Text mining: résumé de *Dunes*

Pierre Prablanc et Julien Mirval

***Abstract*— Dans ce rapport, nous présentons une méthode de résumé automatique, basée sur des techniques issues du cours de Text Mining, applicable à une série de livres spécifiques. Il s'agit en l'occurence de la série de livres *Dunes*, de Frank Herbert.**

I. INTRODUCTION

Le résumé est une des formes les plus connues utilisée pour condenser l'information littéraire. Elle permet au lecteur d'assimiler une connaissance en un temps bien plus court que s'il lui était demandé de lire l'intégralité du texte résumé. La production littéraire depuis les dernières décennies a littéralement explosé (fig. 1). Ansi, le choix de lectures peut s'avérer difficile étant donnée cette masse littéraire. Le résumé est donc un excellent outil pour lire ou analyser un grand nombre de textes.

Cependant, le résumé peut être également vu comme un procédé de compression avec perte d'information puisqu'il n'est pas possible de reconstruire le texte à partir du résumé. Aussi, peut être existe-t-il d'autres formes pour condenser l'information. On peut citer par exemple le *digest* qui est défini comme un recueil de résumés destiné aux personnes n'ayant pas le temps de lire des livres entiers [1]. Ces formes de résumés citées ci-dessus nécessitent d'avoir lu le texte avant de pouvoir le résumer.

Aujourd'hui, il est désormais possible d'analyser des documents de manière automatique via l'informatique et de manière sophisitiquée grâce aux techniques développées en Text Mining. C'est donc par ces techniques que nous allons explorer une autre manière de condenser l'information textuelle.

II. STRUCTURE DES DONNEES

Puisque notre étude porte sur la série des livres *Dunes*. Nous avons à disposition le 1^{er} livre de la série et nous devons trouver une méthode de résumé qui puisse se généraliser sur les autres tomes de la série.

On cherche à tirer un maximum d'informations du livre pour pouvoir ensuite aggréger ces connaissances. Etant donné que nous travaillons sur un roman, dans lequel différents personnages évoluent et interagissent dans différents lieux au cours du temps, nous devons donc trouver un moyen de synthétiser ces informations.

Une première source d'informations est logiquement la structure du livre, puisque l'histoire est généralement racontée dans l'ordre chronologique. Le livre se compose de la façon suivante:

[1] https://fr.wikipedia.org/wiki/R%C3%A9sum%C3%A9

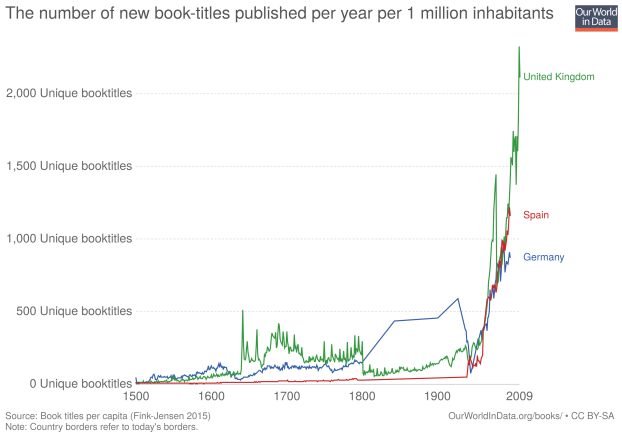

Fig. 1. Le nombre de livre publiés annuellement, par million d'habitants dans un pays donné, sur la période 1500 - 2010. Source: Clio-Infra

- Une page de garde. Elle contient le titre du livre, le nom de l'autre et la date de publication.
- Les différents chapitres. Ces derniers n'ont pas de noms, mais la plupart commencent par une citation d'un livre écrit par un personnage de l'histoire.
- Les appendices. Ils retracent des parties annexes de l'histoire, qui se sont parfois passées avant ou après le fait raconté par le livre.
- Un lexique qui détaille certains des termes et personnages apparaissant dans l'histoire.

On peut également remarquer qu'il existe 2 chapitres au milieu du livre qui servent de séparation entre les grandes parties du livre. Ces chapitres ressemblent à la page de garde et ne contiennent quasi aucune informations.

Dans le livre étudié, les chapitres et appendices sont séparés par une chaine de caractères particulière : "= = = = =". Bien qu'il soit difficile de généraliser en utilisant cette information, on nous a cependant confirmé que ce même caractère de séparation sera utilisé dans tous les ouvrages sur lequel notre méthode sera testé. Il est donc pertinent de l'utiliser pour découper l'oeuvre en chapitres, qui peuvent ensuite former une forme d'unité de temps pour la suite de l'étude.

En revanche, on peut être tenté de ne pas traiter les appendices et le lexique car les informations qu'ils apportent ne s'inscrivent pas dans la chronologie de l'histoire. Cependant, bien qu'ils représentent les 5 derniers chapitres de l'ouvrage étudié, on ne sait pas comment ils se répartissent dans les livres de test.

III. PROBLEMATIQUE

Tout d'abord, il convient de définir les principaux aspects du résumé qui sont son contenu et la forme sous laquelle il est restitué.
Souhaitant résumer un roman, il est nécessaire de considérer les différents éléments de structure de ce type de document. Notamment, les romans présentent généralement une structure temporelle globale, même si il peut exister parfois des narration d'un évènement passé dans le présent (flash-backs). En effet, on peut difficile imaginer fournir un résumé ne respectant pas la structure temporelle. On peut également trouver des liens entre certaines entités (personnages, lieux, organisations, etc. ...). De même, on trouve rarement des objets du roman sans relation les uns avec les autres. On peut penser qu'il est important de conserver ces élements structurant car ils portent une information que nous cherchons à résumer.
Connaissant le contenu du résumé, il existe plusieurs manières de le restituer. La forme classique de restitution d'un résumé se présente sous forme de suites de phrases. Cette forme a l'avantage d'être facilement interprétable localement par phrase (si on sait lire). En revanche, la visualisation est linéaire (à moins de sauter certaines phrases). D'autres types de visualisations commes les graphiques ou les graphes permettent une visualisation plus globale du contenu du résumé sans contrainte de progression linéaire. Cependant, l'interprétabilité peut être moins facile.
Dans notre approche, nous cherchons donc à conserver les éléments de structure cités ci-dessus tout en les représentant sous forme de graphe.

IV. METHODE PROPOSEE

Notre objectif pour résumer le livre est de faire apparaitre les relations entre les personnages et les lieux, ainsi que leurs évolutions au cours du temps. Nous voulons également faire apparaitre dans les grandes lignes les thèmes abordés au fil du livre.

Pour atteindre cet objectif, nous proposons une architecture illustrée en Fig. 2. Le corpus brut est d'abord utilisé pour nourir l'extraction d'entités nommées. Pour l'extraction d'autres éléments, le corpus est segmenté en chapitres qui peuvent être regroupés par ensemble de chapitre (choix de granularité). On calcule ensuite la matrice Terme/Document pour chaque chapitre. Cette matrice utilisée conjointement avec les entités nommées extraites précédemment nous permet de calculer la matrice Entité/Entité. Les topics du corpus sont également extraits mais sur la totalité du corpus prétraité (non segmenté). Les topics et la matrice Entité/Entité sont ensuite utilisés pour la visualisation en graphe.
Le reste du rapport est organisé de la manière suivante. Une description exhaustive de la méthode proposée est donnée en section V. Les résultats de la méthode sont exposés dans partie VI et les améliorations possibles en VII. Enfin nous concluons sur le travail réalisé en partie VIII.

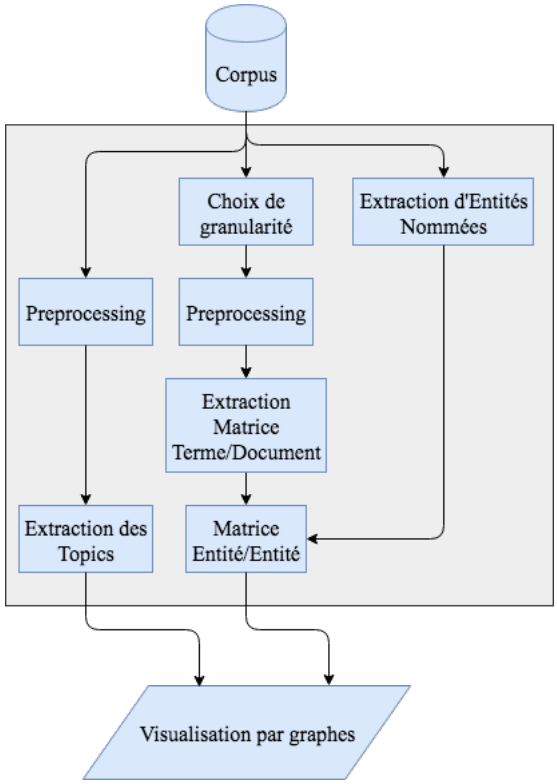

Fig. 2. Vue d'ensemble du système de résumé automatique.

V. SOLUTION DETAILLEE

A. 1^{er} *prétraitement*

Un roman est généralement écrit en séquence: Les premières pages se passent avant les dernières. Il est donc logique de tenter de tirer parti de cette information pour avoir une notion de temps dans notre analyse. Dans cette première partie, nous extrayons donc tout d'abord les informations sur la structure du roman. On découpe donc en chapitres puis chaque chapitre en documents d'une ligne. De la sorte, on dispose alors d'une granularité intéressante pour la suite des traitements.

B. *Détection d'entités nommées*

A partir de ce corpus brut mais découpé, on effectue sans plus de traitement une *détection d'entités nommées*. Cette dernière se basant sur du *Part Of Speech Tagging (PoS)*, elle fonctionne mieux lorsqu'il ne manque pas les mots outils par exemple. On récupère ensuite les *PoS* qui nous intéressent vraiment, c'est-à-dire les personnes, les organisations et les lieux.

C. $2^{ème}$ *prétraitement*

Ensuite, on va traiter le corpus. Le but est d'en enlever les mots outils, la ponctuation et autres, pour ne garder que l'essentiel de chaque document.

Ces opérations ne sont pas sans conséquences, puisqu'on perd par exemple beaucoup d'éléments syntaxiques ainsi que la structure des phrases, mais cela ne nous intéresse que peu dans notre cas.

D. Matrice Termes×Documents

Puis, à partir du corpus traité, pour chaque chapitre, nous extrayons le vocabulaire et les occurences de chaque mots. Grâce à ces informations, nous pouvons ensuite obtenir une matrice Termes×Documents. Cette matrice nous sert de base de mesure pour la proximité entre les entités : deux entités apparaissant dans les mêmes documents vont être jugées proches. Il faut cependant garder en tête qu'ici, les documents vont de quelques mots à plusieurs phrases.

E. Matrice Entité×Entité

En se servant de cette matrice, on en calcule une autre: la matrice Entité×Entité. Pour chaque terme de la matrice, on obtient $0 \leq (a_{ij}) \leq 1$ où le terme vaut 0 quand les deux entités n'apparaissent jamais dans les mêmes documents, et 1 quand elles apparaissent systématiquement ensemble.

Cette matrice est décrite par le terme:

$$(a_{ij}) = \frac{E_i \cdot E_j}{||E_i|| \cdot ||E_j||} \tag{1}$$

avec $(E_i, E_j) \in E$ l'ensemble des entités du chapitre. Cette mesure est fréquemment appellée la similarité cosinus et permet d'obtenir une distance entre deux individus, ici des vecteurs de la matrice Termes×Documents.

F. Topic modelling

Ensuite, on tente de découvrir les thématiques abordées à travers le corpus et de récupérer celles qui sont les plus présentes dans chaque chapitre. Pour cela, on utilise l'algorithme de la *Latent Dirichlet Allocation* (LDA).

La LDA est un algorithme similaire à du clustering : Le but recherché est d'attribuer à chaque document un topic, dépendemment des topics associés aux mots qui le compose. La méthode prend en entrée un hyper paramètre : le nombre de topics souhaités.

Afin d'éviter que les topics ne soient remplis de mots-outils, nous fournissont comme base au modèle le corpus de texte après le $2^{ème}$ traitement. Le modèle LDA est appris sur l'ensemble des chapitres. Nous cherchons ensuite à obtenir une prédiction des topics par chapitre.

Pour déterminer la meilleure valeur pour le nombre de topics, on en essaye plusieurs et on calcule à chaque fois une mesure de cohérence [1]. On retient le modèle qui a obtenu la plus grande mesure de cohérence.

G. Graph des relations

Enfin, on représente par un graphe les différentes entités et les liens qui les unissent. On a un graphe pour chacun des chapitres dans lequel, les noeuds représentent une entité. La longueur des arrêtes entre noeuds est proportionnelle à la similarité entre les entités : Plus elles sont longues, moins les entités sont liées.

De plus, on rajoute un code couleur pour faire état de la dynamique des entités nommées entre les différents chapitres. Ainsi, pour chaque chapitre, on regarde les entités présentes au chapitre suivant et précedent et on distingue 4 cas :

- L'entité n'était présente avant mais le sera après: On la colore en bleu et on la marque "Entrant"
- L'entité était présente avant mais ne le sera plus après: On la colore en rouge et on la marque "Sortant"
- L'entité était présente avant et le sera encore après: On la colore en vert et on la marque "Restant"
- L'entité n'était présente avant et ne le sera pas non plus après: On la colore en jaune et on la marque "Temporaire"

Enfin, on affiche les 10 mots caractéristiques du topic lié au chapitre. Ces mots sont obtenus en fournissant au modèle LDA tout le texte du chapitre. Il nous renvoie une liste de mots classés par probabilités d'apparition dans ce topic.

VI. RESULTATS

Les résultats sont disponibles en intégralité dans les annexes mais dans un souci de clarté, nous avons choisis de n'en faire figurer qu'un seul.

Dans le chapitre 3, 2 personnages (Jessica et Helen Mohia) discutant de l'avenir de Paul, se demandent les raisons qui poussent l'Imperium et la CHOAM Company à venir sur Arrakis. Les choses auraient pu être différentes et cela va les Great Houses du Landsraad. Puis Paul les rejoint et il raconte un rêve prémonitoire qu'il a fait, qui pourrait indiquer qu'il est le Kwisatch Haderach, l'élu. Enfin ils discutent de l'entrainement de Paul, qui doit être fait dans le style de "the Voice".

Dans la Fig. 3, On peut voir comment s'organise tout cela: Jessica, femme d'un duc d'une Great House, mère de Paul et élève Bene Gesserit de Helen Mohia est au centre. Paul n'est pas mêlé au conflit avec l'Impérium, mais un lien est déja créeé avec le Kwisatch et Usul, son futur second nom.

Cependant, dans les annexes, on peut voir de nombreux graphes moins précis. Il y a parfois beaucoup d'entités qui n'ont en réalité que peu de rapport entre elles alors que d'autres liens ne figurent pas, ou sont relayés au second plan.

Enfin, les mots des topics n'offrent que peu d'informations. Par manque de temps, nous n'avons pas pu affiner le procédé autant que souhaité. Les mots ont en effet peu de rapport entre eux et il est difficile, même en connaissant le livre, d'en trouver un.

VII. PERSPECTIVES

Il serait possible d'améliorer la lisibilité en ayant un contrôle plus fin sur la granularité. En effet, lorsque les chapitres sont très denses, le diagramme se retrouve par conséquent peu lisible. En segmentant le chapitre jugé dense, on pourrait obtenir une meilleur visualisation en graphes

Il serait également bénéfique de réussir à faire le liens entre les pronoms personnels et les entités. Ces derniers contribuent à masquer des relations importantes qui pour l'instant nous échappent.

Concernant les topics, il pourrait être intéressant de les incorporer directement dans le graphe en attribuant des entités nommées du graphe à des topics (qui seraient

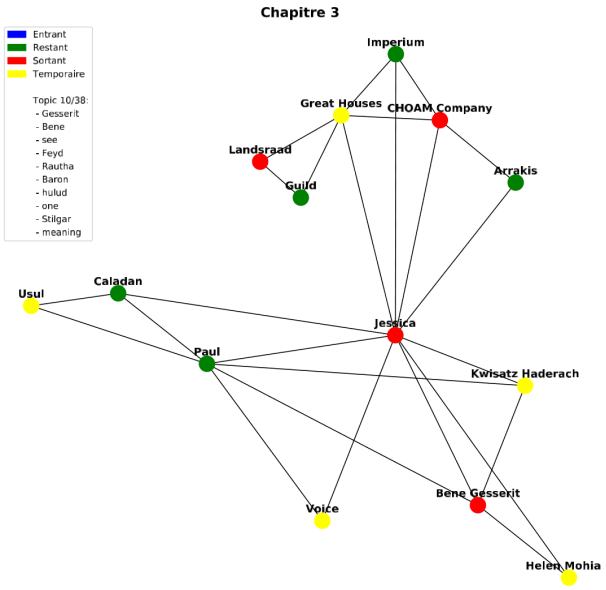

Fig. 3. Graphe de relations entre entités pour le chapitre 3

représentés par exemple avec une forme géométrique pour un topic spécifique). Pour cela, on pourrait récupérer les passages dans lesquels l'entité apparait et grâce au modèle LDA, obtenir le ou les topics prévalent.
Les longueurs d'arêtes du graphe actuel souffrent d'un manque de cohérence d'un chapitre à l'autre. Pour dépasser ce problème, il pourrait être utile de contraindre la valeur des liens entre entitées, et d'introduire une composante en fonction de leurs liens passés.
Les arêtes pourraient également véhiculer l'information de la polarité du lien entre deux entités en utilisant l'analyse de sentiments.

VIII. CONCLUSION

Dans ce projet, nous avons proposé une méthode de résumé centrée sur les graphes d'entités nommées. Les principaux outils utilisés sont donc la reconnaissance d'entités nommées ainsi que le topic modeling. Les résultats obtenus pour certains chapitres montrent que l'approche permet d'obtenir un graphe cohérent. Nous pensons que notre approche pourrait être enrichie en rafinant le choix de la granularité lorsque la lisibilité devient difficile, en ajoutant l'information de polarité entre entités nommées et en incorporant les informations de topics aux graphes.

REFERENCES

[1] Röder, Michael Both, Andreas Hinneburg, Alexander. (2015). Exploring the Space of Topic Coherence Measures. WSDM 2015 - Proceedings of the 8th ACM International Conference on Web Search and Data Mining. 399-408. 10.1145/2684822.2685324.