UN SYSTEME DE SYNTHESE AUTOMATIQUE DE DOCUMENT TEXTUELS BASE SUR LA NOTION D'ABSTRACTION

NOUSSI Roger
KOUSSOUBE Souleymane

Institut Africain d'Informatique
BP 2263
Libreville - GABON -

RESUME

Cet article présente une méthode de synthèse automatique de documents enregistrés sur support électronique. Le système proposé a pour objectif de mettre à la disposition d'un utilisateur un ensemble de moyens (les **opérations de synthèse**) qui lui permettent d'assimiler de façon personnalisée le contenu d'un document.

La méthode utilisée repose sur une **représentation sémantique** adéquate du document à étudier. Cette représentation intègre aussi bien les concepts inhérents au thème traité par le document que l'expertise et les connaissances opératoires nécessaires pour manipuler intelligemment le document. Les opérations de base offertes par ce système sont spécifiées à l'aide de la **notion d'abstraction**. Elle permettent de construire une représentation virtuelle du document de base adaptée aux contraintes d'acquisition, aux objectifs et au profil de l'utilisateur. Le document de base est enregistré sous forme d'hypertexte.

Mots clés :Opérations de synthèse, Représentation sémantique, Abstraction, système expert, représentation centrée objet.

1 - Introduction

L'outil informatique a jusqu'à une période très récente été utilisé dans traitement de l'information textuelle à travers le traitement de texte. Les fonctionnalités de ces systèmes portent essentiellement sur la saisie et l'édition des documents. Le volume de plus en plus croissant de documents aujourd'hui disponibles sur support électronique et le développement de la technologie multimédia tendent à faire évoluer la problématique des systèmes qui utilisent ces documents. On assiste à l'émergence de systèmes capables de prendre en compte le contenu ou le sens d'un texte pour proposer une aide à l'édition ou à l'assimilation [MEUN 94], [GAST 94], [DIVA 85], [HOCH 94]. Les fonctions offertes par de tels systèmes ne sont plus simplement de nature statique telle que afficher, repérer, archiver etc..., mais présentent de plus en plus un caractère dynamique tel que résumer, restructurer, commenter.

Cet article décrit un système de synthèse de document dont l'objectif est de proposer à l'utilisateur, un ensemble de fonctions lui permettant d'assimiler de façon personnalisée un document. Ces fonctions s'appuient sur une expertise profonde du thème développé dans le texte pour assister

l'effort qu'il est prêt à consentir pour l'assimiler et au volume d'informations qu'il peut utiliser sans se sentir submergé. Pour ces différentes catégories d'utilisateurs, le système doit proposer les chapitres appropriés, les documents annexes ou un résumé des concepts pertinents.

Les opérateurs utilisés par ce système pour élaborer la synthèse procèdent par le calcul d'une abstraction [GIOR 90], [PLAI 81], [TENE 87] de la représentation sémantique du texte initial. Cette abstraction est ensuite éditée de façon appropriées pour produire la version synthétique présenté à l'utilisateur. Dans ce document, après avoir présenté l'intérêt et la forme générale de la représentation sémantique d'un document, nous montrons comment la notion d'abstraction peut être utilisée pour formaliser l'ensemble des opérateurs du système.

2 - Représentation sémantique d'un document

Un document est généralement découvert à travers sa forme "typographique" (taille des pages, police, marge ...). c'est en général l'aspect qui intéresse les personnes qui utilisent le document sans s'intéresser à son contenu. certains utilisateurs peuvent aller au delà de la forme typographique et ne s'intéresser qu'aux considérations syntaxiques qui portent sur le respect des conventions d'usage dans une langue donnée. Ces considérations restent encore à un niveau général d'utilisation de la langue et ne concernent pas le message ou le contenu du document. Le problème de synthèse tel que nous l'envisageons ici consiste à modifier la forme ou même le contenu d'un document sans fausser le message qu'il est sensé véhiculer, mais en l'adaptant à l'utilisateur. La validité d'un opérateur de synthèse repose sur son respect de la sémantique du document. La préservation de l'intégrité du message malgré des modifications de forme et même de contenu doit forcément tenir compte des conditions sociales d'acquisition propre à chaque utilisateur. dans ce paragraphe, nous présentons les principaux constituants d'une bonne représentation sémantique en commençant par la notion de sous-language [KITT 78] [KITT 89] dont la caractérisation (informelle) nous offre un cadre méthodologique pour la capture de la sémantique d'un document.

2.1 - Notion de sous-language [KITT 89]

La notion de sous-language s'appuie sur les caractérisation (informelle) suivantes :

1) domaine de référence restreint : l'ensemble des objets et des relations auxquelles se réfèrent les expressions linguistiques est restreint.
2) Finalité restreinte : les échanges linguistiques sont orientés vers certains buts.
3) Communauté d'usagers restreinte : cette communauté est composée d'usagers partageant des connaissances spécialisées et au sein de laquelle l'expression est fortement entachée d'usages.
4) Mode de communication restreint : l'expression est assujettie à des contraintes matérielles ou techniques telles que la limitation du nombre de ligne pour les annonces.

Par rapport à la problématique de la synthèse, un document traite généralement d'un sujet relevant d'un domaine précis. Le premier point de la caractérisation des sous-languages est toujours respecté dès qu'on a choisit un document. Les trois autres points portent sur le profil de l'utilisateur et font référence à sa culture sur le sujet traité (point 3), son projet de lecture (point 2) et ses contraintes d'acquisition (point 4). Ainsi, dès qu'on associe un document au modèle de l'utilisateur qui cherche à l'utiliser, on se retrouve dans le contexte d'un sous-language. Par ailleurs, il apparaît que les caractéristiques sur lesquelles se fondent un sous-language contribuent toutes à restreindre et à préciser son univers sémantique et de ce fait, à en faciliter la compréhension.

2.2 - Compréhension du sous-language des petites annonces immobilières

Ce paragraphe illustre à travers l'exemple des petites annonces immobilières l'ampleur des variations de formes que peut subir un discours en conservant le message véhiculé. Il s'appuie sur la compréhension des petites annonces immobilières à des fins de mise en correspondance, telle qu'elle est réalisée dans le système expert HAVANE [COUR 87]. Cette étude est en partie basée sur l'analyse du corpus linguistique des petites annonces qui se révèlent comme répondant parfaitement à la notion de sous-language. Ceci leur confère une universalité, un naturel et une souplesse favorisant leur utilisation et leur compréhension malgré la diversité des styles et l'hétérogénéité de la communauté des usagers. Pour illustrer ce fait, considérons le texte suivant :

Je cherche à louer un appartement de type 5 ou 6 dont le loyer est de 3000 francs au maximum. Cet appartement devra être situé dans la zone nord de Rennes.

Ce texte, écrit en respectant au mieux les exigences grammaticales du Français apparaît généralement sous une forme différente, contractée et intuitive dans la rubrique annonce des journaux. On peut par exemple trouver les formulations suivantes :

Cherche à louer appartement T5 ou T6, 3000f maximum, rennes nord ou
Cherche Appt T5-6 3000F max rennes nord. ou encore
Cherche rennes nord, T5-6 3000f max.

Cette forme contractée est induite par les contraintes de tarification des journaux qui publient l'annonce. Cependant, bien que les différents usagers n'aient pas de convention de formulation, ils arrivent généralement à se comprendre. En effet, le seul fait de retrouver dans la rubrique "annonce immobilière" limite l'ensemble des ***concepts significatifs*** à ceux qui traitent de l'immobilier. Par ailleurs, on sait aussi que les expression serviront à formuler des offres ou des demandes. En partant de ces deux hypothèses, on arrive à identifier les concepts significatifs dans l'expression de l'annonce et à construire la sémantique de celle-ci malgré la présence d'éléments non significatifs considérés comme du "bruit". Ces hypothèses permettent également de compléter certaines annonces (incomplètes) et à éviter la plupart des contraintes syntaxiques qu'aurait imposées l'usage d'un Français correct pour assurer la compréhension de l'annonce. ces hypothèses qui portent sur l'univers sémantique, permettent généralement d'étoffer un texte d'annonce "dégarni" et serviront dans la synthèse à élaguer puis à reformuler le texte de base tout en garantissant la teneur du message. L'exemple des petites annonces permet d'illustrer de façon presque caricaturale combien la maîtrise de l'univers sémantique permet de transmettre le message en s'affranchissant de nombreuses contraintes de l'expression.

2.3 - Modèle de compréhension d'une annonce

Le modèle de compréhension d'une annonce immobilière dans Havane est basé sur la constitution d'une liste exhaustive de concepts considérés comme représentatifs du domaine de l'immobilier. La structure syntaxico-sémantique d'un tel constituant est spécifié par une grammaire hors contexte appelée dans Havane grammaire d'îlot. la compréhension d'une annonce est basée sur la construction d'une arborescence des concepts reconnus dans le texte de l'annonce, sur laquelle vient s'articuler localement l'expertise utilisée dans la compréhension et la mise en correspondance. La spécification de cette arborescence et de l'expertise associée est faite à l'aide d'une grammaire d'attributs appelée dans Havane grammaire d'archipel. Le calcul d'attributs y est en partie exprimé à l'aide des règles de production.

2.4 - Représentation du document de base.

En observant le cas des petites annonces immobilières, on constate que la représentation sémantique est basée sur le connaissance des éléments suivants :

- les concepts significatifs du domaine dont traite le document
- les relations existant entre ces concepts
- l'expertise permettant d'opérer des transformations sur ces concepts ou sur le document.

[redacted]

2.2 - Les relations entre les concepts

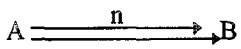
Nous avons relevé dans le paragraphe 2.4 que les concepts ne coexistent pas dans le système de façon indépendante. Dans le cas du document TRPI les relations identifiées sont : les relations d'utilisation, les relations de précédence, et les relations d'héritage.

3.2.1 - Les relations d'utilisation

Dans l'esprit du document TRPI, la présentation de l'ordinateur et de la programmation fait partie des préliminaires. Il y aura donc une relation d'utilisation entre d'une part les préliminaires et d'autre part les concepts *ordinateur et programmation.*

3.2.2 - Les relations de précédence

Les relations de précédence décrivent la culture nécessaire pour aborder un concept. ce sont des relations valuées. La valeur attribuée à une telle relation indique le niveau de culture nécessaire pour l'aborder. Ainsi, la notation.



traduit les situations suivantes :
pour aborder le concept B, si n=0 il n'est pas nécessaire d'être informé sur A
si n=1 il faut avoir au moins lu A avant d'aborder B
si n=2 il faut avoir une bonne pratique de A avant d'aborder B
si n=3 il faut être un spécialiste de A pour aborder B

3.2.4 - Les relations d'héritage

Les relations d'héritage sont de type généralisation et désignent les relations sorte de.
On dira par exemple que algorithmique est une sorte de technique de résolution

3.3 - Les connaissances évaluatoires

La description d'un concept comporte deux parties : la structure du concept (attribut dans la terminologie de la programmation objet) et les connaissances évaluatoires utilisables par cet objet (méthodes). En plus de ces méthodes spécifiques, chaque concept aura les méthodes suivantes :

- les méthodes d'affichages,
- les méthodes de sélection parmis ses composants,
- les méthodes de transformation de ses composants,

etc...

3.4 - Les connaissances expertes

Les connaissances expertes sont des connaissances qui ne sont pas liées à la structure du document mais qui portent sur une bonne pratique du sujet traité par le document et la connaissance des utilisateurs. Les systèmes les utilise pour ajuster la synthèse.

Exemple de connaissance experte

1°) Si l'utilisateur n'exerce pas un métier lié à l'informatique, lui faire une synthèse sur 5 pages.
2°) Si l'utilisateur a déjà suivi avec succès un cours de programmation, procéder comme suit:

- ne pas introduire les préliminaires;
- traiter sur une demie page au plus, l'approche algorithmique
- étudier en profondeur le développement des autres approches.

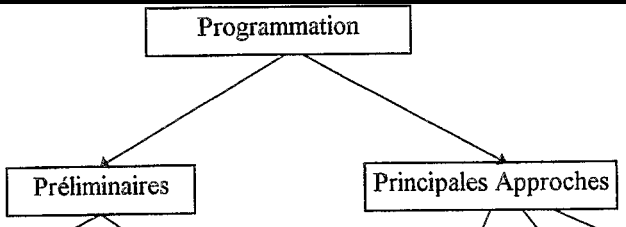
3°) Si l'utilisateur est un enseignant cherchant à constituer un plan de cours, développer tous les concepts avec au plus trois lignes par concept.

4. - L'opération de synthèse

L'opération de synthèse s'effectue en deux temps :
1°) l'extraction d'un sous-graphe correspondant à la représentation sémantique de la synthèse.
2°) l'édition de chaque noeud de la représentation sémantique pour obtenir le texte synthétisé.

Chacune des deux phases utilise des éléments d'expertise portant pour la plupart sur le modèle de l'utilisateur. Le modèle de l'utilisateur est donc l'une des données fondamentales utilisée dans la

En partant de ce modèle, on peut limiter les préliminaires à une présentation succincte des différentes étapes de résolution d'un problème. Parmi les différentes approches, on pourra omettre l'approche algorithmique, considérée comme déjà assimilée. Cette élagage sera itéré à tous les niveaux de l'arbre sémantique. Pour l'exemple introduit au début de ce chapitre, le résultat de cet élagage se présente



5.1. 1 - Synthèse sur un ensemble de clauses

Le précurseur de la formalisation de la notion d'abstraction est D Plaisted. Le but des recherches de

5.1.2 - L'abstraction dans un système de représentation des connaissances centrée objet
[PEUT 92] [ROBI 92]

soit n classes C telles que
- pour tout i {1, ..., n-1} Ci+1 est une sous-classe de Ci
- les instances de la classe Cn sont définies par les attributs des classes C1, C2
des attributs propre à Cn les attributs propre à la classe Ci sont notés A
Soit I une instance de la classe Cn. Elle peut être dénotée par le terme

ou en abrégé
.....
l'attribut A

définition
Une abstraction au sein d'une RCO est une règle de la forme
C:Iav où
Iav décrit l'instance ou l'ensemble des instances initial
Iav est soit un littéral, soit une instance soit un ensemble d'instances
C décrit les relations entre les attributs des instances abstrait et ceux des instances initiales.

Les différentes formes d'abstraction dans le cas d'un héritage multiple

1) la projection d'attribut

Pi: Cette abstraction correspond à l'oubli du ième attribut de Cn

2) la généralisation de classe
Cette abstraction permet d'utiliser une instance d'une classe donnée comme une instance de sa super classe.

3) la généralisation par élément significatif
Cette abstraction s'applique à un ensemble d'instances et permet d'isoler une instance qui possède une propriété qui la distingue des autres instances de l'ensemble considéré.

4) la synthèse simple
Cette abstraction permet de synthétiser une instance d'une classe donnée en une instance d'une autre classe ou en un littéral.

5) la synthèse globale
Sg : ul =

A partir d'un ensemble d'instances d'une classe donnée, cette abstraction crée une instance d'une autre classe ou une valeur littérale.

Les abstractions composées

Les abstractions simples applicables à une instance ou à un ensemble d'instances sont de deux types :
type 1 : une abstraction entière simple de l'instance I (respectivement de l'ensemble d'instances {I} est l'application au terme t représentant I (respectivement {I} de l'une des règles d'abstractions élémentaires.

type 2 : une abstraction partielle simple de l'instance I (respectivement de l'ensemble d'instances {I} est l'application au sous-terme ti représentant la valeur vi de l'attribut Ai de l'une des règles d'abstraction élémentaires.

6 - Utilisation des abstractions dans la formulation des connaissances expertes

Les différentes abstractions utilisées dans le système seront rattachées aux noeuds (concepts) du graphe de la représentation sémantique; la définition de ces abstractions résulte d'une analyse profonde des différents concepts en relation avec les différentes utilisations possibles.
L'expertise permettant au système de réaliser la synthèse du document est exprimée sous forme de règles de production.

De façon générale, la partie condition d'une règle de production spécifie un profil utilisateur et le concept à développer tandis que la partie action précise l'abstraction (attachée au noeud) qu'il faut appliquer pour synthétiser la présentation voulue :

modèle (utilisateur) ⟶ **abstraction** (concept courant)

Pour illustrer cette formulation, nous allons développer l'exemple ci-dessous qui a déjà été introduit au paragraphe 3.4

1°) Si l'utilisateur n'exerce pas un métier lié à l'informatique, lui faire une synthèse sur 5 pages au plus

2°) Si l'utilisateur a déjà suivi avec succès un cours de programmation, procéder de la façon suivante :
- *ne pas introduire les préliminaires*
- *traiter sur une demi page au plus, l'approche algorithmique*
- *étudier en profondeur le développement des autres approches.*

3°) Si l'utilisateur est un enseignant cherchant à constituer un plan de cours, développer tous les concepts avec au plus trois lignes par concept.

Pour des raisons de simplicité, nous nous limitons dans le développement de cet exemple aux relations de compositions qui existent entre les concepts telles que décrites sur la figure du paragraphe 3.2.

La règle 1 fait intervenir une synthèse simple (A1) appliquée à l'unique instance du concept programmation :

A1 : = programmation résume (5 pages) :
Programmation Document

x représentent les valeurs des attributs de l'intance. La fonction résume (méthode de la classe programmation) ramène le résumé de l'instance sous le forme d'un document dont la longueur dépend de l'argument transmis à cette fonction. Ceci correspond à une opération de synthèse simple.

La deuxième règle fait intervenir des abstractions composées. Les abstractions simples impliquées sont les suivantes :

soit I1 = Programmation(prelim,princ_aproch(algo,logiq,fonct))

A3 : Programmation(princ_aproch(algo,logiq,fonc)
Programmation(princ_aproch(logiq,fonct))

A3 est une g-abstraction partielle portant sur le constituant princ-approch. Elle permet d'oublier le premier attribut de ce constituant.

A4 = A3°A2:
A4:I1 Programmation(princ_approch(logiq,fonct))

A4 permet de se focaliser uniquement sur les approches logique et fonctionnelle de la programmation.
A5 : = algo.résume(0,5page):I1 document ( )

A5 est une s-abstraction partielle (c'est une application d'une s-abstraction au constituant algo de l'instance I1 qui permet de faire le résumé en une demi page de l'approche algorithmique.

Les règles 1) et 2) peuvent alos s'écrire :

Règle n°1

Si A_suivi(u,'programmation')
présenter('programmation',u)
instance(p,programmation)
ALORS
APPLY(p,A5), APPLY(p,A4)

Règle n°2

Si présenter ('programmation',u)
Non_Exerc_en(u,'informatique'),
instance(p,programmation)
ALORS APPLY(P,A1)

Conclusion

Nous avons proposé dans cet article une méthode de synthèse de document basée sur une représentation adéquate et enrichie du document. La constitution de cette représentation sémantique nécessite la participation d'un expert du domaine dont traite le document. Par ailleurs, la saisie du document doit être précédée par la saisie de son modèle sémantique, et de l'expertise (sous forme de règles de production) nécessaire pour réaliser la synthèse. la saisie du document lui-même se fait thème

BIBLIOGRAPHIE

[BACK78] J. BACKUS. *Can programming be liberated from the von Neuman style ? A functionnal approch and its algebra of program.* Com. of the ACM aug; 1978 vol. 21 num. 8

[COUR 87] M. COURANT. *La compréhension de petites annonces dans le système havane.* Thèse de doctorat de l'université de Rennes 1, nov 1987

[DIVA 85] M. DIVAY. *Un système expert de traitement de textes écrits* Congrès AFCET, intelligence artificielle et reconnaissance des formes, Grenoble, mars 1985.

[GARG 87] P. GARG *Abstraction mechanism in hypertext.* hypertext'87

[GAST 94] S.B. GASTALDY, L. GIROUX, D.LANTEIGNE, C. DAVID *Les produits et processus congnitifs de l'indexation* in ICO vol. 6 num 1 et 2 printemp 1994

[GIOR 90]] A.GIORDANA, G PERRETO D. ROVERSO L. SAITTA. *Abstraction: an alternative view of concept acquisition.* Elsevier science publishing 1990

[HOCH 94] J.C. HOCHON, F. EVRARD. *Lecture professionnelle et gestion de document textuels* in ICO vol. 6 num 1 et 2 printemp 1994

[KITT 78] R. KITTEREDGE. *Textual cohésion within sublanguage:implication for automatic analysis and synthesis.* 7th international conference on computation linguistic

[KITT 89] R. KITTEREDGE The signifiance of sublanguage for automatic translation in machine learning ed Serguei Nirensburg studies in natural langage processing 1989

[KNOB 89] A. KNOBOCK. *A theory of abstraction for hierarchical planning in change of representation and inductive bias.* Kluwer Publ. Co 1989

[LEVE 93] Level 5 object:reference guide.information builders Inc 1993

[MEUN 94] J.G.MEUNIER,S.B.GASTALDY L.C. PAQUIN *La gestion et l'analyse des textes par ordinateur;leur spécificité dans le traitement de l'information.*in ICO vol.1 et 2 printemp 1994

[NOUS 89] R.NOUSSI. *Un modèle objectal pour la synthèse de dossiers médicaux* Thèse de doctorat de l'université de Rennes1,déc.1989

[PEUT 92] S. LE PEUTREC *Mécanisme d'abstraction dans une représentation de connaissances centrée* rapport interne IRISA 1992

[PLAI 81] D. PLAISTED, *Theorem proving with abstraction.*artificial intelligence 16 1981

[TENE 87] Josh TENENBERG. *Preserving consistency accross abstraction mappings.* IJCAI 87

[ROBI 92] S.ROBIN,M. COURANT. *Mécanisme d'abstraction dans une représentation de connaissances objectales* rapport interne IRISA 1992