Une approche distribuée pour l'extraction de connaissances : Application à l'enrichissement de l'aspect factuel des BDG

Khaoula Mahmoudi*
Sami Faïz ** ***

* Ecole Supérieur des communications de Tunis (SUPCOM)
khaoula.mahmoudi@insat.rnu.tn
** Institut National des Sciences Appliquées et de Technologie (INSAT)
*** Laboratoire de Télédéction et Systèmes d'Informations à Références Spatiales (LTSIRS)
sami.faiz@insat.rnu.tn

Résumé. Les systèmes d'informations géographiques (SIG) sont utilisés pour améliorer l'efficacité des entreprises et des services publics, en associant méthodes d'optimisation et prise en compte de la dimension géographique. Cependant, les bases de données géographiques (BDG) stockées dans les SIG sont restreintes à l'application pour laquelle elles ont été conçues. Souvent, les utilisateurs demeurent contraints de l'existant et se trouvent dans le besoin de données complémentaires pour une prise de décision adéquate. D'où, l'idée de l'enrichissement de l'aspect descriptif des BDG existantes. Pour atteindre cet objectif, nous proposons une approche qui consiste à intégrer un module de fouille de données textuelles au SIG lui même. Il s'agit de proposer une méthode distribuée de résumé de documents multiples à partir de corpus en-ligne. L'idée est de faire coopérer un ensemble d'agents s'entraidant afin d'aboutir à un résumé optimal.

1 Introduction

Le but d'un SIG est de fournir une aide à la décision dans des domaines divers. Souvent, il sert à produire des cartes répondant à un besoin spécifique. Il peut être utilisé pour associer une densité de population à chaque région sur une carte, la représentation de la présence de consommateurs potentiels d'un produit ou d'un service dans une région, etc. Les données sont dans tous les cas restreintes à l'application en cours et parfois on a besoin d'avoir des informations au-delà de ce qui est stocké dans la BDG. A titre d'exemple, une BDG créée pour une application de découpage administratif ne permet pas de fournir une réponse à une requête faisant intervenir des informations d'ordre économique, historique, etc. D'où, l'idée d'offrir des sources complémentaires d'informations sans nuire aux données préalablement fournies (Faïz et Mahmoudi, 2005). Pour atteindre cet objectif, nous avons bâti une approche pour la génération automatique de résumés de documents multiples pour fournir les informations complémentaires relatives aux entités géographiques manipulées par le SIG. Cette approche est basée sur trois types d'agents coopérant afin d'aboutir à un résumé optimal. Il s'agit d'un agent *interface*, des agents *entité* (géographiques) et des agents *tâche*. La communication entre ces agents est assurée par l'envoi de messages. L'approche est

modulaire, elle peut être décomposée en trois grandes phases. Il s'agit de la segmentation et de l'identification des thèmes, l'affectation d'agents délégués, et enfin le filtrage consistant à éliminer toute portion de texte qui s'avère inutile à la compréhension du thème discuté.

Dans cet article, nous présentons à la section 2, un aperçu sur les SIG et l'univers multi-agents. La troisième section est dédiée à la présentation générale des différentes approches de résumés de documents multiples. Dans la quatrième section, nous détaillons notre approche de génération de résumés à partir de corpus en-ligne (plus précisément des documents textuels au format HTML) en vue d'alimenter et compléter l'information descriptive au sein des BDG. Enfin, dans la dernière section, nous présentons notre prototype à travers un exemple d'extraction de résumés textuels se rapportant aux pays méditerranéens.

2 Etat de l'art

2.1 Systèmes d'Informations Géographiques (SIG)

Un Système d'informations géographiques (SIG) permet de gérer des données alpha-numériques spatialement localisées (Faïz, 1999). Il permet à partir de diverses sources, de rassembler, de gérer, d'analyser, et de présenter des informations localisées géographiquement, contribuant à la gestion de l'espace. A l'origine les utilisations des SIG, ont concerné le milieu de la recherche dans le domaine de la géologie et de l'environnement, mais, actuellement elles ont été étendues pour permettre d'améliorer l'efficacité des entreprises et des services publics, en associant méthodes d'optimisation et prise en compte de la dimension géographique. Un SIG est un outil permettant de manipuler des BDG. Ces bases de données sont constituées, à la fois, de *données descriptives ou factuelles* de type chaîne, nombre, date, etc. (par exemple, le nom d'une route, le type d'un bâtiment ou le débit d'un cours d'eau) et également d'autres données dites *spatiales* spécifiques à la position (localisation) des entités géographiques manipulées. Ces entités sont principalement de trois types : point, ligne ou polygone. A titre d'exemple, l'entité rivière peut être considérée comme un objet de type linéaire et un pays peut être vu comme un polygone ou un point selon l'échelle de visualisation.

2.2 Systèmes Multi-Agents (SMA)

Un système multi-agents (SMA) (Briot et Demazeau 2001, Mahmoudi et Ghédira 2000) est un système distribué composé d'un ensemble d'agents. Au sein d'un SMA, il n'y a aucun contrôle global, les données sont décentralisées et le calcul est asynchrone. Ces systèmes abritent ce qu'on appelle agent. Il n'y a pas de définition d'un agent accepté à l'unanimité et cette notion est utilisée dans un bon nombre de domaines à savoir: sociologie, biologie, informatique... Ici nous adoptons la définition de Ferber « Un agent est une entité autonome, réelle ou abstraite, qui est capable d'agir sur elle-même et sur son environnement, qui, dans un univers SMA, peut communiquer avec d'autres agents, et dont le comportement est la conséquence de ses observations, de ses connaissances et des interactions avec les autres agents » (Ferber, 1997). L'organisation des agents au sein de tel système doit passer par le traitement de questions fondamentales, les plus importantes sont la communication et la coopération entre les agents. Les protocoles de communication sont de deux types : par partage d'information (appelé «tableau noir» ou «blackboard») ou par envoi de message. La

coopération peut se ramener à résoudre les différents sous-problèmes que constituent la collaboration par répartition des tâches, la coordination d'actions et la résolution de conflits.

3 Résumé automatique de documents multiples

Résumer, consiste à, partir d'un texte source pour générer un texte cible consistant en une transformation réductrice du texte source vers un résumé par compression du contenu. La nécessité de résumer est née du fait que la masse d'information textuelle existante sous forme électronique ne cesse d'augmenter régulièrement. L'usager, qui n'a plus de temps ni de ressources cognitives pour faire face à un tel volume d'informations, se trouve dans une situation de saturation. La fonction de résumé automatique est alors proposée, elle constitue un moyen efficace et éprouvé pour représenter les contenus textuels et faire des économies de temps. Les premiers travaux dans ce domaine se sont d'abord intéressés au résumé d'un document unique, puis, pour des besoins de veille technologique, les chercheurs se sont penchés sur le résumé multi-documents. Avec la deuxième catégorie, l'usager est confronté à des défis supplémentaires, à savoir : identifier, ce qui est commun et ce qui diffère dans une variété de documents reliés et enlever la redondance du résumé. La littérature a révélé une panoplie de méthodes inspirées de différentes disciplines. On trouve des approches qui adoptent les méthodes statistiques, d'autres utilisant le traitement du langage naturel, une troisième classe, exploitent les techniques d'extraction d'information. Finalement, des approches mixtes combinant toutes ces techniques ont été proposées. Dans ce qui suit, nous présentons un aperçu de quelques approches, appartenant aux catégories susmentionnées.

Mani et Bloedorm (1999) ont proposé une approche transformant un texte en un graphe où chaque mot est un nœud et où les liens conceptuels entre les nœuds sont les arêtes. Cette approche est limitée à deux documents, fournissant en sortie deux graphes. Un algorithme appelé *spreading activation***,** permet d'identifier les nœuds output qui sont en relation avec le thème en question. Ceci est répété jusqu'à épuiser tous les nœuds des deux graphes. L'algorithme *Find Similarities and Diffferences* est appliqué, ayant pour conséquence de dégager les « meilleures » phrases similaires et les « meilleures » phrases différentes entre les deux documents qui feront partie du résumé.

L'approche MEAD (Radev et al., 2000) génère des résumés en utilisant les centroïdes des clusters produits par le système *Topic Detection and Tracking* (Allan et al., 1998). Cette approche, calcule l'utilité des phrases (allant de 0 à 10) pour déterminer le degré de pertinence par rapport au thème générale d'un cluster donné. La valeur 0 signifie que la phrase n'est pas pertinente et la valeur 10 marque une phrase essentielle. Le principe de *cross-sentence informational subsumption,* est appliqué par la suite, afin de marquer les phrases qui rapportent les mêmes informations avec plus ou moins de détails. Ainsi, une seule phrase est maintenue de chaque classe selon le niveau de détails désiré. Enfin, chaque cluster est décrit par des phrases pertinentes et non-redondantes.

L'approche GISTexter (Harabagiu et Finley, 2002) adopte les idées d'Extraction d'Information (EI). Il s'agit de prédéfinir des gabarits (templates) dont les *slots* spécifient les informations importantes à rechercher. Ces templates sont instantiés en utilisant un système EI existant; appelé CICERO, tout en maintenant le lien avec les blocs de textes qui ont servis pour l'instantiation. Ces derniers forment le résumé final. Si jamais le thème n'est pas encodé préalablement dans le CICERO, le GISTexter génère d'une manière ad-hoc, un

ensemble de *templates* et un ensemble de *patterns d'extraction* qui sont des règles qui guident le CICERO à détecter les informations pertinentes.

Gees et al. (2000) décrivent une approche qui génère des résumés pour chaque document du corpus. Ces derniers sont regroupés en clusters selon la similarité de leurs thèmes. Enfin, un résumé représentatif est choisi pour chacun de ces clusters.

L'approche connue sous l'acronyme MMR pour Maximal Marginal Relevance (Goldstein et al., 2000) segmente les documents en passages. Les passages pertinents sont identifiés en mesurant la similarité, par rapport, à une requête d'un usager. Les passages en dessous d'un seuil donné sont écartés. Pour le reste, la métrique MMR est appliquée. Selon cette mesure, un passage textuel a une haute *marginal relevance* s'il est pertinent à la requête, en même temps ayant une similarité minimale avec les passages précédemment sélectionnés. Les passages sélectionnés constituent le résumé du corpus.

NEATS (Lin et Hovey, 2002) est un système résultant d'un mixage de plusieurs techniques prouvées efficaces dans le résume de document unique, tel que, la fréquence des termes, la position des phrases, les mots *stigma*, et également, une version simplifiée de MMR. Etant donné un groupe de thèmes, un ratio de probabilité λ est utilisé pour identifier les concepts clés en unigrams, bigrams, et trigrams (Kraaij et al., 2002). La signature résultante est sauvegardée sous forme d'un arbre et ceci après élimination de mots ou phrases sous un seuil bien déterminé. L'ordonnancement de ces phrases est accompli ainsi qu'un filtrage de contenu selon le critère simplifié de MMR. Finalement, pour s'assurer de la cohérence chronologique, les expressions temporelles sont mentionnées explicitement dans le texte en indiquant les dates effectives.

(McKeown et al., 1999) propose une méthode par reformulation. L'idée est d'extraire des fragments ou clauses à partir des phrases les plus saillantes puis les fusionner pour générer un nouveau texte. Cette méthode transforme les informations en résumé dans des formats d'entrée de modules de génération de langue en utilisant le système FUF/SURGE (Elhadad et Robin, 1996).

A l'opposé de cet ensemble de méthodes qui proposent toutes une résolution centralisée du problème de génération de résumés multi-documents, notre méthode est **totalement distribuée**. La distribution est justifiée par le fait que le problème peut être vu comme naturellement distribué si on suppose que chaque document est un agent qui cherche sa satisfaction et coopère avec les autres pour atteindre un but commun. Nous procédons, ainsi, à des économies de temps qui résultent du traitement parallèle tout en respectant les problèmes inhérents à la génération de résumés automatiques à partir de documents multiples.

4 Une approche distribuée de résumé de documents multiples

4.1 Processus général

Tout le travail se déclenche lorsqu'un utilisateur cherche des informations concernant des entités géographiques affichées sur une carte. Ainsi, l'utilisateur peut manipuler la carte en localisant par pointé ou fenêtrage respectivement, une entité ou une zone composée d'un ensemble d'entités. Une autre alternative est envisageable, il s'agit d'utiliser un système de requêtage permettant d'identifier les entités désirées. Une dernière alternative étant la

combinaison de ces deux techniques : Nous pouvons exécuter, tout d'abord, une requête pour déterminer les entités répondant à certaines contraintes utilisateurs. Puis, en manipulant la carte, nous choisissons les éléments qui nous intéressent. Ces manipulations font intervenir la BDG pour extraire des informations relatives aux entités sélectionnées.

Si les résultats escomptés ne sont pas atteints, notre prototype de génération de résumés est lancé : Le processus se déclenche en effectuant un accès Web pour ramener les documents en relation avec les entités géographiques. Deux cas de figures sont envisageables, selon qu'on manipule une entité ou toute une zone. Dans le premier cas, un agent *interface* qualifié de coordinateur est créé par le système. Il collecte les documents relatifs à l'entité en question. Cet agent crée à son tour des agents *tâche* qui se chargent de traiter les documents.

Dans le cas d'un ensemble d'entités géographiques, l'agent *interface* déclenche une collecte d'informations en relation avec les entités appartenant à la zone. Cette collecte se fait d'une manière parallèle. En fait l'*interface*, crée des agents *entité* chacun responsable d'une entité géographique faisant partie de la zone. Les documents ramenés par chaque agent *entité* seront confiés à des agents *tâche* pour traiter leurs contenues.

En fait, quel que soit le cas, chaque agent *tâche* segmente le texte de son document en fixant les frontières thématiques. Une fois les segments cernés, chaque agent *tâche* identifie les thèmes correspondants à chacun de ses segments. Si on manipule une seule entité, les thèmes sont envoyés à l'agent *interface* qui les regroupe ensemble selon le degré de similarité des thèmes. Par la suite, il affecte un délégué pour chaque thème et ce en tenant compte du coût engendré par sa décision. Il s'agit en fait de minimiser la surcharge de chaque délégué tout en distribuant au maximum la tâche de génération de résumé entre le plus grand nombre d'agents. Si on manipule un ensemble d'entités géographiques, le traitement est le même sauf que la communication s'effectue entre les agents *tâche* et les agents *entité*. Autrement dit, tout le traitement fait par l'agent *interface* sera fait par chacun des agents *entité*.

A l'issue de la décision de délégation, chaque délégué doit interagir avec ses accointances (les agents qui traitent le(s) thème(s) sous sa responsabilité) par envoi de messages afin de collecter les segments correspondants. Désormais, le délégué tient à sa disposition un ou plusieurs documents virtuels. Un document virtuel ne résulte pas d'une tâche de recherche d'informations, mais de la collecte de segments issus des différents agents *tâche* et traitant des thèmes similaires. Enfin, chaque délégué extrait les informations les plus saillantes qui résument mieux ses documents virtuels. Le résumé final résulte de l'assemblage des résumés partiels générés par chaque agent délégué.

Ce résumé est stocké dans la BDG. Nous stockons les thèmes avec les fragments de textes correspondants avec une référence aux documents qui les traitent. Dans différents cas, le stockage des résultats est justifié : (i) résultats disponibles à l'exploitation sans avoir recours à la ré-exécution du processus global, (ii) l'ensemble de la méta-données constituée devient une base de réflexion, voire un support de prédiction puisque tous les résultats à différentes dates d'extraction y sont stockés. Le pseudo-code décrit par la figure 1 récapitule le processus de résumé à partir d'un corpus de documents.

Les sections suivantes détaillent les différentes phases de génération automatique de résumés.

Algorithme d'enrichissement

Utilisateur : Entité (s) Géographique (s) (EG) (pointé / fenêtrage / requêtage)

Téléchargement des pages web

☺ **Agent *tâche***

- Pré-traitement des pages web (suppression des mots vides, lemmatisation)
- Segmentation et identification de thème
- *Notifier(interface, liste-thèmes)* (si une EG) *Notifier (entité, liste-thèmes)* (si une zone={EG})

☺ **Agent *interface* (si une EG) ou Agent *entité* (si une zone={EG})**

- Regroupement par similarité des thèmes
- Prise de décision selon la valeur de la fonction de coût *f*
- *Notifier (tâche, liste-thèmes) //désormais l'agent concerné est appelé délégué*

☺ **Agent *délégué***

- Création des documents virtuels
- $\Re_i \leftarrow$filtrer (Documents virtuels)
- *Notifier (interface,* $\Re_i$*)* (si une EG) ou *Notifier (entité,* $\Re_i$*)* (si une zone={EG})

☺ **Agent *interface* (si une EG) ou Agent *entité* (si une zone={EG})**

Output: résumé$\leftarrow \bigcup_{i=1}^{n} \Re_i$

Utilisateur : Intégration du résumé dans la BDG

FIG. 1 – *pseudo-code du processus de génération de résume pour l'enrichissement du BDG*

4.2 Segmentation et identification de thème

Les agents *tâche* déterminent en premier lieu, les limites entre les blocs textuels et leurs thèmes associés. Cette tâche est à entreprendre simultanément par les différents agents *tâche* et ce pour chaque document appartenant au corpus. Une phase de pré-traitement est accomplie au préalable. Il s'agit d'éliminer les mots non porteurs de sens (les prépositions, les articles…) et l'application d'un algorithme de lemmatisation afin de retenir seulement les racines des mots. Dans ce qui suit nous parlons de tokens plutôt que de mots. Afin de fixer les frontières entre les blocs textuels, nous avons adopté l'algorithme de TextTiling (Hearst, 1997). L'entité de base pour cet algorithme est le bloc de texte qui est défini par un nombre fixe de phrases. Un score est accordé pour chacun des blocs en fonction du bloc qui le suit, déterminant à quel point les blocs adjacents sont similaires. Le calcul de la similarité est défini par la formule suivante et ce pour chaque «*gap*» *i* séparant deux blocs adjacents.

$score(i) = \frac{\sum_t w_{t,b1} w_{t,b2}}{\sqrt{\sum_t w_{t,b1}^2 \sum_t w_{t,b2}^2}}$ où t est un token ; $w_{t,b}$ est le poids (fréquence) de t dans le bloc b; b_1 et b_2 sont deux blocs textuels (séquence de tokens) adjacents. La profondeur de chaque «*gap*» est évaluée pour déterminer les vallées les plus profondes qui marquent les segments.

Une fois les blocs détectés nous identifions leurs thèmes. Nous associons à chaque segment les *n* plus fréquents tokens. Ces derniers sont utilisés pour déterminer approximativement les concepts à partir du thésaurus *WORDNET* (Miller, 1990). Ce thésaurus est le plus

couramment utilisé pour l'extraction des relations sémantiques (synonymes, hyponymes…). Par exemple, « agriculture », « farming » et « plantation » sont considérés faire référence au même concept. Les thèmes ainsi dégagés sont envoyés à l'agent *interface* ou à un agent *entité* via le message *listeThème (tâche, Int|entité, ensemble-thème)*. En fait, dans cet article nous présentons un message en respectant la syntaxe suivante : *IDmessage (émetteur, récepteur, paramètre)*. *IDmessage* est l'identifiant du message, *émetteur* et *récepteur* peuvent avoir comme valeur un type d'agent *(tâche*, *délégué*, *entité* ou *interface* (*Int))* et *paramètre* définit ce qui est envoyé par l'émetteur au récepteur.

4.3 Délégation

Cette tâche a pour finalité d'affecter un délégué responsable de la condensation d'un ou de plusieurs documents virtuels. En fait, *l'interface* (ou l'agent *entité*) en recevant les thèmes procède à leur regroupement selon leur similarité. La décision de délégation est prise en tenant compte du coût engendré par chaque affectation et qui est présenté formellement par la fonction suivante : $f(s) = \alpha\ workload + \delta\ communication$

où s est une allocation possible. α et δ sont les coefficients de pondération déterminés expérimentalement. La charge du travail noté "*workload*" est définie pour chaque délégué k comme suit: $workload = \sum_i \sum_j segment - size_{ij}$

où i est un des thèmes affectés à l'agent k. j représente un des agents traitant i. $segment\text{-}size_{ij}$ désigne la taille (nombre de phrases) d'un segment issu de l'agent j traitant le thème i.
La communication est définie par :

$$communication = \sum_j \beta \text{ où } \beta = 1 \text{ et } j \in \text{ accointances de l'agent } k.$$

communication vaut 0 si l'agent k est un délégué d'un des thèmes qu'il traite dans son document original, sinon, *communication* vaut le nombre d'agents traitant le thème. En fait, chaque agent *tâche* déjà affecté est écarté de la liste des candidats pour distribuer au maximum la charge du travail entre la société d'agents. Dans le cas où certains thèmes restent non affectés avec un nombre de candidats égal à 0, on reconsidère tous les agents *tâche* (traitant ou non le thème en question) et on procède à l'affectation des thèmes non affectés aux agents tout en minimisant le coût *f*. Ainsi, un agent peut être responsable de la condensation de 0, 1, ou plusieurs thème(s) qu'il traite ou pas dans son propre texte. Une fois la décision de délégation prise par l'agent *interface(entité),* une notification est envoyée aux agents concernés via le message, *Délégation(Int|entité, Délégué, liste-thèmes)*. Ainsi, un délégué peut être responsable de(s) thème(s) qu'il traite dans son document original et/ou d'autre(s) thèmes en provenance d'autres agents. Dans le premier cas, le délégué sollicite ses accointances (selon les thèmes) à fournir leurs segments en envoyant le message *demandSegment(Délégué, tâche, liste-thèmes)*. La réponse est expédiée via : *textThème(tâche, Délégué, liste-segments)*. Au cas où le délégué est responsable d'un thème qu'il ne traite pas dans son document original, il doit identifier les agents concernés. Il doit contacter l'agent *interface(entité)* à travers le message*: membres(Délégué, Int|entité, liste-thèmes)*. Ce dernier connaît tous les agents du système. L*'interface(entité)* communique la liste des agents à travers *listeMembres(Int|entité, Délégué, membres)*. En effet, un délégué ne se met à condenser ses documents virtuels que s'il a reçu tous les messages *textThème* concernant tous les thèmes sous sa responsabilité.

4.4 Extraction de textes

Une fois l'opération de délégation accomplie par l'agent *interface ou entité*, chaque délégué, selon le choix de l'utilisateur, peut considérer un seul segment comme représentant du document virtuel, tout le document virtuel, le plus long segment en espérant l'obtention d'un maximum de détails ou un extrait à partir des segments. Pour chaque segment sélectionné, le délégué dérive les portions de textes les plus pertinentes et ce en construisant l'arbre rhétorique qui décrit sa structure. Inspiré des idées de Mann et Thompson (Mann et al. 1988, Marcu 1999), cette théorie annonce que le texte, peut être vu comme un arbre binaire dont les nœuds sont des satellites ou des nucléus (appelés aussi noyaux). Un nucléus est la partie d'une phrase ou d'un segment textuel qui expose l'idée importante de l'auteur, alors que le satellite est la partie subordonnée qui vient soutenir le nucléus. En effet, le satellite augmente la crédibilité du noyau aux yeux du lecteur. En plus, chaque nœud est caractérisé par le type de la relation rhétorique (objectif, exemple, élaboration...) qui tient entre le nucléus et le satellite, et l'ensemble de promotions (l'ensemble des unités textuelles les plus pertinentes). En fait, générer un résumé peut être réduit à l'extraction des promotions en les classifiant par ordre d'importance selon la proximité de la racine qui est considérée comme la clé du segment, et afin de promouvoir les unités textuelles qui contiennent les thèmes identifiés lors de la phase précédente, nous augmentons leurs scores pour augmenter leur chance de faire partie du résultat final. Nous illustrons ce propos à travers un extrait de texte qui traite une stratégie d'irrigation.

[To conserve water resources and encourage demand management in the irrigation sector,^{A1}] [a national water saving strategy was implemented.^{B1}] [As part of the strategy, a number of reforms were introduced in the past few years,^{C1}] [for instance, the promotion of water users' associations, the increase in the price of irrigation water, etc.^{D1}]

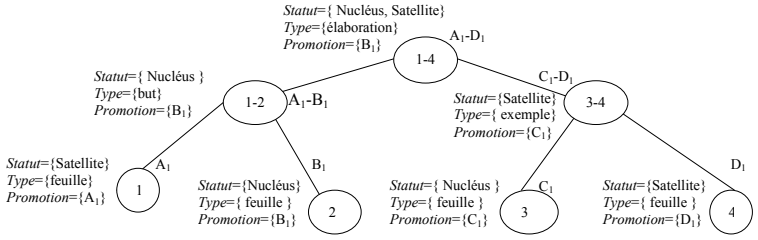

FIG. 2 – *Arbre rhétorique du texte*

Dans la figure 2, la relation *but* tient entre le satellite A_1 et le nucléus B_1. D_1, illustre par des exemples ce qui est annoncé dans C_1 qui est la plus importante pour l'auteur. A_1-B_1 (plus précisément sa *promotion*: B_1) est relié à C_1-D_1 (plus précisément sa *promotion*: C_1) par une relation d*'élaboration*. En fait, C_1 fournit des détails relative à B_1. Ainsi, A_1-B_1 est le nucléus supporté par le satellite C_1-D_1. La racine (1-4) qui représente tout l'extrait de texte; A_1-D_1, peut être un nucléus ou un satellite par rapport aux autres extraits (ou unités textuelles) du texte initial. La partie la plus pertinente de tout l'extrait est B_1.

En procédant à l'extraction des ensembles de promotions nous obtenons l'ordre suivant : $B_1 > C_1 > A_1, D_1$. Selon le degré de compression du résumé, nous pouvons obtenir un résume de taille 1 en extrayant la racine (B_1), de taille 2 en ajoutant à B_1, C_1 et ainsi de suite. En

appliquant le même raisonnement, chaque délégué extrait les informations qui résument mieux ses documents virtuels.

Si l'utilisateur sélectionne plus qu'un segment à partir du document virtuel, nous devons éliminer la redondance qui peut éventuellement se produire. Pour ce faire, nous calculons la similarité entre les segments comme suit : $\text{Similarité}(S_1, S_2)=\frac{\vec{S_1} \times \vec{S_2}}{|\vec{S_1}| \times |\vec{S_2}|}$, où chaque segment ($S_1$ et S_2) est vu comme un vecteur de nombres (fréquence des tokens au sein du segment), avec ses composants : les tokens qui y sont détectés. De cette manière, le délégué ne considère que les segments qui sont non similaires selon le seuil de similarité préalablement fixé.

5 Réalisation

Un prototype démonstratif avait été développé. Il s'agit de manipuler la carte des pays de la méditerranée (cf. la figure 3). Nous avons utilisé Java comme langage de développement, étant donnée qu'il est multi-threading supportant bien la concurrence et le parallélisme. La figure 3 montre l'intérêt de l'utilisateur vis à vis des informations relatives à l'entité pays (dans notre cas, la Tunisie). La fenêtre est principalement formée par deux onglets ; un pour l'affichage des documents virtuels et l'autre pour les résumés. Suite à la consultation du résumé, l'utilisateur peut afficher à sa guise l'intégralité du document s'il le désire. En fait, l'utilisateur a le choix entre sauvegarder les résumés, un des segments, ou juste un extrait.

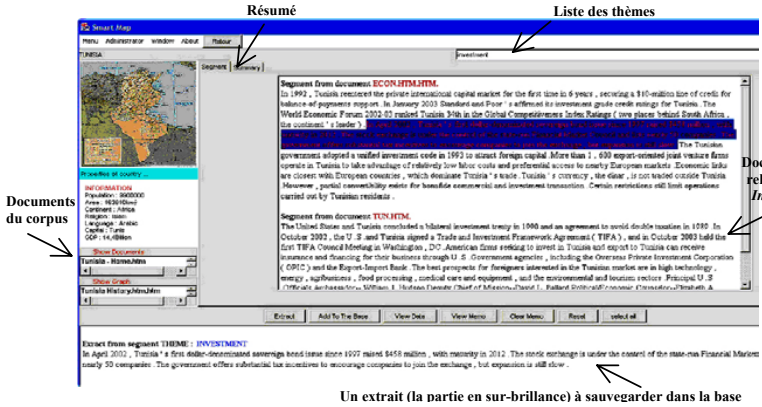

FIG. 3 – *Capture d'écran de notre prototype de génération de résumés multi-documents*

5.1 Exemple d'illustration

Dans ce qui suit, nous donnons un exemple de résultats fournis par notre prototype. Les résultats obtenus ont été validés par des experts humains qui les ont jugés satisfaisants. Cet exemple est un extrait d'un document virtuel relatif au thème *Trade* se rapportant à l'entité pays : la Tunisie. Ci-dessous, un extrait de ce document virtuel DOC_1.

Bloc 1 : http://www.worldbank.org/fandd/english/0996/articles/040996.htm#author (28/09/2005)

During the past decade, Tunisia has adopted an economic reform strategy aimed at establishing a market-based and private-sector-driven economy that is increasingly open to world goods and capital markets. A gradual liberalization of Tunisia's trade was launched in the mid-1980s. This was complemented by liberalization of the exchange system, which culminated in 1993 in the establishment of current account convertibility and the adoption of the obligations under Article VIII of the International Monetary Fund's Articles of Agreement. Tunisia became a full member of the General Agreement on Tariffs and Trade (GATT) in 1990 and is a founding member of the World Trade Organization. The opening of Tunisia's economy was reflected in its growing integration into the world economy. Led by the rapid growth of its non-energy exports, particularly in the textile sector, Tunisia's export share in its traditional markets (mainly Europe) increased steadily during 198595, and the share of exports of goods and nonfactor services in its GDP trended upward, to an average of 41 percent during 199195 from 35 percent during 198285.

Bloc 2 : http://www.infoprod.co.il/country/tunis2f.htm (28/09/2005)

Tunisia has entered into trade agreements with forty-one developed and developing countries, which granted Tunisia most-favored-nation status. Tunisia has entered into bilateral and regional trade preference agreements with the European Union and the Arab Maghreb Union as well as certain agreements under the framework of the Inter-Arab Cooperation, the Inter-African Cooperation and the Organization of the Islamic Conference. Furthermore, Tunisia is a member of the world trade organization (WTO) and is a signatory to the Global System on Trade Preferences. In 1995, the Tunisian government and the European Union negotiated a major economic agreement on free trade. The pact establishes the framework for free trade between Tunisia and the European Union.

Bloc 3 : http://www.state.gov/e/eb/ifd/2005/43042.htm (28/09/2005)

With few exceptions, domestic trading can only be carried out by a company established under Tunisian law with majority capital ownership and management held by Tunisians. An additional barrier regarding investments by non-European Union (EU) investors is contained in Tunisia's Association Agreement with the EU. The EU provides massive funding to Tunisia, especially for infrastructure development, but such funding often contains conditions prohibiting non-EU member investors from participation. In order to boost bilateral trade and to address U.S. investment issues in Tunisia, the U.S. government has begun a dialogue on free trade with the government of Tunisia. The first, formal step was launched in 2002 with the signing of a Trade and Investment Framework Agreement (TIFA) to formalize discussions on bilateral trade and investment. A TIFA Council convened in October 2003 as an initial step, but progress on trade liberalization with the U.S. has not moved forward significantly. Recently, however, Tunisian government officials, including the head of the Central Bank of Tunisia, have emphasized the importance of diversification of external investment sources. A large share of Tunisia's FDI in recent years has come from a privatization program to sell off state-owned or state-controlled enterprises.

Ci-dessous, un exemple de résumé généré par notre prototype à partir de DOC_1 :

A gradual liberalization of Tunisia's trade was launched in the mid-1980s. This was complemented by liberalization of the exchange system. Tunisia became a full member of the General Agreement on Tariffs and Trade (GATT) in 1990 and is a founding member of the World Trade Organization.

Tunisia has entered into trade agreements with forty-one developed and developing coutries. Tunisia has entered into bilateral and regional trade preference agreements with the Eurpean Union and the Arab Maghreb Union as well as certain agreements under the framework of the Inter-Arab Cooperation, the Inter-African Cooperation and the Organization of the Islamic Conference.

With few exceptions, domestic trading can only be carried out by a company established under Tunisian law with majority capital ownership and management held by Tunisians. The EU provides massive funding to Tunisia, especially for infrastructure development, but such funding often contains conditions prohibiting non-EU member investors from participation.

The U.S. government has begun a dialogue on free trade with the government of Tunisia. A TIFA Council convened in October 2003 as an initial step, but progress on trade liberalization with the U.S.

has not moved forward significantly. A large share of Tunisia's FDI in recent years has come from a privatization program to sell off state-owned or state-controlled enterprises.

6 Conclusion

Afin d'enrichir le contenu des bases de données, nous avons proposé une approche distribuée pour la génération automatique de résumés à partir d'un corpus textuel. En conformité avec l'univers multi-agents, un ensemble d'agents s'entraident pour atteindre un but commun (le résumé optimal) dans des délais très courts. L'approche est modulaire, elle se compose de trois grandes phases : Une segmentation et identification thématique, une délégation et enfin un filtrage. Le résumé résultant de notre méthode vient s'ajouter aux données existantes au préalable dans la BD. A des fins de validation, nous avons bâti un prototype qui manipule et enrichi le contenu descriptif d'une base de données géographiques.

Références

Allan, J., J. Carbonell, G. Doddington, J. Yamron, et Y. Yang (1998). *Topic Detection and Tracking Pilot Study: Final Report*. Proceedings of the DARPA Broadcast News Transcription and Understanding Workshop, Morgan Kaufmann, San Francisco.

Briot, J.P. et Y. Demazeau, (2001). *Principes et architecture des systèmes multi-agents*. Edition Hermes Science Publications, Paris, France, 266 p.

Elhadad, M. et J. Robin (1996). *An overview of SURGE: A re-usable comprehensive syntactic realization component*. Proceedings *of the 8th International* Workshop *on Natural Language generation (INLG'96)*, Brighton, UK.

Faïz, S. (1999). *Systèmes d'Informations Géographiques : Information Qualité et Data mining*. Editions C.L.E, 362 p.

Faïz, S. et K. Mahmoudi (2005). *Semantic Enrichement of Geohraphical Databases*. Encyclopedia of database tehnologies and applications. Idea group reference Hershey, London. Melbourne. Singapore, 587-592.

Ferber, J. (1997). *Les Systèmes Multi-Agents vers une intelligence collective.* Edition InterEditions, 522 p.

Gees, C.S., T. Strzalkowski, G.B. Wise, et A. Bagga (2000). Evaluating Summaries for Multiple Documents in an Interactive Environment. General Electric, Corporate R&D, United Stated.

Goldstein, J., V. Mittal, J. Carbonell, et M. Kantrowitz (2000). Multi-document summarization by sentence extraction. In Proceedings of the ANLP/NAACL-2000 Workshop on Automatic Summarization, pages 40--48, Seattle, WA, May 2000.

Harabagiu, S., et L. Finley (2002). Generating single and multi document summaries with GISTEXTER. In Proceedings of the Workshop on Multi-Document Summarization Evaluation of the 2nd Document Understanding Conference at the 4Oth Meeting of the Association for Computational Linguistics, Philadelphia, PA, July.

Hearst, M.A. (1997). *TextTiling: Segmenting Text into Multi-paragraph Subtopic Passages*. Computational linguistics, vol. 23, No. 1, pp. 33-46.

Kraaij, W., M. Spitters, et A. Hulth (2002). *Headline extraction based on a combination of uni-and multidocument summarization techniques*. In Proceedings of the ACL workshop on Automatic Summarization/Document Understanding Conference (DUC 2002), June.

Lin, C. Y. et E.H. Hovy (2002). *From Single to Multi-document Summarization: A Prototype System and its Evaluation.* In Proceedings of the Association for Computational Linguistics (ACL) conference. Philadelphia, PA.

Mahmoudi, K. et K. Ghédira (2000). Distributed Rescheduling for the Workforce Management Dynamic Aspect. 3rd Ibero American Workshop On Distributed Artificial Intelligence And Multi-agent Systems, Atibaia, Sao Paulo, Brazil.

Mani, I. et E. Bloedorn (1999). Summarizing Similarities and Differences Among Related Documents. *Information Retrieval, Vol. 1, No. 1, pp. 35-67.*

Mann, W. C. et S.A. Thompson (1988). Rhetorical structure theory: Toward a functional theory of text organization. *An Interdisciplinary Journal for the Study of Text 8(2):243-281.*

Miller, G. (1990). Wordnet: An on-line lexical database. *International Journal of Lexicography (special issue), 3(4):235-312.*

Marcu, D. (1999). *Discourse Trees are good Indicators of Importance in Text.* Mani and Maybury editors, advances in Automatic Text Summarization, p. 123-136, the MIT Press.

McKeown, K., J. Klavens, V. Hatzivassiloglou, R. Barzilay, et E. Eskin (1999). *Towards Multidocument Summarization by Reformulation: Progress and Prospects.* AAAI/IAAA, 1999.

Radev, D.R., H. Jing, et M. Budzikowska (2000). *Centroid-based Summarization of multiple Documents: sentence extraction, utility-based evaluation, and user Studies*. In ANLP/NAACL Workshop on Automatic Summarization. Seattle, p. 21-19, April.

Summary

A Geographic Information System (GIS) is a tool that handles Geographic Databases (GDB). The use of the GIS are multiple and concerns many domains. In fact, in many situations we have to add complementary data to be able to make adequate decisions. In this context, we have proposed a method to generate automatic summaries for multiple documents to extract the essentiel of the information from an on-line corpus to enrich the descriptive component of the GDB. The method is a distributed resolution among a set of autonomous agents to lead jointly to an optimal summary. This distribution is intended to speed up the summary generation while holding the main issues inherent to the problem. Our approach is modular. It consists of three main steps, namely, segmentation and theme identificaion, delegation and text filtering.