VEILLE ET EXTRACTION DE CONCEPTS: ETUDE DE CAS DANS LE DOMAINE DES TELECOMMUNICATIONS

Hamid MACHHOUR(*), Khalid EL HIMDI(**), Ilham BERRADA(*), Ismail KASSOU(*)
machhourhamid@yahoo.fr ; elhimdi@menara.ma ; iberrada@ensias.ma, kassou@ensias.ma

(*) ENSIAS, Université Mohammed V Souissi, BP 713, Agdal, Rabat, Maroc
(**) Faculté des Sciences, Université Mohammed V Agdal, BP 1014 RP, Rabat, Maroc

Mots clefs :
Catégorisation, classification, collecte d'informations, gestion stratégique de l'information, innovation, système de veille, text mining, veille scientifique et technologique.
Keywords:
Categorization, clustering, information gathering, strategic information management, innovation, watching system, text mining, Scientific and technical observation.
Palabras clave :
Clasificación, reunir de información, estratégica Gerencia De Información, innovación, sistema de vigilancia, text mining, Escudriñar científico y tecnológico..

Résumé

Dans cet article nous présentons une application du processus de veille stratégique appliqué au secteur des télécommunications au Maroc. Le but de cette application est de veiller sur les sites web des opérateurs concurrents afin de détecter leurs mouvements. Dans la première partie nous introduisons la problématique de veille en relation avec l'extraction de connaissance et le text mining en général. Puis, dans la deuxième partie, nous présentons un modèle conceptuel pour l'extraction des concepts ainsi qu'un algorithme choisi pour l'enrichissement d'un dictionnaire de concepts spécifique au domaine étudié. Ensuite, dans la troisième partie, nous présentons le processus de veille appliqué à des documents collectés à partir de sites web et fortement liés au secteur des télécommunications au Maroc. Puis nous nous présentons le processus de catégorisation de nouveaux documents, et nous montrons qu'on améliore de plus en plus nos modèles de segmentation et de classement lorsqu'on utilise respectivement la représentation par valeur booléenne, fréquence absolue et par le poids des concepts.

1 Introduction

La veille est «un processus par lequel l'entreprise se met à l'écoute anticipative de son environnement socio-économique dans le but créatif d'ouvrir des fenêtres d'opportunités et de réduire les risques liés à l'incertitude» [9].

En fait, bien que la veille soit une activité volontaire, née dans le monde de l'entreprise, et dont l'objectif est d'aider les décideurs à prendre des décisions stratégiques, on peut distinguer entre une veille manuelle appliquée à des sources d'informations hors ligne (journaux papier, sources bibliographiques, événements professionnels, salons, colloques, fédérations, clubs,…) et une veille automatique appliquée à des sources de données en ligne par le biais de logiciel de veille ou agent intelligent. L'information publique contenue sur Internet représente environ 20 % de l'information disponible [2] et sa plus grande partie (85%) est sous la forme de texte généralement non structuré [13]. De ce fait, l'information recherchée par un processus de veille [2] est le résultat de l'analyse des données collectées et de celles déjà stockées dans des systèmes hétérogènes dans l'entreprise. L'analyse de ces données nécessite des techniques de text mining et d'extraction de connaissances.

De nos jours, le secteur de télécommunication au Maroc a connu un grand essor et plusieurs changements grâce à la concurrence acharnée entre les principaux opérateurs qui sont Maroc Telecom, Meditel et Wana. Les nouveaux clients d'un opérateur sont issus des autres opérateurs qui sont ainsi soumis à une forte concurrence. Les offres proposées par chaque concurrent suscitent toujours des réactions de la part des autres. Pourtant, chaque opérateur doit s'ouvrir vers l'extérieur tout en gardant la maîtrise de son environnement. Sa réussite dépendra en grande partie de sa capacité à gérer la collecte, le traitement et la diffusion de l'information à des fins stratégiques. C'est dans ce contexte que s'inscrit ce papier. Il s'agit de l'identification de la dynamique du secteur de télécommunication au Maroc à travers l'analyse des mouvements des opérateurs concurrents, des fournisseurs, des clients, des produits et des technologies. Cette application se base sur le processus de veille appliqué à des documents collectés du web. Ces documents sont fortement liés au secteur de télécommunication au Maroc.

2 Extraction de concepts

2.1 Prétraitement de texte

Pour faciliter l'analyse de grand corpus de documents, le prétraitement de texte consiste à préparer les documents dans une structure de données qui est plus appropriée aux techniques de data mining et d'extraction des connaissances. Pour ce faire, il existe plusieurs méthodes qui permettent d'exploiter aussi la structure syntaxique et sémantique du texte [3], la plupart des approches du text mining sont basées sur l'idée qu'un document peut être représenté par un ensemble de mots [8], [5], [1], c-à-d, un document est décrit par l'ensemble de mots qu'il détient. L'une des approches principales basées sur cette idée est le modèle d'espace vectoriel (VSM : Vecteur Space Model) [10].

Par sa simple structuration des documents textes (Figure 1), le modèle VSM est le plus utilisé comme approches de text mining. Il représente un document par un vecteur dans l'espace m-dimensionnel, tel que pour un document d du corpus, d est décrit par $V_d = (x(d,t_1),...,x(d,t_m))$, où $x(d,t_i)$ est la représentation du concept t_i dans le document d, m étant le nombre de concepts extraits du corpus. Les documents peuvent donc être comparés à l'aide d'opérations vectorielles simples (calcul de similarité et de distance).

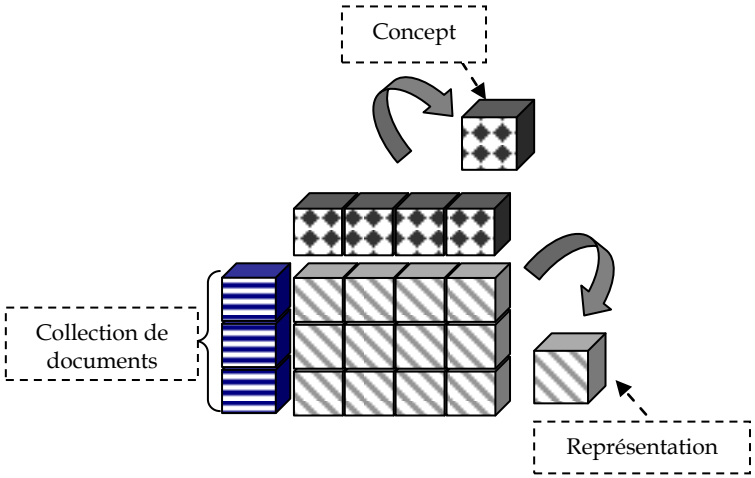

Figure 1: La représentation vectorielle des documents

Les composantes du vecteur V_d peuvent être représentées de façon simple dans l'espace B^m (avec $B = \{0,1\}$ l'espace binaire), tel que x(d, t_i) =1 si le concept t_i est présent dans le document d et x(d,t_i) =0 dans le cas contraire. Cette représentation donne une importance similaire pour tous les concepts du vecteur. Pour améliorer la performance de comparaison et de recherche de documents, et pour donner de l'importance aux concepts les uns par rapport aux autres dans un document spécifique du corpus considéré, la fréquence absolue f(d, t_i) ou encore le « poids » d'un concept (concept weighting) sont utilisés [12]. Le poids w(.,t) d'un concept t est défini par :

$$w(d,t) = \frac{f(d,t)\log(N/n_t)}{\sqrt{\sum_{j=1}^{m} f(d,t_j)^2 (\log(N/n_{tj}))^2}} \qquad (1)$$

où N est le nombre de documents, n_t représente le nombre de documents contenants le concept t et f(d,t) la fréquence du concept t dans le document d.

Il ressort de l'équation (1) que les poids élevés sont assignés aux concepts les plus fréquemment utilisés dans les documents mais rarement pertinents dans la collection entière. Reste à décrire la méthodologie d'extraction des concepts dans ce qui suit.

2.2 Modèle conceptuel d'un concept

L'extraction de concepts se base généralement sur le traitement de la langue naturelle (NLP : Natural Language Processing). La reconnaissance de la langue constitue la première étape du processus NLP. Sur la base d'analyses lexicales et statistiques, la langue du document est identifiée et pour obtenir tous les mots qui sont utilisés dans un texte donné, un processus d'extraction des formes graphiques 1 de base est exigé, on parle de la tokenisation (phase A de la figure 2). Le vocabulaire de mots extraits est soumis à un filtrage qui permet d'enlever les mots portant moins d'information, comme les conjonctions, les prépositions, etc., ou

[1] Forme graphique : Suite de caractères non délimiteurs entourée par des caractères délimiteurs.

ceux qui, en se basant sur la fréquence absolue des mots, se produisent souvent dans tous les documents, car il portent moins d'information pour distinguer entre les documents. Par filtrage, 40 à 50 % de mots sont supprimés [11]. La méthode standard du filtrage est le filtrage par liste de mot d'arrêt « Stopword list »

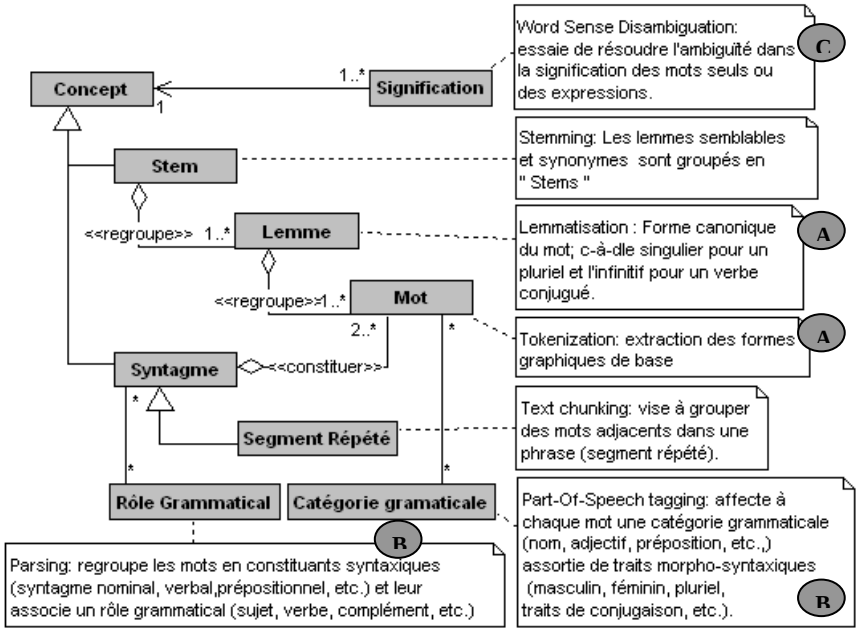

Figure 2: Modèle conceptuel d'un concept

La seconde étape du processus NLP est l'analyse syntaxique (phase B) ; pour chaque mot une catégorie grammaticale est affectée (*POS : Part-of-speech tagging)*. Les mots ramenés à leur forme canonique de base (*Lemmatisation*) peuvent être regroupés en constituants syntaxiques et leur associer un rôle grammatical (*parsing*), parmi ces constituant syntaxiques, des phrases nominales (ou segments répétés) peuvent être extraites comme des concepts (*text chunking*). La troisième étape du processus est l'analyse sémantique (phase C), elle consiste à identifier le sens d'un mot dans une phrase. Le principal défi de cette étape est l'ambiguïté dans le sens d'un concept ou d'une expression (*Word Sense Disambiguation)*. La signification peut représenter un concept dans le VSM au lieu d'utiliser le concept lui-même. Cela mène à un plus grand nombre de concepts extraits mais considère la sémantique d'un concept dans la représentation.

Le processus NLP se trouve au cœur des outils text mining d'extraction des connaissances. Néanmoins, cette extraction ne répond pas souvent aux attentes des utilisateurs; tel que la liste des concepts extraits, en utilisant le processus NLP et des dictionnaires internes de ces outils, nécessite un affinement supplémentaire. Par exemple, Il existe plusieurs concepts extraits et qui indiquent la même chose dans la problématique étudiée, et dans ce cas un groupement de ces concepts est nécessaire dans un seul "*concept clé*" : les concepts « MMS » et « SMS », pour notre cas, peuvent être groupés dans le concept clé « Service » (voir la partie 2). Ou

au contraire, il existe des concepts auxquels on s'intéresse et qui n'apparaissent pas dans la liste finale des concepts extraits, d'où la nécessité d'imposer leur extraction et de les rendre visibles. Ou encore, il existe des concepts extraits et qui ne sont pas intéressants d'où la possibilité de les exclure et de les rendre invisibles.

Le modèle conceptuel d'un *concept clé* donné dans la figure 3 offre une flexibilité dans l'affinement des concepts. Il permet de présenter la liste des concepts extraits sous forme d'un arbre où les *concepts clés* constituent les nœuds. Cette représentation en arbres permet de grouper plusieurs nœuds en un seul (zoom arrière), de dissocier un nœud en plusieurs (zoom avant), de monter ou de descendre un nœud dans tous les niveaux de l'arbre et enfin de rendre un nœud visible ou invisible.

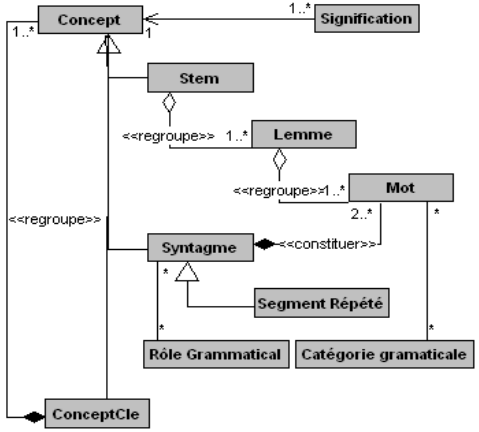

Figure 3: Modèle conceptuel d'un concept clé

2.3 Processus d'enrichissement des dictionnaires des concepts clés

L'élaboration du dictionnaire de concepts clés est un processus itératif semi-automatique et relativement complexe comme l'illustre la figure 4 suivante

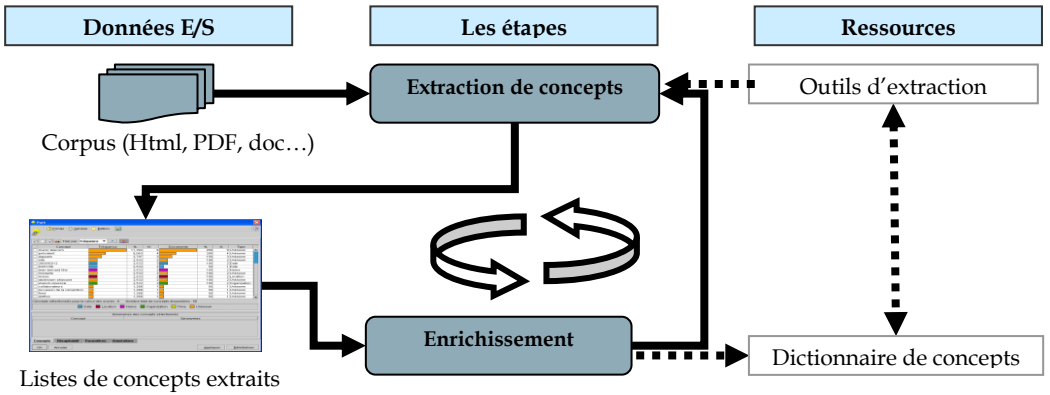

Figure 4: Processus d'enrichissement des dictionnaires.

L'ensemble des concepts existant est alors extrait à partir de tous les documents collectés, en enrichissant le dictionnaire selon l'algorithme suivant.

Faire jusqu'à ce que le résultat d'extraction soit satisfaisant
Début
1. Extraire la liste de concepts par un outil se basant sur la technique NLP.
2. La liste est bien affinée ?
    2.1. Si oui
        2.1.1. Chercher des concepts qui devraient être extraits alors qu'ils n'étaient pas
        2.1.2. Les ajouter au dictionnaire
    2.2. Sinon
        2.2.1. Choisir le concept à affiner
        2.2.2. Exclure ce concept ?
            2.2.2.1 Si oui: Ajouter le concept au dictionnaire et le rendre invisible
            2.2.2.2. Sinon: Regrouper le concept avec ses synonymes, s'ils existent, dans le dictionnaire en spécifiant le concept clé.
3. revenir à 1
Fin

Le dictionnaire de concepts clés doit avoir une structure convenable afin de permettre son intégration et sa participation dans le processus d'extraction en cours.

3 Processus de veille stratégique en télécommunication

3.1 Définition des besoins

L'identification des organisations actives dans le domaine de télécommunication au Maroc dévoile le besoin d'identifier la terminologie technique et industrielle utilisé dans le secteur. Ceci nécessite la conception d'un dictionnaire de tous les concepts utilisés. Un tel dictionnaire devrait normalement permettre de dégager des informations à propos de l'emplacement des opérateurs, leurs dirigeants, leurs partenaires stratégiques, ce qui est acheté ou vendu récemment par chaque opérateur, leurs principales compétences, leurs profil financier ou leurs situation financière, leurs principaux clients, la part de marché par opérateur, les nouveaux produits ou services en développement, leurs stratégies et activités de commercialisation, leurs fournisseurs, etc.

Afin d'établir les thématiques de la veille, nous nous sommes basé sur les profils des opérateurs concurrents sur le marché. Le profil d'un opérateur doit contenir les données nécessaires pour définir, classer et surveiller ce concurrent et ses comportements. Dans le tableau suivant, nous présentons une liste de données dont nous devrons tenir compte dans la définition des thématiques de veille.

A partir des données citées dans le tableau 1, et en fonction de la disponibilité et de la véracité de ces données sur le web, les thèmes de veille suivants ont été sélectionnés :

- ***Offres de produits ou de services* :** Détection de nouvelles offres de services et/ou de produits, ainsi que tout signe d'amélioration ou de détérioration de ces offres, etc.
- ***Ressources financières et de gestion*** : détection de noms de dirigeants ou l'embauche de nouvelles compétences, détection de noms de partenaires (banques, universités, bourses, …), etc.
- ***Concurrents*** : détection des opérateurs concurrents sur le marché du Telecom.
- ***Distributions* :** détection d'ajout de nouveaux emplacements, points de ventes, centres, etc.
- ***Stratégies et démarches de commercialisation*** : détection de nouvelles opérations d'investissement, de partenariat et coopération, de vente et/ou d'achat et établissement de prix, etc.

Tableau 1 : Profil d'un opérateur

Renseignements sur l'opérateur	Nom, Coordonnées, Structure de l'entreprise, Propriété, Historique, Culture d'entreprise
Stature et crédibilité de l'opérateur	Taille, Stabilité, Réputation, Crédibilité
Actifs de marque	Brevets d'invention, Caractéristiques des actifs, Stratégie de marque
Conception des produits/services et innovations	Offre, Qualité (point de vue client), Activités de R&D
Opérations (capacités de production et de prestation des services)	Capacité et ressources internes, Ressources externes
Démarche commerciale	Ventes, Part de marché, Équipe de vente, Volume des ventes
Stratégie de commercialisation et établissement des prix	Marchés cibles, Projets d'expansion, Communication, Établissement des prix
Distribution	Couverture géographique, Réseau commercial, Soutien aux clients

Ressources de gestion	Décideurs clés et personnel d'exécution, Conseil d'administration, Ajouts prévus à l'équipe de gestion
Finances	Ressources financières

3.2 Choix des sources d'information

Il s'agit d'une étape de collecte de l'information du web (source en ligne). Actuellement, il existe 3 opérateurs Telecom au Maroc : Maroc Telecom, Meditel et récemment Wana (Maroc Connect). Les sites de ces trois opérateurs ainsi que tout site web pertinent des entreprises, des administrations publiques et des organisations actives qui sont en relation avec le secteur Telecom au Maroc et qui parlent de ces trois opérateurs ont été consultés. Selon la véracité des sites web consultés, nous avons sélectionné la liste suivante:

- *www.iam.ma*: site de Maroc Telecom,
- *www.meditel.ma*: site web de Meditel,
- *www.wana.ma*: site de wana,
- *www.mobileiam.ma*: site du pôle mobile de Maroc Telecom,
- *www.menara.ma*: site du fournisseur Internet de Maroc Telecom,
- *www.itmaroc.com*: organisation veillant sur le secteur Telecom au Maroc.

3.3 Collecte et organisation de l'information

Cette étape a consisté à rassembler et à classer les documents textuels contenant le maximum possible de concepts et du vocabulaire métier. Pour cela, l'outil Robot v1.2 ([2]) a été utilisé. Robot v1.2 fait partie des outils d'aspiration de pages Web, appelés aussi en anglais *crawler* ou *spider*, qui collecte et maintient de l'information issue d'Internet sur le disque dur de façon intelligente. Contrairement à certains aspirateurs qui consistent à aspirer à chaque requête une collection de pages web, Robot v1.2 télécharge seulement les pages qui ont subi des modifications. Le résultat donné dans le tableau 2 est sous la forme d'une arborescence de dossiers contenants les documents textuels ([3]) collectés :

Tableau 2 : Nombre de documents collectés du Web

Site web	Nombre de pages téléchargées
www.iam.ma	88
www.meditel.ma	120
www.wana.ma	59
www.mobileiam.ma	132
www.menara.ma	184
www.itmaroc.com	183

[2] Outil gratuit de l'éditeur Acetic : www.acetic.fr
[3] Selon le filtre utilisé dans la configuration des options d'analyse de l'outil Robot v1.2, ces documents peuvent être de différents types (html, PDF, doc…).

Total	766

3.4 Analyse et traitement de l'information

Cette étape est la plus longue du processus, elle intègre les trois principales phases du processus du Data mining qui sont : la préparation des données, la modélisation et l'évaluation.

3.4.1 Préparation des données : Elaboration d'un dictionnaire de concepts clés

En vue d'élaborer le dictionnaire de concepts clés, le processus d'enrichissement décrit dans la section 1.3 a été utilisé. L'extraction des concepts a été réalisé par l'outil Text Mining for Clementine (TM4C) de SPSS ([4]) et le processus NLP est au cœur de cet outil. Via un accès direct au corpus, TM4C procède à une extraction automatique des catégories de concepts et leurs fréquences à partir d'un ou plusieurs fichiers textes du corpus en utilisant un dictionnaire interne non éditable ayant des composants différents tel que une liste de formes de phrases prédéfinie pour divers langages (Anglais, Français, Chinois, Japonais,…), des indexes pour des domaines spécifiques (télécommunication, domaine médicale,…) et un ensemble de noms propres pour les besoins de typage des termes extraits (organismes, noms, lieus et produits) ([4]).

Après la constitution d'une liste initiale de termes extraits, cette dernière est affinée afin que les concepts soient plus utiles et répondent bien aux attentes des utilisateurs. En effet, on commence par regrouper les concepts ayant la même signification et qui ont été extrait par l'algorithme NLP comme des concepts distincts. Puis, on exclut les termes non utiles en tant que concepts n'ayant pas une relation avec la problématique considérée. Et enfin, on inclut les concepts que l'on souhaite utiliser et qui n'ont pas été extraits à partir du texte tels que des verbes ou des adjectifs. Un tel affinement de concepts requiert souvent une bonne connaissance de la terminologie du domaine spécifique étudié.

Notons par ailleurs qu'une série de dictionnaires externes est utilisé par TM4C pour affiner les résultats d'extraction. Il s'agit notamment des dictionnaires d'extraction, des dictionnaires de Synonymes, des dictionnaires de Type, dictionnaires de mots clefs,…) ([5]). L'enrichissement de ces dictionnaires est relativement manuel et complexe. Ainsi, il peut être gourmand en terme de temps. Par soucis d'optimalité et plus de rigueur, l'algorithme décrit dans la section 1.3 a été appliqué.

Après un affinement itératif de la liste de concepts extraits, et en fonction des thématiques de veille fixées au départ, nous nous sommes limités à un modèle d'extraction de 12 concepts clefs que nous énumérons ci-après :

1. ***Service* :** le concept le plus fréquent dans le corpus (31.67%) et correspond à tout concept de service offert au client.
2. ***Produit* :** (23.61%) correspond à tout concept de produit ou technologie offerte au client.
3. ***Distribution* :** (9.60%) correspond à tout concept de nouveaux emplacements : villes, régions etc.
4. ***Stratégie commerciale*** : (9.40%) détection de nouvelles opérations d'investissement, de coopération, de vente et/ou d'achat et établissement de prix, etc.
5. ***Ressource financière*** : (6.58%) correspond à toute opération financière telles que des partenariats avec de nouveaux partenaires (les banques, les universités, bourses, etc.), la communication sur les chiffres annuels, etc.

[4] www.spss.com
[5] www.spss.com

6. ***Amélioration*** : (3.70%) regroupe toute désignation d'augmentation, d'ajout, etc.
7. ***Ressource gestion*** : (3.50%) correspond à tout ce qui est en relation avec le noms de dirigeants, de nouvelles compétences et brevets, etc.
8. ***Innovation*** : (1.50%) en relation avec toute nouveauté, innovation, offre, etc.
9. ***Détérioration*** : (0.53%) indiquant toute diminution et/ou affaiblissement.

Les 9 concepts décrits ci-dessus concernent les trois opérateurs Maroc Telecom (3.82%), Meditel (3.90%) et Wana (2.30%) qu'on doit extraire aussi comme concepts cibles.

3.4.2 Modélisation : Elaboration d'un modèle de classification des documents

Le but de la modélisation consiste à structurer automatiquement le corpus sous formes de groupes de documents similaires afin de simplifier considérablement l'accès à ce dernier et d'en tirer des connaissances non triviales. Après avoir élaboré le dictionnaire de concepts clés, spécifique aux objectifs et aux thématiques de la veille fixées au départ, un modèle de classification des documents provenant seulement des sites de Maroc Telecom « www.iam.ma », de Meditel « www.meditel.ma » et de Wana « www.wana.ma » a été élaboré. Les documents du corpus sont représentés dans l'espace vectoriel de dimension 12 (le nombre de concepts clés). Durant la modélisation, trois représentations des documents ont été utilisées dans cet espace. La première utilise la représentation en valeurs booléennes indiquant la présence ou non d'un concept clé dans le document. La deuxième utilise la représentation basée sur la fréquence d'un concept clé dans le document et la troisième représentation utilise le poids d'un concept clé dans le document selon l'équation (1) donnée dans la section 2.1.

Comme dans tout processus de modélisation, les documents du corpus ont été divisés en deux parties : la première partie, dite d'apprentissage et représentant 70% du corpus, est utilisée pour la génération du modèle. La deuxième partie, dite de test et représentant 30% du corpus, est utilisée pour le test.

Pour expliquer les relations existantes entre les 9 concepts dégagés précédemment et 3 opérateurs, nous avons créé un nouveau champ nommé « Opérateur » qui peut avoir les 3 modalités suivante : 'Maroc Telecom', 'Meditel' ou 'Wana'. Ce champ est considéré comme variable cible pour les techniques supervisés. Puis, nous avons appliqué des modèles de règles d'associations, en particulier le modèle GRI (Generalised Rule Induction) [Lallich, 2003]. Avec la représentation en valeurs booléennes nous avons obtenu 7 règles avec un degré de confiances variants entre 70% et 90%, puis nous avons sélectionné les règles suivantes (Cf. Tableau 3):

Tableau 3 : les règles d'associations générées par le modèle GRI

Conséquence	Antécédent	Support	Confiance
Operateur = meditel	Concept_distribution Concept_produit Concept_stratégie-commerciale	52.200	72.630
Operateur = wana	Concept_amélioration Concept_distribution Concept_détérioration	3.850	71.430
Operateur = maroc_telcom	Concept_innovation Concept_ressource-financière Concept_ressource-gestion	4.950	77.780

En text mining, les concepts sont souvent fortement corrélés et lorsqu'on applique des modèles de règles d'associations on abouti souvent à un nombre très important de règles. Aussi, un document doit apparaître si nécessaire, dans plus qu'un seul cluster lors de l'application des techniques de classification non supervisés. De ce fait, la technique de classification non supervisée la plus appropriée est l'analyse en composante principale (ACP) [4] qui, d'une part et contrairement aux autres techniques, ne force pas un document d'appartenir à un et un seul cluster. Les documents ainsi obtenus peuvent être classés selon des facteurs multiples et par conséquent peuvent se trouver à la fois dans plusieurs clusters [6]. D'autre part, les composantes trouvées sont indépendantes et permettant d'améliorer les résultats des techniques supervisées sur les facteurs.

Le tableau 4 regroupe le résultat des modèles obtenus de l'ACP selon le type de représentation utilisée. Dans le cas de la représentation par des valeurs booléennes, le modèle obtenu a généré 4 composantes principales dans le cas de la représentation par des valeurs booléennes (Cf. Tableau 4). Alors que, dans les cas de la représentation par la fréquence absolue et par le poids, le modèle a généré 5 composantes principales correspondant aux cinq axes principaux (ou facteurs) qui sont pondérés par les 9 champs de base (les concepts clés sélectionnés à l'exception des 3 opérateurs télécoms).

Les modèles ont généré :

- Une corrélation élevée du concept « *Meditel* » avec les concepts « *Stratégie-commerciale* », « *Service* » et « *Distribution* ». Cela signifie que l'opérateur Meditel adopte une stratégie commerciale qui se base sur l'offre de services aux différents points de vente.
- Une corrélation élevée du concept « *Maroc Telecom* » avec les concepts « *Ressource-financière* », « *ressource-gestion* » et «*innovation* ». Ce qui indique que le mouvement actuel de l'opérateur Maroc Telecom s'oriente vers l'innovation et la communication sur les bilans annuels et les bénéfices en hausse ainsi que la vente d'actions en bourse.
- Une corrélation élevée du concept « *Wana* » avec les concepts « *amélioration* » et « *produit* ». Ce qui indique que l'opérateur Wana s'oriente vers l'offre de produits améliorés et différents par rapport à la concurrence.

Tableau 4 : Les différents modèles de classification générés par l'application d'une ACP

	Composante			
	1	2	3	4
Concept_distribution	.843			
Concept_service	.756			
Concept_stratégie-commerciale	.743		-.356	
Concept_ressource-financière		.878		
Concept_ressource-gestion		.778		
Concept_détérioration			.873	
Concept_amélioration				.858
Concept_produit	.551		.407	.555
Concept_innovation				

	Composante				
	1	2	3	4	5
Concept_service	.811				
Concept_stratégie-commerciale	.741				
Concept_distribution	.678				
Concept_innovation		.832			
Concept_amélioration		.726			
Concept_ressource-gestion		.518			.475
Concept_ressource-financière			.880		
Concept_détérioration				.915	
Concept_produit					.484

	Composante				
	1	2	3	4	5
Concept_distribution	.809				
Concept_service	.685				
Concept_stratégie-commerciale	.600				.484
Concept_détérioration		.795			
Concept_produit		.755			
Concept_ressource-financière			.506		
Concept_innovation				.789	
Concept_amélioration				.661	
Concept_ressource-gestion					.796

Valeur booléenne	Fréquence absolue	Poids

Après la génération des modèles ACP, nous avons appliqué à nouveau des modèles de règles d'associations GRI, avec les facteurs comme antécédents et le champ *Opérateur* comme conséquence et nous avons eu les résultats suivants :

Tableau 5 : les règles d'associations générées par le modèle GRI

Conséquence	Antécédent	Support	Confiance
Operateur = meditel	$F-Facteur-2 < 0.051378	22.530	100.000
Operateur = meditel	$F-Facteur-1 > 0.952326 $F-Facteur-2 < 0.038054	21.430	100.000
Operateur = meditel	$F-Facteur-1 > 0.952326 $F-Facteur-1 < 0.969492	20.330	100.000
Operateur = maroc_telcom	$F-Facteur-4 < 0.135953	16.480	100.000
Operateur = meditel	$F-Facteur-1 > 0.952326	27.470	94.000
Operateur = maroc_telcom	$F-Facteur-2 > 0.882028	10.990	100.000
Operateur = wana	$F-Facteur-4 > 0.788063	2.200	100.000

Par rapport au tableau 3, on remarque que nous avons amélioré le degré de confiance et aussi le support du modèle de règles d'association GRI.

3.4.3 Test: classement et catégorisation de documents

L'analyse en composante principale permet de donner un score pour chaque document en calculant la valeur de chaque facteur puis de décider les classes (les composantes) gagnantes auxquelles le document doit appartenir. Suivant les composantes principales générées par les modèles de classification ACP précédemment, on a assigné un nom de classe pour chaque document ayant participé dans l'élaboration de ce modèle. Une classe Ci est assignée à un document d s'il contient tous les concepts définissants cette classe. Après l'assignement de classes, les documents du corpus ont été divisés en deux parties : la première partie d'apprentissage, représentant 70% du corpus, est utilisée pour la génération du modèle de classement. Alors que la deuxième, représentant 30% du corpus, est utilisée pour le test.

Le modèle de classement utilisé est l'arbre de décision à base de l'algorithme C5.0, disponible à partir du nœud « C5.0» de Clementine ([6]). Ce nœud génère un arbre de décisions ou un ensemble de règles. Le champ cible à prédire par cette modélisation est le champ classe, alors que les champs d'entrée sont les facteurs calculés lors de l'étape de scoring, par le modèle de classification ACP.

Pour un nouveau document, les concepts clés sont extraits et signalés présents ou non par le modèle d'extraction, puis le modèle de classification est appliqué dans le but de calculer les facteurs sur lesquels se base le modèle de classement qui, à son tour, applique l'ensemble de règles de décision générés pour classer le nouveau document. Ce flux a été appliqué aux documents collectés du site « www.itmaroc.com », et le résultat obtenu est comme suit :

- En utilisant les modèles de classification et de classement générés dans le cas de la représentation par valeurs booléennes, 65% des documents ont été classé parmi ceux qui parlent de « Maroc Telecom » et des « ressources-financières ».
- Alors que l'utilisation des modèles de classification et de classement générés dans le cas de la représentation par la fréquence absolue a donné :
    a. 30.6% seulement de documents qui parlent de « Maroc Telecom » et des « ressources-financières »,
    b. 25.14 % parlent des concepts « Meditel», « Stratégie-commerciale », « Service » et « Distribution »,
    c. Et 17.5% de documents parlent de « Wana », « ressources-gestion » et « Produit ».
- Enfin, l'utilisation des modèles de classification et de classement générés dans le cas de la représentation par le poids a donné :
    a. 36.61% de documents qui parlent de « détérioration » de « produit »,
    b. 28,42% parlent des concepts « Meditel», « Stratégie-commerciale », « Service » et « Distribution »,
    c. 12% parlent de « stratégie-commerciale » et des « ressources-gestion »,
    d. seulement 11% de documents qui parlent de « Maroc Telecom » et des « ressources financières »

4 Conclusion

Dans cet article nous avons présenté une application du processus de veille stratégique appliqué au secteur des télécommunications au Maroc. Le but de cette application et de veiller sur les sites web des opérateurs concurrents afin de détecter leurs mouvements. Dans la première partie nous avons présenté la problématique de veille et sa relation avec l'extraction de connaissance et le text mining en général. Dans la deuxième partie nous avons présenté un modèle conceptuel de concepts clé pour l'extraction des concepts, puis nous avant présenté un algorithme que l'on a adopté pour l'enrichissement d'un dictionnaire de

[6] Documentations SPSS: Référence des nœuds Clementine 11.1, Clementine Algorithms Guide.

concepts spécifique au domaine étudié. Ensuite dans la troisième partie nous avons présenté le processus de veille appliqué à des documents collectés du web et selon la véracité des informations extraites. Nous avons pu détecter que l'opérateur Meditel adopte une stratégie commerciale qui se base sur l'offre de services aux différents points de vente, que le mouvement actuel de l'opérateur Maroc Telecom s'oriente vers l'innovation avec des services nouveaux (TV sur ADSL par exemple) et la communication sur les bilans financiers et les bénéfices en hausse, alors que l'opérateur Wana s'oriente vers l'offre de produits améliorés et différents des concurrents. Puis nous avons présenté le processus de catégorisation de nouveaux documents, en montrant que l'on améliore de plus en plus nos modèles de segmentation et de classement lorsqu'on utilise respectivement la représentation par valeur booléenne, fréquence absolue et par le poids d'un concepts.

5 Bibliographie

[1] BLOEHDORN S. et HOTHO A., *Text classification by boosting weak learners based on terms and concepts*, Proceedings IEEE International Conference on Data Mining (ICDM 04), 1-4 November 2004, P 331-334.

[2] DEROUET D., LEPOIVRE F., *veilles, processus et méthodologie*, NEVAOCONSEIL, 2005.

[3] HAARSLEV V., *Introduction to Artificial Intelligence*, Department of Computer Science and Software Engineering, Concordia University, Montreal, 2004.

[4] HAIR J.F., TATHAM J.L., ANDERSON R.E., BLACK W., *Multivariate data analysis* (3rd edition), New York Macmillan, 1992

[5] HOTHO A., STAAB S., et STUMME G., *Ontologies improve text document clustering*, Proceedings IEEE International Conference on Data Mining (ICDM 03), 19-22 November 2003, p 541-544

[6] KONGTHON A., *A Text Mining Framework for Discovering Technological Intelligence to Support Science and Technology Management*, PhD Thesis, Georgia Institute of Technology, April 2004.

[7] LALLICH S., TEYTAUD O., *Évaluation et validation de l'intérêt des règles d'association*, n° spécial *Mesures de qualité pour la fouille des données*, Revue des Nouvelles Technologies de l'Information, RNTI-E-1, 2004, p. 193-218.

[8] LEOPOLD E. et KINDERMANN J., *Text categorization with support vector machines. How to represent texts in input space?*, Machine Learning, Volume 46, Issue 1-3, 2002, p 423-444

[9] LESCA H., *Veille stratégique : concepts et démarches de mise en place dans l'entreprise*, Guide pour la pratique de l'information scientifique et technique. Ministère de l'Education Nationale, de la Recherche et de la Technologie, 27 p., 1997.

[10] SALTON G., WONG A., et YANG C. S., *A vector space model for automatic indexing*, Communications of the ACM, Volume 18, Issue 11, 1975, p 613-620.

[11] SALTON G., *Introduction to Modern Information Retrieval*, McGraw-Hill, 1983.

[12] SALTON G. et BUCKLEY C., *Term weighting approaches in automatic text retrieval*, Information Processing & Management, Volume 24, Issue 5, 1988, p 513-523.

[13] WHITE C., *Consolidating, Accessing and Analyzing Unstructured Data*, Business Intelligence Network, December 12, 2005.