que :

1. V_T est l'ensemble Φ_T des phrases du texte T,
2. E_T est l'ensemble de paires $\{\varphi_i, \varphi_j\}$ d'éléments de Φ_T tel que $\varphi_i \cap \varphi_j \neq \emptyset$
3. L_T est une fonction définie sur E_T par $L_T(\varphi_i, \varphi_j) = |\varphi_i \cap \varphi_j|$

G_T correspond simplement à la matrice carrée $S \times S^t$. Outre son adaptation à la représentation informatique de larges matrices creuses, l'intérêt de cette représentation sous forme de graphes est aussi de suggérer l'utilisation d'une large famille de calcul bien connus de centralité[3] de nœuds dans un graphe issus de l'analyse des réseaux sociaux (ARS) (*Social Network Analysis*[4]). Bien sûr, la fonction de valuation des arêtes L_T peut être remplacée par toute mesure de similarité entre deux ensembles cependant la structure du graphe, c'est à dire le couple (V_T, E_T), reste inchangée contrairement au cas général des graphe seuils considérés en analyse de similarité[5]. Nous allons montrer que c'est cette structure joue un rôle fondamental en REAP.

Cependant l'algorithme TEXTRANK[MIH 04] le plus répandu de REAP qui repose sur un parcours du graphe G_T n'a pas été inspiré par l'ARS, mais par l'algorithme PAGERANK[6] [PAG 98] utilisé par le moteur de recherche Google pour calculer l'importance des pages web liées par des hyperliens. De façon intuitive, une page aura un score PAGERANK haut s'il existe plusieurs pages qui la signalent ou s'il y a quelques unes mais avec un score élevé. PAGERANK prend en compte le comportement d'un surfeur aléatoire qui, à partir d'une page choisie au hasard, commence à cliquer sur les liens contenues dans ce site. Éventuellement il peut sortir de ce chemin et recommencer aléatoirement dans une autre page. Vu d'un autre angle, l'algorithme PAGERANK correspond à une variante de la méthode de calcul par puissance successives (*Power Iteration*[7]) pour calculer le premier vecteur propre de la matrice correspondant au graphe des liens entre pages. Le vecteur des scores R correspond ainsi aux composantes du premier vecteur propre de la matrice carrée M de liens entre pages [PAG 98]. L'algorithme PAGERANK a été transposé au traitement de textes par [MIH 04]. L'auteur a assimilé les phrases aux pages web et les liens aux ensembles de termes partagés. Leur système, connue sous le nombre de TEXTRANK, calcule les rangs des phrases dans les documents. Différentes variantes de PAGERANK existent selon la définition de la fonction L_T de valuation et l'initialisation de la méthode. Cependant le principe reste le calcul approximatif du premier vecteur propre d'une matrice creuse dont le graphe correspondant a la même structure que le graphe G_T défini précédemment. Or, si le calcul par puissances successives du premier vecteur propre reste pertinent pour de très larges matrices creuses, dans le cas de textes relativement courts, tels que ceux qui ont été utilisés dans les conférences DUC/TAC, on peut directement procéder au calcul de ce vecteur.

3. L'Énergie textuelle

Si l'on voit maintenant la matrice S comme un système magnétique de spins tel que en [FER 07], alors cette nouvelle analogie conduit à calculer les énergies E d'interaction, ce qui dans le modèle le plus simple correspond au carré de la matrice de G_T : $E = (S \times S^T)^2$. Cette matrice E donne le nombre de chemins de longeur au plus deux entre deux sommets de G_T.

Dans ce cas on prend comme pondération des phrases la vecteur $E.\vec{1}$. Cette nouvelle approche a été appelée ENERTEX. Par rapport à TEXTRANK elle est plus simple puisque elle se limite aux deux premières itérations,

3. http ://en.wikipedia.org/wiki/Centrality
4. http ://en.wikipedia.org/wiki/Social_network_analysis
5. Un graphe seuil (V_s, E_s) découle d'une fonction de similarité f définie sur toute les paires de V_s et d'un seuil s avec $E_s = \{\{u, v\} : f(u, v) > s\}$. Dans ce cas la structure du graphe dépend du seuil s.
6. Marque déposée de la société Google
7. http ://en.wikipedia.org/wiki/Power_iteration

alors que TEXTRANK décrit un processus itératif (de 30 pas approximativement) basé sur le calcul du premier vecteur propre de la matrice de liens entre phrases. Il se trouve que d'après les résultats présentés en [FER 07], ENERTEX atteint les mêmes performances que TEXTRANK du moins en ce qui concerne les mesures ROUGE sur la distribution des mots et des bi-grames couramment utilisées lors des campagnes du NIST. Comme TEXTRANK procède par ailleurs à des post-traitements pour améliorer la lisibilité du résumé, traitements qui peuvent avoir un impact sur ces mesures, une comparaison directe s'avère intéressante. Pour cela nous avons réalisé deux types d'expériences : sur un ensemble de matrices aléatoires en calculant directement le vecteur propre principal de ces matrices et sur un ensemble de textes réels, en calculant les scores TEXTRANK. Dans les deux cas, les rangs obtenus seront comparés à ceux induits par $E.\vec{1}$.

3.1. *Comparaison théorique sur des matrices aléatoires*

Pour comparer les classements issus du vecteur propre principal avec ceux obtenus par l'énergie textuelle, nous avons réalisé l'expérience suivante. En utilisant le logiciel statistique R [8], nous avons défini un ensemble de matrices entières positives M de taille arbitraire P comme le produit matriciel $S \times S^T$ où S est une matrice binaire de D lignes et N colonnes. Nous supposons que pour quelque $0 < i \leq P, 0 < j \leq N$, la probabilité d'avoir $S_{i,j} = 1$ est une constante p. Il est clair que : P est le nombre de phrases à pondérer, N est le nombre de mots différentes et p est la probabilité qu'un terme t se trouve dans une phrase μ. Pour chaque matrice nous avons calculé la matrice d'énergie $E = (S \times S^T)^2$ et le vecteur $\overrightarrow{e} = E.\overrightarrow{1}$, où $\overrightarrow{1} = (1, ..., 1)$ et $\overrightarrow{e}$ est le score donné par l'énergie textuelle. Additionnellement nous avons obtenu le vecteur propre principal $\overrightarrow{r}$ de M. Enfin, nous avons comparé les scores induits pour chaque vecteur en utilisant le test τ de Kendall[9]. Nous avons expérimenté les triplets suivants de valeurs (P, N, p) : (100 ;100 ;0,01), (500 ;100 ;0,01), (500 ;100 ;0,001), (1000 ;100 ;0,01) et (1000 ;100 ;0,001) en répétant le processus 30 fois pour chaque triplet. Nous avons alors calculé la valeur minimale obtenue pour le τ de Kendall. Nous avons obtenu $|\tau| > 0,8$, ce qui induit une p valeur inférieure à 10^{-5}.

Ce résultat indique que pour des matrices entières aléatoires, les rangs basés sur le calcul de l'énergie textuelle ($E = (S \times S^T)^2$), sont fortement corrélés avec les rangs induit par le vecteur propre principal de la matrice $S \times S^T$.

3.2. *Comparaison sur des textes*

Pour mener une comparaison sur des textes réels nous avons implémenté l'algorithme TEXTRANK [MIH 04]. Nous avons utilisé les deux systèmes, TEXTRANK et énergie textuelle, pour classer les phrases d'une vingtaine de documents issus du corpus DUC 2002[10] choisis aléatoirement. Les classements obtenues sont très similaires surtout dans l'assignation des premières places. Un exemple est montré au tableau 1. Il correspond au document sur l'ouragan *Gilbert* utilisé par [MIH 04] pour illustrer le fonctionnement du système. À gauche, les scores normalisés des 25 phrases du texte obtenus pour la méthode d'énergie et TEXTRANK. À droite, les quatre phrases les plus pertinentes qui seraient sélectionnées pour produire, par exemple, un résumé d'environ 100 termes.

4. Discussion

L'explication de ce phénomène semble se trouver dans le degré de connectivité du graphe. Plus ce facteur est grand, plus efficacement sera calculé la relation entre sommets depuis les premières itérations. En fait, un texte mono-thématique est un système où le haut degré de corrélation entre les phrases, produit du partage de termes, donne lieu à un graphe avec un haut degré de connectivité. La connectivité est clairement plus forte entre les

8. http ://www.r-project.org
9. Le coefficient τ de Kendall est proche de 0 dans le cas d'une indépendance totale entre les scores, proche de 1 pour une concordance parfaite et -1 pour des rangs opposés. La p-value donne la probabilité de l'hypothèse nulle d'indépendance statistique.
10. Les conférences DUC 2002 ont concerné aux tâches de résumé automatique monodocument. Le corpus contient ≈ 600 documents généralistes d'une trentaine de phrases chacun.

Rang	Phrase	Score Energie	Phrase	Score TextRank
1	**9**	1,00	**9**	1,00
2	**15**	0,89	**16**	0,90
3	**18**	0,83	**18**	0,86
4	**16**	0,73	**15**	0,74
5	20	0,32	5	0,66
6	14	0,31	14	0,60
7	10	0,31	21	0,56
8	17	0,28	10	0,54
9	5	0,23	12	0,51
10	13	0,21	20	0,46
11	21	0,20	23	0,44
12	11	0,19	13	0,42
13	22	0,18	4	0,39
14	4	0,17	8	0,38
15	23	0,13	17	0,38
16	24	0,12	22	0,38
17	12	0,11	11	0,31
18	8	0,10	24	0,27
19	7	0,03	6	0,08
20	19	0,02	7	0,08
21	3	0,01	19	0,08
22	6	0,00	3	0,00
23	2	0,00	2	0,00
24	1	0,00	1	0,00
25	0	0,00	0	0,00

Les deux systèmes classent en premières places les même quatre phrases :

9 Hurricaine Gilbert Swept towrd the Dominican Republic Sunday, and the Civil Defense alerted its heavily populated south coast to prepare for high winds, heavy rains, and high seas.

15 The National Hurricaine Center in Miami reported its position at 2 a.m. Sunday at latitude 16.1 north, longitude 67.5 west, about 140 miles south of Ponce, Puerto Rico, and 200 miles southeast of Santo Domingo.

16 The National Weather Service in San Juan, Puerto Rico, said Gilbert was moving westard at 15 mph with a "broad area of cloudiness and heavy weather" rotating around the center of the storm.

18 Strong winds associated with the Gilbert brought coastal flooding, strong southeast winds, and up to 12 feet to Puerto Rico ?s south coast.

TAB. 1. *Score des 25 phrases d'un des documents du corpus DUC 2002 obtenus par le calcul de l'énergie textuelle et le système* TEXTRANK. *Les résultats sont similaires, surtout pour les phrases classées en premières places.*

phrases d'un même texte qu'entre les hyperliens entre pages web, ce qui explique les bonnes performances de notre méthode.

5. Bibliographie

[FER 07] FERNÁNDEZ S., SANJUAN E., TORRES-MORENO J. M., Energie textuelle des mémories associatives, ET PHILIPPE MULLER N. H., Ed., *Actes de TALN 2007*, Toulouse, June 2007, ATALA, IRIT, p. 25–34.

[FER 08] FERNÁNDEZ S., VELÁZQUEZ P., MANDIN S., SANJUAN E., MANUEL J. T.-M., Les systèmes de résumé automatique sont-ils vraiment des mauvais élèves ?, *Actes de Journées internationales d'Analyse statistique des Données Textuelles JADT 2008*, 2008.

[IBE 07] IBEKWE-SANJUAN F., *Fouille de textes : méthodes, outils et applications*, Paris , Hermes Science Publications, Lavoisier, Paris, France, 2007.

[MIH 04] MIHALCEA R., Graph-based ranking algorithms for sentence extraction, applied to text summarization, *Proceedings of the ACL 2004 on Interactive poster and demonstration sessions*, Morristown, NJ, USA, 2004, Association for Computational Linguistics, page 20.

[PAG 98] PAGE L., BRIN S., MOTWANI R., WINOGRAD T., The PageRank Citation Ranking : Bringing Order to the Web, rapport, 1998, Stanford Digital Library Technologies Project.

[TOR 02] TORRESMORENO J.-M., VELÁZQUEZMORALES P., MEUNIER J., Condensés de textes par des méthodes numériques, *JADT*, vol. 2, 2002, p. 723–734.