Trouver de quoi parle un article sans le comprendre

Finding what an article is about without understanding it

Ihab Mallak Henri Prade Mohand Boughanem

Institut de Recherche en Informatique de Toulouse
Université Paul Sabatier
118 route de Narbonne, 31062 Toulouse cedex 09

Résumé :
Le cerveau humain est capable d'identifier les thématiques d'un texte en le parcourant, sans vraiment chercher à comprendre son contenu. C'est ce que cherche à faire la procédure décrite dans cet article. Elle procède en deux temps. Tout d'abord les mots importants et significatifs sont identifiés à partir de regroupements thématiques, et sur la base de plusieurs critères. Ensuite quelques phrases contenant un maximum de ces mots sont extraites du texte, et proposées comme représentatives de son contenu. La procédure qui s'appuie sur des évaluations en termes d'ensembles flous, est illustrée et testée sur une série d'exemples.

Mots-clés :
Contenu sémantique, relation floue.

Abstract:
The human brain is capable of identifying the topics of a text by taking a quick look of it, without really seeking to understand the details of its content. That's what the procedure described in this article is trying to do. It proceeds in two steps. First, words that are important and significant are identified from thematic groupings, and on the basis of several criteria. Then a few sentences containing a maximum of these words are extracted from the text, and taken as representative of its contents. The procedure, based on evaluations in terms of fuzzy sets, is illustrated and tested on examples.

Keywords:
Semantic contents, fuzzy relation.

1 Introduction

La représentation du contenu des documents par mots-clés est à la base des méthodes de recherche d'information. Un document est alors vu comme un sac de mots pondérés [23]. La pondération permet de mesurer le degré d'importance d'un mot du document dans la description du document. Plusieurs techniques ont été développées, les plus répandues et reconnues sont celles basées sur le schéma *tf. idf* (où *tf* est la fréquence du terme dans le document, et *idf* la fréquence inverse du terme dans une collection de référence) [17], [23]. Une telle approche, même si elle reste la plus effective pour retrouver prioritairement le maximum de documents répondant à une requête dans une collection, ne permet d'avoir qu'une vue approximative du contenu des documents. Or, une fois retrouvés des documents présumés pertinents, et avant même d'en prendre entièrement connaissance, il apparaît utile de disposer de vues significatives de leur contenu. C'est ce que d'une façon rudimentaire, les «snippets» de Google permettent en offrant à l'utilisateur de courts extraits de pages où sont mis en évidence les termes de la requête, pour l'aider à ensuite mieux sélectionner les pages susceptibles de l'intéresser.

Dans cet article, on cherche à donner une image du contenu d'un document, indépendante de toute requête, et qui permette d'identifier les thèmes principaux du document, sans recourir à une analyse linguistique détaillée de son contenu (qui permettrait non seulement de déterminer les sujets abordés, mais quels sont les faits rapportés, les thèses discutées, défendues par le document). Ce qui est présenté ci-après s'inscrit en continuation d'un travail récent [2]-[3] qui propose une approche de ce problème conduisant à la construction de groupements stratifiés de termes significatifs. Les termes sélectionnés sont alors choisis non seulement à partir d'informations fréquentielles, mais aussi de leurs niveaux de spécificité et de « centralité » parmi des sous-ensembles de termes du document qu'on peut relier entre eux.

A la différence du travail précédent, on cherche à identifier dans un document les quelques phrases les plus représentatives afin de communiquer une idée de la diversité des sujets abordés. En effet, quelques phrases constituent un matériel plus facilement appréhendable que des strates de termes regroupés thématiquement (comme c'était le cas précédemment). De plus, ces phrases peuvent apporter de l'information sur la façon dont les thèmes sont abordés. La procédure s'appuie dans une première étape sur un affinement de la méthode précédemment proposée pour extraire les termes significatifs d'un document. Dans un second temps, on recherche un petit sous-ensemble de phrases qui contient un maximum de ces termes.

Le contenu de l'article est organisé comme suit. On rappelle tout d'abord les travaux existants en matière de représentation du contenu d'un document, en particulier ceux qui s'appuient sur des méthodes issues de la logique floue. La procédure de sélection des phrases significatives est ensuite présentée et discutée. Un compte-rendu de premières expérimentations montre le caractère prometteur de l'approche.

2 Travaux existants

Extraire automatiquement les caractéristiques importantes d'un texte est un problème clé pour le traitement des tâches liées à la gestion automatique de textes. Le contenu de documents textuels peut être indexé à partir des mots (ou des expressions) présents dans le document, ou des concepts mis en œuvre, ou des thématiques abordées. La plupart des approches sont basées sur les mêmes techniques. Elles consistent à voir un document comme un sac de mots pondérés [23], [12], [15]. La pondération estime le degré d'importance d'un mot à l'égard de la description du document. Plusieurs techniques ont été développées, les plus utilisées sont fondées sur des indices fréquentiels du type *tf. idf* [17], [24], [26].

Diverses approches ont été proposées pour aller au-delà de ces techniques qui voient les documents comme de simples sacs de mots. Elles tentent d'identifier et d'extraire le sens des mots, ou les concepts présents dans le document. Deux grands types d'approches ont été ainsi proposés en recherche d'information (RI). Le premier essaie d'extraire le sens des termes (mots ou expressions), par l'analyse des mots dans leur contexte [16, 25, 29], ou en utilisant une distribution de paires de termes dans des collections de documents, comme cela est fait en indexation sémantique latente (LSI) [8].

Le deuxième type d'approche tente d'identifier les concepts figurant dans les documents. Ces concepts sont souvent obtenus à partir d'une ressource externe (dictionnaire, thesaurus, ontologie). Dans [19], ou dans [27], des graphes sont utilisés pour représenter les documents et les requêtes. Les auteurs proposent une méthode de mesure de la similarité de phrases ainsi représentées. Gonzalo *et col.* [11] proposent une méthode d'indexation fondée sur les 'synsets' de WordNet. Le modèle vectoriel est employé ; les 'synsets' sont utilisés comme espace d'indexation, au lieu des mots lématisés. Dans le même esprit, Richardson *et col.* [22] représentent les documents et les requêtes sur la base de concepts extraits de WordNet. Un des premiers exemples de l'utilisation des ontologies pour l'affinage des requêtes a été proposé dans [28]. Des relations de similarité, ou de généralisation entre termes, sur la base d'un dictionnaire flou de synonymes et d'une ontologie floue, sont utilisées par Olivas *et col*. [20] pour faire un réajustement des pondérations des termes des documents (et générer de nouvelles requêtes). Baziz *et col.* [4] ont proposé un modèle de RI basé sur les concepts, où les documents et les requêtes sont représentés par des sous-arbres de concepts (nœuds de l'ontologie), obtenus à partir d'une ontologie. Les représentations du document et de la requête ne s'appuient pas seulement sur un ensemble de concepts explicitement présents, mais sont complétées dans cette approche par des concepts intermédiaires.

Les travaux de recherche liés à l'extraction de termes dans les documents sont principalement motivés par des tâches de récupération, ou de classification. Par exemple, Boone [4]

décrit une procédure d'apprentissage de traits conceptuels en termes de mots clés, et de distances à des groupes d'exemples de mèls, pour ensuite déterminer les mesures à prendre sur les mèls une fois que leur sujet a été identifié.

L'approche présentée ici vise plutôt à l'extraction de phrases de documents afin de donner à l'utilisateur une idée de leur contenu. Elle se rapproche donc de tâches de résumé [13], mais est basée sur la recherche de mots significatifs.

3 Identification dans un texte des mots et des phrases représentatifs

Comme il vient d'être dit, afin de faciliter l'appréhension du contenu d'un document (sans chercher à le comprendre vraiment), on cherche à identifier des phrases présumées significatives dans ce document et à les retourner à l'utilisateur. L'approche met en œuvre des critères représentés par des ensembles flous [30]. Une méthode dont le principe a déjà été proposé [2] permet l'extraction des termes significatifs d'un document, ce qui donne une meilleure idée du contenu d'un texte que le simple emploi de l'indice *tf* . *idf*, comme de premières expérimentations l'ont montré [3]. Des phrases contenant un maximum de ces termes sont ensuite sélectionnées.

3.1 Relations entre termes

La base de données lexicale WordNet [18], permet de répertorier, classifier et mettre en relation de diverses manières le contenu sémantique et lexical de la langue anglaise. Six relations de base peuvent exister entre deux mots (ou expressions) w et w', et être obtenues à partir de WordNet. Ce sont :

- w et w' sont synonymes (w S w') ;
- w figure dans la définition de w', relation dite de glossaire (w G w') ;
- w spécialise w', c.-à-d. que un w "est-un" w' (w I w'); w' I^{-1} w se lit alors w' généralise w ;
- w est une partie de w' (w P w'); à l'inverse w' P^{-1} w se lit : w' comporte w comme partie ;
- w et w' sont dans le même domaine thématique (w D w') ;
- w est en relation sémantique avec w' (wRw').

Les relations S, D, et R sont symétriques, tandis que G, I et P sont antisymétriques. Par ailleurs, S, I et D sont transitives. De plus, il est possible de définir de nouvelles relations à partir de ces relations en prenant leurs unions: w ($R^i \cup R^j$) w' signifie que w est en relation R^i ou R^j avec w'; on peut aussi penser à composer les relations: w $R^i o R^j$ w' ssi $\exists w°$, w R^i $w°$ et $w°$ R^j w', mais cela conduirait à mettre en relation des mots qui sont sémantiquement distants.

3.2 Grouper les mots qui sont en relation

L'ensemble (fini) des mots/expressions d'une catégorie grammaticale considérée (on ne s'occupe ici que des groupes nominaux), présents dans un document peut être partitionné en clusters, à l'aide d'une relation X, en prenant la fermeture transitive. Dans la suite on prend $X = S \cup I \cup I^{-1} \cup P \cup P^{-1} \cup D \cup R \cup G^*$ (où G* est une restriction de G à des paires dont la «vector measure» [21] est supérieure à un seuil — 0.4 dans l'expérimentation ci-après). Formellement parlant, un cluster, (qui n'est pas un singleton) est tel qu'il existe un chemin entre n'importe quelle paire de mots du cluster, constitué d'une séquence de mots X-reliés: on recherche ainsi les parties connexes du graphe créé par X. L'obtention des clusters est la partie la plus coûteuse de la procédure, mais le calcul de l'ensemble X(w) des mots en relation X avec w peut se faire hors ligne pour chaque mot w.

En cas de mots ou d'expressions polysémiques, les sens multiples présents dans l'ontologie sont discriminés en utilisant les relations, et désambiguïsés (e. g. [1]) afin de choisir le sens le plus approprié dans le contexte.

L'idée de construire des 'clusters conceptuels' a déjà été proposée dans [14] comme une nouvelle méthode d'indexation, mais sans procédure d'extraction de termes. On associe à un cluster sa taille, et sa fréquence globale dans le texte, calculée comme la fréquence cumulée dans le texte des mots du cluster.

3.3 Critères pour sélectionner les termes

A partir de l'ensemble des clusters construits pour un document, nous extrayons un ensemble de sous-ensemble (flou) de mots, qui

restent structurés en clusters, chaque sous-ensemble correspondant à un cluster. Les mots doivent pouvoir être considérés comme représentatifs du contenu du document. Pour opérer cette sélection, nous faisons appel à des critères, dont nous détaillerons la mise en œuvre dans la sous-section suivante.

Ces critères sont la *taille* des clusters, la *fréquence* des mots qui les composent, mais aussi la *spécificité* de ces mots, leur *centralité* dans les clusters, et éventuellement leur *idf* par rapport à une collection de textes de référence. A l'intérieur d'un cluster, chaque mot a en effet un niveau de "centralité", fonction du nombre de mots du cluster avec lesquels il est en relation directe. Comme d'habitude en RI, la "fréquence" d'un mot est évaluée par son nombre d'occurrences dans le texte (et est éventuellement normalisée), les mots dans le même "synset" (mots synonymes au sens de la relation S) étant comptabilisés ensemble.

La spécificité d'un terme est estimée au moyen de sa "profondeur" dans WordNet dans l'arbre conceptuel induit par la relation "est-un". Notons que la mesure de spécificité d'un mot est absolue, puisqu'elle est estimée par sa profondeur dans l'ontologie. Mais, elle n'est pas complètement décorrélée de l'idée d'*idf*, puisqu'un mot spécifique est souvent moins fréquent qu'un mot plus général, tout au moins pour de très grands corpus constitués d'une grande variété de textes. Cependant, ce nombre peut être qualitativement différent d'une mesure *idf* sur des corpus limités de textes spécialisés.

3.4 Sélection progressive des termes

A ce stade, les clusters de mots sont identifiés, et leur "poids" (qui est la fréquence cumulée de leurs mots) calculés, et chaque mot d'un cluster est lui-même associé à trois (ou quatre) éléments d'information: sa fréquence, sa spécificité, sa centralité, (et éventuellement son *idf*). L'idée est de fournir une image du contenu d'un texte sous la forme d'un ensemble de sous-ensembles flous des mots significatifs de chaque cluster ayant un poids *important*.

A chaque itération de la procédure, les clusters qui ont la fréquence cumulée la plus haute (en termes des mots non encore retenus) sont choisis. Pour chaque cluster sélectionné nous cherchons les mots représentatifs sur la base des critères: fréquence (TF), centralité (C), spécificité (S), et rareté dans la collection (IDF). Ces derniers sont représentés par des ensembles flous sur des partitions floues '*grand*', '*moyen*' et '*petit*'. La Fig. 1 donne un exemple d'une telle partition pour la centralité.

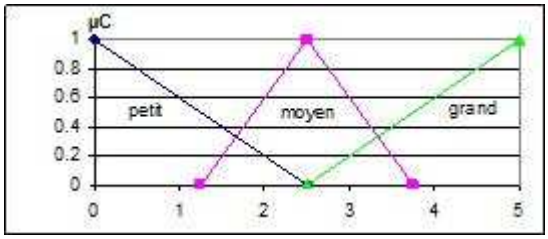

Fig. 1: Partition floue des valeurs de centralité

Les mots représentatifs qui sont choisis dans chaque cluster sont des mots qui satisfont suffisamment au moins deux des critères, par exemple les mots qui sont les plus centraux et les plus spécifiques, ou qui ont un *tf.idf* suffisant. Clairement, déterminer l'exacte balance entre les critères est une question d'expérimentations. Différentes fonctions d'agrégation ont été testées précédemment [3]. Celles qui donnent les meilleurs résultats sont:

- $max(\mu_{\text{TF.IDF-assezgrand}}, min(\mu_{\text{S-grand}}, \mu_{\text{C-grand}}))$,
- $min(\mu_{\text{TF-assezgrand}}, max(\mu_{\text{S-grand}}, \mu_{\text{C-grand}}))$,

Elles peuvent faire émerger d'autres mots que le seul index *tf.idf* [3]. Ici,'assez-grand' a le même support que 'grand', et un noyau élargi égal à la moitié de ce support. La seconde fonction, utilisée dans la suite, donne des résultats comparables, sans avoir besoin d'*idf*.

La procédure de sélection est ensuite itérée sur les mots restants, en ajoutant progressivement de nouvelles couches de sous-ensembles de mots, de moins en moins significatifs, dans les clusters qui sont progressivement sélectionnés. Au début, seuls les mots les plus significatifs des clusters les plus "lourds" sont choisis, puis davantage de clusters sont considérés, et plus de mots dans chaque cluster. Le résultat peut être considéré comme un ensemble d'ensembles flous, où chaque ensemble flou

correspond à l'ensemble des mots qui sont plus ou moins représentatifs (au sens de la fonction de sélection) pour un cluster donné. En effet chaque itération fournit un ensemble (éventuellement vide) qui peut être considéré comme une section de l'ensemble flou des termes représentatifs de chaque cluster. Notons que chaque sous-ensemble flou n'est pas nécessairement normalisé. En effet, le noyau de cette structure est constitué seulement de représentants du/des cluster(s) ayant la fréquence cumulée la plus haute. En fait, nous ne sommes intéressés ici qu'à la structure stratifiée obtenue par itération de la procédure, sans avoir besoin de lui associer des degrés.

3.5 Discussion

On pourrait voir différemment le problème de la sélection des termes représentatifs. Supposons qu'on dispose d'une ressource qui permette d'associer à chaque mot (nom ou verbe) le (ou les) concept(s) qu'il peut servir à exprimer, ses concepts étant organisés dans une ontologie structurée avec une relation de spécialisation/généralisation. Le problème reviendrait alors à chercher un ensemble minimal de concepts les plus spécialisés possibles qui couvrent un maximum de mots (de la catégorie considérée) présents dans le texte. Dans l'approche développée dans cet article, les termes choisis dans les clusters sont, de par leur spécificité, aussi spécialisés que possible, et sont de par leur centralité susceptibles de « couvrir » en un certain sens un maximum d'autres mots (les seuls mots laissés à l'écart étant ceux à la fois peu fréquents et appartenant à de très petits clusters). Comprendre précisément les liens et les différences entre les deux types d'approche est une question à étudier dans le futur.

3.6 Extraction de phrases significatives

A l'étape précédente, nous avons sélectionné l'ensemble des mots du texte qui sont supposés les plus significatifs. Ces mots sont extraits des différents clusters, et leur caractère significatif est stratifié en différents niveaux.

Le but de cette étape est de détecter un ensemble des phrases du texte qui illustre au mieux les sujets abordés. Pour atteindre ce but, nous utilisons les mots sélectionnés à l'étape précédente, les phrases supposées représentatives d'un texte sont choisies à partir de plusieurs critères. Idéalement, un ensemble de phrases sera d'autant plus convenable pour donner une bonne idée du contenu d'un texte que i) un nombre maximum de mots parmi les plus représentatifs y figureront, ii) les mots d'un maximum de clusters y apparaîtront, iii) l'ensemble des phrases sélectionnées restera court (la longueur des phrases étant mesurée par le nombre des mots constituant la phrase).

On pourrait envisager pour cette étape, comme pour la précédente, de représenter en termes de critères flous l'objectif à atteindre. Dans la suite une procédure heuristique plus simple est expérimentée qui sélectionne d'abord la phrase contenant *la proportion de mots significatifs la plus grande*», et qui cherche ensuite des phrases contenant un maximum de mots appartenant au(x) cluster(s) non encore représentés.

4 Expérimentation

Les expérimentations sont menées sur des documents différents portant sur le naufrage de Erika qui a eu lieu en 1999 et ses conséquences. Chaque article traite l'événement d'un point de vue différent. Le but de cette expérimentation est d'extraire à partir de ces documents des phrases qui les représentent au mieux, et de comparer ces phrases à des thèmes extraits à la main par des utilisateurs.

Doc n°	Nb de termes	Résultat fct min($\mu_{TF\text{-}largeenough}$, max($\mu_{s\text{-}large}$,$\mu_{c\text{-}large}$))			Liens vers les documents
		N° d'itérations	Nb de clusters sélectionnés	Termes extraits du texte	
1	446	6	18	1:{C6:{trial} C11:{company} C15:{tanker} C22:{sea} } 2:{C6:{plaintiff} C15:{ship} C56:{total} } 3:{C6:{court} C9:{million} C15:{oil} C19:{accident} C30:{disaster} C31:{damage} C37:{pollution} } 4:{C4:{responsibility, impunity} C6:{appeal, case} C7:{owner} C9:{ten_thousand, hundred} C10:{world} C11:{industry} C13:{lawyer} C15:{structure} C21:{december, february} C22:{coast} C34:{warning} C48:{dollar} C55:{eleven, manager} } 5:{C6:{defendant} C11:{organisation} C15:{fuel}	http://english.aljazeera.net/news/europe/2008/01/200852512481572296.html

				C22:{france, bay} } 6:{C11:{party} C22:{shoreline, state} }	
7	342	5	3	1:{C13:{matter, compound} } 2:{C13:{water, fuel, sediment} } 3:{C13:{coast, sulfur, oil} } 4:{C13:{shoreline} } 5:{C21:{contamination} C23:{mollusc} }	http://cybertheses.francophonie.org/index.php/record/view/22342
10	106	2	6	1:{C1:{oil, tanker} C2:{france, atlantic} C4:{amendment} C8:{tonne} C10:{april, december} C13:{disaster} } 2:{C2:{brittany, sea-coast} }	http://www.imo.org/Environment/mainframe.asp?topic_id=231
15	518	10	7	1:{C3:{ship} } 2:{C3:{trial, case} } 3:{C3:{charge, damage, tanker} } 4:{C3:{government, company, oil} } 5:{C3:{proceeding, tribunal, faith, individual, contract, article, witness, official, council, friend, certification, crew_member, lawyer, critic, law} } 6:{C0:{france} } 7:{C0:{sea} } 8:{C0:{euro} C3:{group} } 9:{C6:{feb} C14:{plaintiff} C21:{pollution} C61:{spill} C62:{total} } 10:{C0:{bay} C3:{organisation} C6:{june} }	http://www.alertnet.org/thenews/newsdesk/L11816302.html
17	175	3	12	1:{C0:{breton} C6:{oil} C25:{disaster} C29:{spill} } 2:{C3:{business, tourism} C5:{responsibility} C6:{tanker} C7:{precedent} C12:{pollution} C13:{seabird, bird} C19:{ton} C22:{mile} C23:{february, december} } 3:{C6:{fuel} }	http://celticcountries.com/webmagazine/environment/erika-oil-spill-brittany-legal-precedent-maritime-pollution/
20	295	5	15	1:{C0:{sea} C3:{prosecution} C5:{oil, tanker} } 2:{C0:{euro} C3:{case, prosecutor} C5:{ship} } 3:{C0:{france} C3:{trial} C7:{week, year} C10:{company} C13:{plaintiff} C15:{monday} C17:{damage, charge} C18:{pollution} C22:{sea_bird, seabird} C24:{sinking} C25:{beginning} C27:{subsidiary} C29:{check} C35:{disaster} } 4:{C0:{bay} C3:{verdict, court} C5:{fuel} C25:{individual} } 5:{C0:{paris, brittany, coast} }	http://www.reuters.com/article/environmentNews/idUSL0427183620070604

Tableau 1

Le Tableau 1 représente les résultats pour 6 des documents utilisés dans l'expérimentation. La colonne « **Doc n°**» donne le numéro du document, la colonne « **Nb of terms**» le nombre des termes réellement présents dans le document. La colonne « **Résultat fct** » représente les résultats de la fonction d'agrégation $min(\mu_{\text{TF-assezgrand}}, max(\mu_{\text{S-grand}}, \mu_{\text{C-grand}}))$, choisie pour l'expérimentation. Les itérations sont notées par leur numéro, les nouveaux représentants des clusters (non singletons) introduits à chaque itération sont notés {C1: {w_1, …, w_i}, …, Ck:{w_1,…, w_j}}. La colonne « **Liens vers les documents**» représente les liens vers les documents sur le web.

Doc n°	**Lignes extraits du texte**	**Nb de termes dans les lignes extraits**	**Nb de termes dans le document**	**Thèmes générale du document extrait manuellement**
1	* Total guilty over Erika oil spill. ** French oil giant Total has been ordered to pay millions of dollars in damages after being found responsible for the 1999 sinking of the tanker Erika, one of France's worst environmental disasters. ** Eleven others, including the ship's captain, were found not guilty. *** 'Severe warning' - The defendants could face hundreds of millions of dollars in further damages after the court said environmental organisations could sue them over the ecological impact of the disaster. * Toxic fuel ** The case finally came to trial in February 2007. - 'Murky world'	88	446	* French oil giant Total responsible for the 1999 sinking of the tanker Erika, ** French court had held the charterer of a tanker responsible for pollution caused through shipwreck. *** Plaintiffs accused the company of negligence in hiring the ship and of acting too slowly when the accident happened.
7	* A qualitative and quantitative assessment was conducted of this contamination in water, suspended particulate matter, sediments, and in intertidal molluscs. * (4) consistent and visible temporal decline in concentrations for water, SPM and molluscs,	33	342	*An investigation was carried out into the PAH chemical contamination resulting from the "Erika" tanker fuel spillage ** The results of this study demonstrated that heavily oil-contaminated shorelines… *** The increase in the contamination levels before and after the spill, together with the significant change in the pattern of PAH composition
10	*The Erika disaster and revised single-hull phaseout schedule ** On 12 December 1999 the 37,238-dwt tanker ** Erika broke in two in heavy seas off the coast of Brittany, France, while carrying approximately 30,000 tonnes of heavy fuel oil	37	106	* The Erika disaster and revised single-hull phaseout schedule ** circumstances of the Erica accident
15	*Total on trial over 1999 French oil disaster * MONSTER TRIAL * The French government alone is	85	518	* Opening of the trial on the Erica disaster - Erica accident and effect on the environment

	seeking 153 million euros. * Besides Total and two of its subsidiaries, the ship's Indian captain, its management company, four French maritime officials and the Italian maritime certification company RINA, which classified the ship as safe, are also on trial. ** Critics, including the environmental group Friends of the Earth, which is one of the plaintiffs in the trial, say Total took cynical risks with the ship to meet a tight contract deadline.			- Political are among some 70 plaintiffs **Total accused of marine pollution and endangering human lives
17	* Erika oil spill in Brittany could set legal precedent for responsibility in maritime pollution ** Brittay's worst-ever environmental disaster	18	175	* Erika oil spill in Brittany could set legal precedent for responsibility in maritime pollution **Brittay's worst-ever environmental disaster - Fishing and tourism business were severely damaged
20	* Prosecutor wants Total convicted for Erika disaster ** PARIS (Reuters) - French oil giant Total should be convicted of maritime pollution for its role in the sinking of the oil tanker Erika, which provoked one of France's worst environmental disasters, prosecutors said on Monday. - The company denies the charges	48	295	* Prosecutor wants Total convicted for Erika disaster - Erica history and effects seabirds ** Total failed to conduct proper checks before chartering the ageing ship. ** Total had faced pollution and negligence charges as well as complicity in endangering human lives over the incident. * Prosecution convict six other individuals and organizations

Tableau 2

La sélection des phrases supposées représentatives d'un texte est basée sur la proportion de mots significatifs présents dans la phrase. A titre indicatif, toutes les phrases (ou sous-titres) dont la proportion de ces termes significatifs est supérieure à 1/4 sont données.

Le Tableau 2 permet de comparer pour un texte les lignes extraites automatiquement par rapport à des thèmes qui sont extraits manuellement (phrases du texte ou des idées reformulées manuellement). La première colonne donne le numéro du document. La colonne « **Lignes extraits du texte**» donne les phrases (ou sous-titres) qui sont extraits automatiquement, tandis que la dernière colonne présentent les thèmes (qui peuvent être des phrases) extraits manuellement du même document. Pour marquer la compatibilité entre thèmes automatiques et thèmes manuelles, nous avons marqué les thèmes identiques par le même nombre d'étoiles dans le tableau. La colonne «**Nb de termes dans les lignes extraits** » compte le nombre des termes dans les lignes extraites automatiquement. La colonne « **Nb de termes dans le document**» donne le nombre réel des termes dans un document.

Les résultats obtenus dans le Tableau 2 montrent une grande compatibilité entre les thèmes extraits manuellement et les phrases du document qui sont extraites automatiquement. Malgré le fait que tous les termes significatifs extraits par la méthode de base ne soient pas présents dans les phrases extraites, les résultats obtenus sont de bonne qualité (83,3% des thèmes manuellement extraits ont été identifiés) et donnent une bonne idée du texte en général. De plus des phrases complètes extraites d'un document sont beaucoup plus faciles à lire et à comprendre pour un utilisateur que des clusters de termes isolés.

Il est à noter que les (sous-)titres sont ici retrouvés par la méthode, ce qui montre son efficacité, puisque en général un titre reflète le contenu d'un document ou d'une section.

5 Conclusion

L'esprit humain est très habile à repérer d'un coup d'oeil les mots significatifs dans un texte afin d'identifier ce dont un texte parle sans vraiment avoir à le lire. Nous avons proposé une sorte d'"algorithme flou" [30] qui tente de faire cela, afin d'avoir une vue du contenu d'un texte qui est sémantiquement plus riche que celle que fournirait une approche purement statistique du type *tf.idf*. Pour ce faire, nous avons exploité les différentes relations sémantiques entre les mots que peut fournir une ontologie telle que WordNet [17]. L'utilisation de structures hiérarchiques de type "petits mondes" [10] construites, e.g. à partir de relations de glossaire entre les mots, pourrait aussi être exploitée.

Références

[1] M. Baziz, M. Boughanem, G. Pasi, H. Prade, A fuzzy logic approach to information retrieval using an ontology-based representation of documents. In: *Fuzzy Logic and the Semantic Web* (E. Sanchez, ed.), Elsevier, 363-377, 2006.

[2] M. Boughanem, H. Prade, O. Bouidghaghen. Extracting topics in texts: Towards a fuzzy logic approach. *Proc. Inter. Conf. on Inform. Processing and Manag. of Uncert. in Knowledge-based Systems (IPMU 2008)*, *Málaga, 22-27/06/2008*, (L. Magdelena, M. Ojeda-Aciego, J. L. Verdegay, eds.), 1733-1740, 2008.

[3] O. Bouidghaghen, M. Boughanem, H. Prade. Extraire les thématiques des textes: Vers une approche par la logique floue. *Actes Rencontres Francophones sur la Logique Floue et ses Applications (LFA 2008)*, Lens, *16-17/10/2008*, Cépaduès Editions, 34-41, 2008.

[4] G. Boone, Concept features in Re:agent, an intelligent email agent. *Proc. Conf. on Autonomous Agents*, Minneapolos/St Paul, May 10-13, 1998.

[5] D. A. Buell, D. H. Kraft, A model for a weighted retrieval system. *J. of American Society for Information Science*, 32, 211-216, 1981.

[6] D. Choi, Integration of document index with perception index, *Soft Computing*, 6, 300-307, 2002.

[7] F. Crestani, Exploiting the similarity of non-matching terms at retrieval time, *Information Retrieval*, 2, 25-45, 2000.

[8] S. Deerwester, S. T. Dumais, T. K. Landauer, G. W. Furnas, R. A. Harshman, Indexing by latent semantic analysis, *J. of the Society for Information Science*, 41(6), 391-407, 1990.

[9] P. J. Garcés, J.A. Olivas, F.P. Romero, Concept-matching IR systems versus word-matching information retrieval systems: Considering fuzzy interrelations for indexing Web pages. *J. of the Amer. Soc. for Information Science and Technology*, 57, 564-576, 2006.

[10] B. Gaume, Mapping the forms of meaning in small worlds, *Int. J. of Intellig. Syst.,* 23, 848-862, 2008.

[11] J. Gonzalo, F. Verdejo, I. Chugur, J. Cigarrán, Indexing with WordNet synsets can improve text retrieval, *Proc. COLING/ACL'98 Workshop on Usage of WordNet for Natural Language Processing*, 1998. http://vldbarc.org/dblp/db/journals/corr/corr9808.html#cmp-lg-9808002.

[12] T. Hisamitsu and J.I. Tsujii, Measuring Term Representativeness, In *Extraction in the Web Era*, (M. T. Pazienza, ed.), Springer, LNAI 2700, 45-76, 2003.

[13] E. Hovy, Text summarization. Chap. 32 in *The Oxford Handbook of Computational Linguistics*, (R. Mitkov, ed.), Oxford Univ. Pr., 583-598, 2005.

[14] B.-Y. Kang, D.-W. Kim, S.-J. Lee, Exploiting concept clusters for content-based information retrieval, *Inform. Sciences*, 170, 443-462, 2005.

[15] D. H. Kraft, G. Bordogna, G. Pasi, Fuzzy set techniques in information retrieval, In *Fuzzy Sets in Approximate Reasoning and Information Systems,* (J. C. Bezdek et al., eds.), Kluwer, 469-510, 1999.

[16] R. Krovetz, B. Croft, Lexical ambiguity and information retrieval, *ACM Trans. on Information Systems*, 10(2), 115-141, 1992.

[17] H. P. Luhn, A Statistical Approach to Mechanized Encoding and Searching of Literary Information, *IBM J. Research and Development,* 1 (4), 309-317, 1957.

[18] G. Miller, R. Beckwith, C. Fellbaum, D. Gross, K. Miller, Introduction to WordNet: An on-line lexical database. *J. of Lexicography*, 3**,** 235-244, 1990.

[19] M. Montes-y-Gómez, A. López-López, A. Gelbukh. Information retrieval with conceptual graph matching. *Proc. DEXA'00*, Greenwich, 2000. Springer, LNCS 1873, 312–321, 2000.

[20] J. A. Olivas, P. J. Garcés, F. P. Romero, An application of the FIS-CRM model to the FISS metasearcher: Using fuzzy synonymy and fuzzy generality for representing concepts in documents. *Int. J. of Approximate Reasoning*, 34, 201-219, 2003.

[21] T. Pedersen, S. Patwardhan, J. Michelizzi, WordNet: Similarity-measuring the relatedness of concepts. *Proc. of NAACL*, 2004.

[22] R. Richardson, A. Smeaton, J. Murphy. Using WordNet as a knowledge base for measuring semantic similarity between words. *Proc. of AICS Conf.*, Dublin, 1994. http://citeseer.ist.psu.edu/187048.html.

[23] G. Salton, M. McGill, *Introduction to Modern Information Retrieval*, McGraw-Hill, 1987.

[24] G. Salton, C.S. Yang, On the specification of term values in automatic indexing, *J. Documentation,* 351-372, 1973.

[25] M. Sanderson. Word sense disambiguation and information retrieval. *Proc. of ACM SIGIR'94 Conf.*, 17, 142-151, 1994.

[26] K. Sparck-Jones, Index term weighting, *Information storage and retrieval,* 9, 619-633, 1973.

[27] R. Thomopoulos, P. Buch, O. Haemmerlé. Representation of weakly structured imprecise data for fuzzy requiting. *Fuzzy Sets & Sys.*, 140, 111-128, 2003.

[28] D. Widyantoro, J. Yen, A fuzzy ontology-based abstract search engine and its user studies. *Proc. 10th IEEE Inter. Conf. on Fuzzy Systems*, Melbourne, Australia, 1291-1294, 2001.

[29] D. Yarowsky, Unsupervised word sense disambiguation rivaling supervised methods. *Proc. 33rd Annual Meeting Assoc. for Computational Linguistics*, Cambridge, Ma, 189-196, 1995.

[30] L. A. Zadeh, Fuzzy sets, *Information and Control*, 8, 338-353, 1965.

[31] L. A. Zadeh, Fuzzy algorithms. *Information and Control*, 12, 94-102, 1968.