



IFSIC
Université de Rennes 1
Campus de Beaulieu
35042 RENNES CEDEX

Alcatel Alsthom Recherche
Route de Nozay
91460 MARCOUSSIS

Filtrage Intelligent de documents textuels

Sous la responsabilité de :
Hélène Bernard (Alcatel)
Annie Morin (IFSIC)

Frédéric Le Mouël
DIIC 3ème année, filière LSI
1996-1997

ALCATEL ALSTHOM RECHERCHE

Renseignement Documentaire

Titre : **Filtrage intelligent de documents textuels**

Résumé : Ce document présente la reprise et les améliorations de la partie Filtrage de l'étude d'EXtraction d'Informations STructurées (EXIST). Ces améliorations consistent principalement à connaître l'utilisateur, en particulier son profil, à le conseiller de manière intelligente pour classer ses documents, à apprendre interactivement des choix que celui-ci fait, à le décharger d'une partie de son classement.

This document presents the re-using and the improvements of the filtering part of the EXIST Study (EXtraction of STructured Information). These improvements consist in the user recognition, and especially his profile, consist in giving user advice to classify his documents in the right topic, also consist in learning interactively from user's choice, and finally in releasing user from a part of this classification.

Mots clés : Filtrage, Intelligence Artificielle, Profil Utilisateur, Conseils, Apprentissage, Interactivité, Autonomie.

Filtering, Artificial Intelligence, User's profile, Advice, Learning, Interactivity, Autonomous.

Référence AAR : UAR/C/97/0280	Version : 1.2	Auteurs : F. Le Mouël
Référence client :	Date : 25/08/1997	Visa :

Nombre de pages du document : 84

APPROBATION

Fonction	Resp. d'étude	Chef de groupe	Chef d'unité
Visa			
Nom	H. Bernard	M. Mautref	E. Daclin

FLM

VERSIONS

1.2	25/08/1997	F. Le Mouël	
Version	Date	Auteurs	Approbation

Documents référencés :

Co-auteurs :

Resp. d'étude : H. Bernard

Table des matières

1 Introduction **9**
1.1 Objet du stage 9
1.2 Contenu du document 9
1.3 Les documents de référence 10

2 Présentation de l'entreprise Alcatel Alsthom **11**
2.1 Alcatel Alsthom 11
2.1.1 Secteur Télécommunications 12
2.1.2 Secteur Câbles et Composants 12
2.1.3 Secteur Engineering et Systèmes 13
2.1.4 Secteur GEC Alsthom 13
2.2 Alcatel Alsthom Recherche 14
2.2.1 Recherche Informatique 14
2.2.2 Recherche Systèmes optiques 15
2.2.3 Recherche Énergie 15
2.2.4 Recherche Matériaux et procédés associés 16

3 Présentation de l'étude EXIST **17**
3.1 Objet de l'étude 17
3.2 Contexte de l'étude 18
3.2.1 L'individualisation de la société 18
3.2.2 Accroissement du volume d'informations disponibles 18
3.2.3 Positionnement d'Alcatel 20
3.2.4 Identification des besoins 21

4 État d'avancement du projet **23**
4.1 Rappel des contraintes générales 23
4.1.1 Contraintes de réalisation 23
4.1.2 Contraintes de performances 24
4.2 Architecture visée 24
4.3 Architecture système globale actuelle 25

4.4 Rappel des améliorations déjà effectuées 27
4.5 Améliorations et nouvelles fonctionnalités 28

5 État de l'Art **30**
5.1 Agents intelligents 30
5.1.1 Origines 30
5.1.2 Typologie 31
5.1.3 Agents d'interface 33
5.2 Filtrage de documents 34
5.2.1 Origines 34
5.2.2 Cas d'études 36
5.2.3 Théorie de conception 37

6 Conception **38**
6.1 Conception générale de l'agent filtreur 38
6.2 Modélisations 40
6.2.1 Modélisation d'un document 40
6.2.2 Modélisation d'un utilisateur 40
6.2.3 Modélisation du profil utilisateur 41
6.2.4 Modélisation d'un centre d'intérêt 41
6.3 Comparaison d'un document avec un profil utilisateur 42
6.3.1 *The Sub-String Indexing* 42
6.3.2 *The Cosine Measure* 42
6.3.3 Métrique inspirée de celle de Salton 43
6.4 Reconnaissance de l'utilisateur 44
6.5 Conseils de classement d'un document 44
6.6 Interactivité avec l'utilisateur 44
6.7 Apprentissage 45
6.7.1 Apprentissage vectoriel simple 45
6.7.2 Apprentissage statistique 45
6.7.3 Apprentissage génétique 46
6.8 Autonomie 48

7 Résultats **50**
7.1 Apprentissage vectoriel :
Conseils / Nombre de textes appris 51
7.1.1 Texte t_1 : pertinent 51
7.1.2 Texte t_2 : non pertinent 52
7.1.3 Conclusion 53
7.2 Apprentissage statistique :
Conseils / Nombre de textes appris 54

7.2.1 Texte t_1 : pertinent 54
7.2.2 Texte t_2 : non pertinent 55
7.2.3 Conclusion 56
7.3 Apprentissage génétique :
Conseils / Nombre de textes appris 57
7.3.1 Texte t_1 : pertinent 57
7.3.2 Texte t_2 : non pertinent 58
7.3.3 Conclusion 59
7.4 Apprentissages :
Performance en temps / Nombre de textes 60
7.5 Conseils / Nombre de mots dans le Profil P 61
7.5.1 Texte t_1 : pertinent 61
7.5.2 Texte t_2 : non pertinent 63
7.5.3 Conclusion 64
7.6 Conseils / Nombre de mots dans les textes 65
7.6.1 Texte t_1 : pertinent 65
7.6.2 Texte t_2 : non pertinent 67
7.6.3 Conclusion 68
7.7 Conseils :
Performance en temps / Nombre de mots 69
7.8 Conclusions des résultats 70

8 Conclusion **71**

A Exemple d'application d'un algorithme génétique **73**
A.1 Tirage aléatoire 73
A.2 Première étape : la sélection 74
A.3 Seconde étape : la recombinaison 75
A.4 Troisième étape : la mutation 76

Table des figures

4.1 Architecture globale du prototype 96 26
4.2 Nouvelle partie Filtrage 29

5.1 Typologie d'un agent 32
5.2 Agent d'interface dans le cas du filtrage 33
5.3 Filtrage, inverse de la Recherche d'informations 35

6.1 Diagramme des objets 39
6.2 Définition du degré d'autonomie de notre agent 49

7.1 Apprentissage vectoriel : texte pertinent 51
7.2 Apprentissage vectoriel : texte non pertinent 52
7.3 Apprentissage statistique : texte pertinent 54
7.4 Apprentissage statistique : texte non pertinent 55
7.5 Apprentissage génétique : texte pertinent 57
7.6 Apprentissage Génétique : texte non pertinent 58
7.7 Performances des différents apprentissages 60
7.8 Conseils pour le texte t_1 / Nombre de mots pris dans le profil P . . . 61
7.9 Zoom sur le début du graphique précédent 62
7.10 Conseils pour le texte t_2 / Nombre de mots pris dans le profil P . . . 63
7.11 Zoom sur le début du graphique précédent 64
7.12 Conseils pour le texte t_1 / Nombre de mots pris dans le texte t_1 . . . 65
7.13 Zoom sur le début du graphique précédent 66
7.14 Conseils pour le texte t_2 / Nombre de mots pris dans le texte t_2 . . . 67
7.15 Zoom sur le début du graphique précédent 68
7.16 Durée de calcul des différents conseils 69

Liste des tableaux

5.1 Origine des agents intelligents 31
5.2 Différences entre la recherche et le filtrage d'informations 35

6.1 Application d'un algorithme génétique dans le cas du filtrage 47

Remerciements

Je voudrais tout d'abord remercier Eric Colaviti pour sa sympathie, et pour l'aide compétente qu'il m'a apportée.

Je tiens aussi à remercier Marc Mautref pour m'avoir accueilli au sein de l'*Unité Automatismes et Systèmes de Renseignements*, groupe *Multimédia*, ainsi que Jean-Paul Rossazza et Hélène Bernard pour m'avoir intégré dans l'étude Renseignement Documentaire.

Je tiens enfin à souligner les très bonnes conditions de travail d'Alcatel Alsthom Recherche au niveau environnement, horaires, avantages liés au comité d'entreprise.

Chapitre 1

Introduction

1.1 Objet du stage

C'est à Alcatel Alsthom Recherche, dans l'*Unité Automatismes et Systèmes de Renseignements*, groupe *Multimédia* que j'ai effectué le stage nécessaire à la validation de mon Diplôme d'Ingénieur en Informatique et téléCommunication (DIIC).

Le travail, à effectuer au cours de mon stage, consistait à reprendre la partie filtrage de l'étude d'EXtraction d'Informations STructurées (EXIST). Cette partie n'avait pas été approfondie dans l'étude, l'accent ayant été mis sur l'analyse de documents. J'ai donc eu à faire un état de l'art pour déterminer les solutions existantes, et les plus adaptées à une mise en œuvre dans le cadre cette l'étude.

En premier lieu, je vais présenter l'entreprise Alcatel Alsthom, et plus particulièrement son centre de recherche de Marcoussis. Ensuite, je détaillerai l'étude EXIST, avec ce qui est fait, ainsi que les objectifs pour mon stage. Enfin, j'exposerai les différentes étapes et choix effectués au cours de mon stage : état de l'art, choix de conception, mises en œuvre ...

1.2 Contenu du document

Ce document présente les résultats de l'analyse et de la conception de la partie filtrage de la maquette logicielle de l'étude EXIST. Il comporte 7 chapitres en plus du présent chapitre d'introduction.

Le chapitre 2 présente l'entreprise Alcatel Alsthom et son centre de recherche de Marcoussis.

Le chapitre 3 présente l'étude EXIST, nature, objet, les objectifs fixés.

Le chapitre 4 présente l'état d'avancement de l'étude EXIST, contraintes, architecture, fonctionnalités, améliorations.

Le chapitre 5 présente un état de l'art sur les agents et le filtrage de documents.

Le chapitre 6 présente la conception de l'agent de filtrage.

Le chapitre 7 présente les résultats de notre agent.

Le chapitre 8 conclut ce rapport.

1.3 Les documents de référence

[1]	**Dossier d'Analyse des Besoins en Renseignement Documentaire**	UAR/RT/95/119/V1.2
[2]	**Dossier d'État de l'Art**	UAR/RT/95/143/V1.2
[3]	**Rapport complémentaire sur l'État de l'Art**	UAR/RT/95/234/V1.1
[4]	**Dossier de Spécifications Logicielles**	UAR/RT/95/206/V1.1
[5]	**Première spécifications EXIST 96**	UAR/C/96/0110/V1
[6]	**Outil de recherche documentaire avancée sur Internet**	UAR/C/96/0136/V1
[7]	**Deuxièmes spécifications EXIST 96**	UAR/C/96/0334/V1

Chapitre 2

Présentation d'Alcatel Alsthom

2.1 Alcatel Alsthom

Alcatel Alsthom est un fournisseur mondial d'équipements et de systèmes de haute technologie dans les domaines des télécommunications, de l'électronique et de l'électromécanique. Il occupe des positions internationales de tout premier plan dans chacun de ses secteurs.

Pour s'adapter à l'évolution de ses marchés et à leur globalisation, et répondre plus efficacement aux nouvelles exigences de ses clients, Alcatel Alsthom s'est organisé par métiers. Ses activités sont désormais réparties en grands secteurs : Télécommunications, Câbles et Composants, Engineering et Systèmes, GEC Alsthom (rassemblant Énergie et Transport).

Avec 190.600 employés répartis dans le monde, Alcatel Alsthom a réalisé un chiffre d'affaires de 162,1 milliards de francs en 1996. A l'avant-garde des technologies, le Groupe consacre 10,2 % de ses ventes (plus de 20 % dans certains secteurs-clés des télécommunications) à la Recherche et au Développement. Cette stratégie s'accompagne d'une politique de partenariats et d'alliances dans des domaines complémentaires, afin de renforcer les positions du groupe dans certaines activités.

Alcatel Alsthom est présent dans plus de 20 pays à travers des implantations industrielles ou des partenariats locaux et entretient des relations commerciales dans plus de 130 pays. Si deux tiers de son activité sont encore réalisés en Europe, son marché d'origine, le Groupe se développe de façon soutenue en Asie, dans la zone Pacifique ou encore dans les Amériques. Avec des produits performants et des équipes internationales, Alcatel Alsthom est en mesure d'anticiper les besoins du marché et de proposer des solutions innovantes à ses clients partout dans le monde.

2.1.1 Secteur Télécommunications

Alcatel Alsthom, à travers Alcatel Telecom, est l'un des leaders mondiaux des systèmes de télécommunication. Son activité couvre l'ensemble des besoins publics ou privés, de l'opérateur à l'utilisateur final, avec une gamme de produits et services recouvrant les systèmes de réseaux fixes, notamment avec la plus forte base installée au monde, les communications mobiles, la communication d'entreprise, les câbles sous-marins ainsi que les marchés de la radiocommunication, de l'espace et des communications pour satellites et de la défense.

Alcatel Telecom occupe une position de premier plan sur le marché très concurrentiel des communications mobiles cellulaires où son offre couvre aussi bien les commutateurs que les stations de base et les terminaux.

Grâce à ses investissements en recherche et développement, Alcatel Telecom est un des premiers acteurs mondiaux dans les technologies des autoroutes de l'information que sont la commutation large bande ATM et la transmission SDH, appelée SONET aux États-Unis. L'intégration de réseaux et le service aux clients sont également des composantes essentielles de l'offre d'Alcatel Telecom.

2.1.2 Secteur Câbles et Composants

Le secteur Câbles et Composants rassemble les activités manufacturières du Groupe à partir de l'expérience et des positions de leader mondial d'Alcatel Alsthom dans les câbles. A ce titre, il intègre désormais les activités batteries et composants. Largement axé vers les marchés des équipements de télécommunication, d'électronique et d'électrotechnique, il est organisé en cinq divisions : énergie, métallurgie télécommunications, composants et batteries.

Avec Alcatel Cable, Alcatel Alsthom occupe la place de numéro un mondial dans le secteur des câbles. Il produit une gamme complète de câbles en cuivre et fibre optique pour les télécommunications terrestres, sous-marines ou aériennes, ainsi que des câbles d'énergie haute, moyenne et basse tension, des câbles spéciaux, des équipements de connexion ou des antennes d'émission.

Saft, leader mondial des accumulateurs et systèmes de secours d'énergie, offre une gamme complète de solutions adaptées aux marchés des télécommunications, notamment des applications portables, des transports et équipements industriels. Saft joue un rôle clé dans le domaine aéronautique et spatial ainsi que dans celui des batteries pour véhicules électriques.

2.1.3 Secteur Engineering et Systèmes

Engineering et Systèmes est l'entrepreneur ensemblier d'Alcatel Alsthom. Une proportion croissante des services offerts à la clientèle nécessite, en effet, la combinaison de nombreuses compétences allant de la direction de projet au développement de produits, en passant par la planification de réseaux, l'intégration, l'installation, la mise en service, la maintenance et, de plus en plus, l'ingénierie financière. Ce secteur du Groupe, largement axé conduite de projets clés en main, réunit Cegelec, Alcatel Contracting, Alcatel Siette et Sogelerg-Sogreah.

2.1.4 Secteur GEC Alsthom

GEC Alsthom, société détenue à 50/50 avec le britannique GEC, est un leader mondial en production, transmission et distribution d'énergie. GEC Alsthom innove dans des domaines de haute technologie : centrales à cycle combiné, chaudières à lit fluidisé et turbines hydrauliques à vapeur et à gaz. Certaines de ces turbines sont utilisées dans des centrales nucléaires pour lesquelles Framatome (détenue à 44% par Alcatel Alsthom) fournit les réacteurs et les combustibles.

GEC Alsthom exerce aussi une forte activité dans le domaine des transports. Il conçoit et réalise une large gamme de matériels ferroviaires, notamment trains à grande vitesse (comme le TGV français, l'AVE espagnol et le TGV coréen), locomotives électriques et diesels-électriques, rames automotrices, métros et tramways, ainsi que des systèmes ferroviaires de signalisation et d'automatisme.

À travers les Chantiers de l'Atlantique, il fournit des paquebots parmi les plus grands du monde ainsi que des navires de grande complexité, comme les paquebots de croisière ou les méthaniers.

2.2 Alcatel Alsthom Recherche

La recherche à Alcatel Alsthom s'effectue dans quatre centres : Anvers, Madrid, Stuttgart et Marcoussis. Les trois premiers centres font de la recherche pour Alcatel Telecom dans un domaine bien précis (Réseaux & Services à Anvers, Énergie à Madrid, Technologies à Stuttgart).

Le centre d'Alcatel Alsthom Recherche, situé à Marcoussis, est le principal centre de recherche pour les différents secteurs, et travaille avec les autres centres de recherche en partenariat ou comme support. Les axes de recherche d'Alcatel Alsthom Recherche peuvent être découpés en 4 domaines : Informatique, Systèmes optiques, Énergie, Matériaux et procédés associés.

L'*Unité Automatismes et Systèmes de Renseignements*, dans laquelle j'ai effectué mon stage, s'occupe de la recherche informatique, et l'étude EXIST se situe dans le multimédia et dans le renseignement.

2.2.1 Recherche Informatique

Logiciels de télécommunications :

- plates-formes de contrôle temps réel
- gestion de réseaux
- ingénierie des services, architecture TINA
- technologies orientées objets
- multimédia
- sécurité

Logiciels d'automatisme :

- processus automatisés : contrôle, commande, supervision, sûreté de fonctionnement

Logiciels pour l'Espace :

- Espace : ingénierie des systèmes spatiaux, technologies embarquées, supervision du segment spatial, traitement d'images
- Renseignement : exploitation d'image/compression, traitement de données documentaires, aide à la décision
- gestion des réseaux et sécurité des systèmes

2.2.2 Recherche Systèmes optiques

Photonique et électronique :

- transmissions optiques à longue distance
- systèmes à amplification optique
- routage et commutation photonique

Composants optoélectroniques :

- lasers
- amplificateurs optiques
- filtres et multiplexeurs
- commutateurs optiques
- circuits intégrés optoélectroniques

Fibres optiques et composants à fibres :

- fibres dopées erbium, composants photoréfractifs

2.2.3 Recherche Énergie

Génie électrique :

- câbles d'équipement et câbles d'énergie
- supraconductivité
- contrôle-commande
- systèmes d'isolation
- maintenance
- systèmes électroniques

Électronique de puissance :

- architecture/topologie
- compatibilité électromagnétique
- gestion thermique

Électrochimie : sources d'énergie pour les applications portables et industrielles : véhicule électrique, spatial ...

- batteries Nickel-Cadmium, Nickel-hydrures, Lithium-Carbone ...
- supercapacités
- analyse et modélisation des mécanismes
- contrôle et gestion du fonctionnement des batteries

2.2.4 Recherche Matériaux et procédés associés

Polymères

Ablation laser

Acoustique et vibrations

Chapitre 3

Présentation de l'étude EXIST

3.1 Objet de l'étude

L'étude 'EXtraction d'Informations STructurées' (EXIST) a pour but la mise au point d'outils de haut niveau d'aide à l'exploitation intelligente de documents textuels. Ces outils sont définis en vue d'une participation à une offre globale de services pour l'exploitation de données multimédia.

L'étude EXIST concerne l'exploitation intelligente des documents textuels répondant ainsi à un besoin réel des entreprises. En effet, l'explosion du nombre d'ordinateurs, d'utilisateurs connectés à Internet provoque un flux croissant d'informations disponibles. Le besoin d'acquisition, d'exploitation de documents textuels se retrouve dans les banques, les administrations, les entreprises, pour diverses raisons : veille économique, veille financière, veille industrielle, veille technologique. Ce besoin se traduit par la nécessité d'avoir des outils de recherche de haut niveau, c'est à dire des outils capables de trouver de façon pertinente les documents recherchés, capables de classer ces documents en fonction de ce que désire, de ce qui est intéressant pour l'utilisateur.

Actuellement de nombreux moteurs de recherches existent sur Internet (Alta Vista (DEC), Yahoo!, Lycos, Excite, Webcrawler, ...) et offrent leurs services gratuitement. Mais des analystes prévoient qu'à terme ces services seront payants, les fabriquants cherchant actuellement à occuper le terrain et à provoquer une dépendance.

Ces services de recherche textuelle sont encore trop simplistes, l'information recherchée (documents pertinents) est généralement noyée au milieu d'informations inutiles, le bruit (documents non pertinents). Si certains moteurs de recherche (ex :

Yahoo!) commencent à proposer une recherche basée sur le contenu d'une image et non plus sur sa légende textuelle, cette solution d'avenir reste actuellement marginale vue la complexité du problème. L'indexation par le texte est et devrait donc rester, pour un certain nombre d'années encore, le moyen le plus efficace de retrouver des informations.

3.2 Contexte de l'étude

Actuellement, il est possible d'observer de plus en plus de "services Multimédia". Cette recrudescence de services est imputable à la conjonction d'au moins deux grands phénomènes :

un phénomène socioculturel : l'individualisation

un phénomène technique : le développement de la télématique et un volume croissant d'informations disponibles

3.2.1 L'individualisation de la société

L'évolution démographique des sociétés industrielles se caractérise par le vieillissement de la population et la réduction de la taille des ménages. Un autre tendance forte de la société est l'individualisation des consommations et des activités. Les gens ont de plus en plus tendance à mener des activités de manière isolée (baladeur, télévision) ou à rester isolé même pour des activités de groupe (vidéo-phone, téléconférence, dialogue par minitel, télé-travail, télé-enseignement, etc, ...). Ces facteurs entraînent une augmentation des besoins en communications.

3.2.2 Accroissement du volume d'informations disponibles

L'explosion du parc informatique remonte aux années 60 et concerne principalement le monde de l'entreprise, et aussi de grandes administrations comme l'armée. Ces organisations mettent en route d'ambitieux projets comme ARPANET, Xerox PARC. Dès le début des années 80, des analystes prévoyaient l'explosion de la télématique pour les années 1980-2000, et ils ne s'y trompent pas puisqu'au début des années 80, les recherches lancées dans les années 60 arrivent "à maturité" : Internet se retrouve dans les universités mondiales, les laboratoires de recherche, les grosses entreprises ... On assiste bien là à cette explosion de la télématique, à cette

mise à disposition de moyens informatiques à travers des réseaux de télécommunication. C'est la base des services multimédia. Les principaux avantages qu'apporte la télématique sont les suivants :

- une connaissance accrue, c'est-à-dire l'accès rapide à beaucoup plus d'informations pertinentes;
- une puissance de calcul amplifiée, c'est à dire pouvoir faire plus de choses plus complexes;
- une augmentation globale de la productivité des entreprises et des administrations;
- une amélioration de la vie courante et des conditions de travail.

Actuellement, le volume d'information potentiellement disponible, pour le grand public et les applications professionnelles, devient de plus en plus important, du fait de la demande croissante et des développements techniques (fibres optiques, ATM, transmissions satellitaires, ADSL) et notamment du fait de l'ouverture d'Internet au grand public (fournisseurs d'accès). Les informations utilisables sont de natures variées :

- textes structurés ou non,
- images de natures différentes,
- vidéos,
- graphiques,
- sons, ...

Dans ce contexte de mise à disposition massive d'informations par le développement des réseaux mondiaux de télécommunication, un impact essentiel va être le développement d'une industrie logicielle de consultation et d'exploitation de l'information. L'explosion des informations multimédia au travers d'Internet a particulièrement mis en avant la nécessité d'avoir des moteurs de recherche efficaces permettant d'accéder rapidement aux informations pertinentes. Les outils actuellement proposés ne sont pas suffisants parce que trop simplistes. En effet, des études montrent que seulement 5 % du temps de travail fourni sur Internet consiste en l'exploitation d'informations pertinentes, 70 % du temps est perdu en recherche et parcours, 25 % du temps est perdu à lire des informations inutiles. De plus en plus d'outils plus ou

moins "intelligents" sont étudiés, et leurs services sont, pour beaucoup, distribués gratuitement. La stratégie des fabriquants consiste à occuper le terrain, à analyser et créer le besoin, et à créer une dépendance. Ces services, gratuits pour l'instant, sont destinés à devenir payants dans l'avenir.

Parmi les divers types de documents, les documents textuels électroniques sont actuellement une source majeure d'information même si les progrès technologiques permettent de stocker, de transmettre et de consulter de plus en plus facilement d'autres types de documents (images, vidéo, son, ...). Cependant, le texte est à l'heure actuelle et vraisemblablement pour quelques années encore, le média le plus propice à la recherche d'information. Celle-ci a, en effet, atteint un degré de maturité qui la rend, bien qu'encore insuffisante, réellement utilisable en pratique. La recherche d'information relative aux autres média, quoique d'avenir, en est aujourd'hui à ses débuts.

3.2.3 Positionnement d'Alcatel

L'évolution des réseaux de communication va concerner principalement les architectures et les logiciels d'exploitation, avec une plus value importante apportée particulièrement par les services. Le groupe ne peut plus se contenter d'une position de fournisseur d'équipement de télécommunication pour maintenir ses positions, et puis accroître ses parts de marchés. Du fait de la dérégulation et de l'ouverture des marchés, le groupe se positionne en tant qu'*intégrateur* et propose une *offre intégrée et complète*, ce que demandent de plus en plus les clients. Cette demande client concerne en particulier le domaine multimédia. L'offre proposée doit alors être complète, jusqu'aux services terminaux proposés aux clients finals.

L'objectif, concernant les services multimédia, est dans un premier temps :

- d'analyser et de comprendre les besoins dans ce nouveau domaine où les évolutions sont très rapides,
- de se positionner et de conserver une avance par rapport aux concurrents,
- de maîtriser les technologies nécessaires,

le tout afin de conserver le contrôle des partenariats et de maintenir une position forte d'intégrateur.

3.2.4 Identification des besoins

Les besoins identifiés dans le domaine du traitement intelligent de documents textuels sont :

- la recherche de documents dans des bases locales ou au travers d'un réseau d'information; ceci couvre des besoins en :
  - indexation automatique des documents, afin de pouvoir indexer les documents non plus seulement sur les mots qu'ils contiennent (ce qui est limitatif), mais sur les idées, les concepts qu'ils contiennent,
  - langage de requête puissant permettant de rechercher les documents pertinents à partir d'idées et de concepts et non pas qu'avec des mots,
  - facilités de consultation des bases de renseignement documentaire en plusieurs langues. Clairement sur Internet, la majorité des informations est en anglais; l'utilisateur français doit pouvoir effectuer une recherche au travers des documents anglais, même s'il ne sait pas toujours, au début, comment formuler en anglais ce qu'il cherche.
- le filtrage automatique afin de ne retenir automatiquement, parmi tous les documents saisis ou reçus, que ceux qui sont pertinents en fonction du contexte de travail. Ou, à l'opposé, un outil qui sache éliminer les documents "vides", sans intérêt en fonction du contexte. De tels outils de filtrage peuvent constituer la partie "intelligente" des agents (robots) de recherche sur Internet,
- la classification automatique des documents par thèmes reconnus. Ce service est du même genre que le précédent, mais le service dispose ici de plusieurs filtres et de moyen de décider quel filtre est le meilleur,
- le routage intelligent qui permet de diriger automatiquement les documents vers les utilisateurs voulus,
- l'extraction d'information dont le but est de rechercher et d'extraire des informations très précises et structurées (*qui* achète *tant d'actions* de *telle société* à *tel prix* par exemple ...).

D'autres besoins importants mais plus périphériques sont :

- des outils d'analyse et de fusion des informations extraites des documents textuels. Ce besoin est identifié comme important dans le domaine militaire ou civil. La gestion du renseignement de documentation est une fonction critique notamment pour la détection et le suivi de prolifération. L'objectif est alors

de détecter et de suivre la production et la diffusion, dans certains pays, des produits nucléaires, biologiques, chimiques (NBC). Cette détection et ce suivi ne peuvent se faire que par analyse, exploitation et recoupement des informations de renseignement documentaire. Ce besoin apparaît maintenant comme de plus en plus important au niveau des entreprises avec l'émergence du data-mining,

- des outils qui permettent de constituer, tenir à jour et compléter automatiquement les index et les thesaurus utilisés pour indexer et consulter les bases de documents.

Chapitre 4

État d'avancement du projet

4.1 Rappel des contraintes générales

4.1.1 Contraintes de réalisation

Plusieurs contraintes doivent être prises en compte :

matérielle : fonctionnement sur station Sun et sur PC. Le prototype est développé et démontré sur Sun mais doit être portable sur PC pour pouvoir être intégré dans des produits des filiales Alcatel.

logistique : découlant des contraintes matérielles : système UNIX sur Sun et Windows 32 bits (NT) sur PC, utilisation de compilateurs C/C++ adéquats, de logiciels portables Sun/PC.

conception : la conception doit être modulaire et ouverte par la nécessité de s'intégrer dans des plates-formes de travail développées par des filiales, ou pour pouvoir intégrer aisément de nouvelles fonctionnalités.

multilinguisme : la conception doit prévoir que le traitement des documents peut être fait dans des langues diverses. Il faut donc également rester ouvert pour l'intégration de nouveaux analyseurs linguistiques.

intégration : la maquette doit pouvoir être intégrée dans des produits des filiales Alcatel.

4.1.2 Contraintes de performances

Les contraintes de performances sont les suivantes :

le temps de traitement : le système doit supporter un flux documentaire de l'ordre d'un millier de pages par jour (habituellement le flux est de quelques centaines de documents de 2 à 3 pages de texte par jour; il monte à plusieurs milliers par jour lors de périodes de crise).

la réactivité : le prototype visé est un système interactif. Le temps de réponse doit être acceptable dans le cadre de cette interactivité, moyennant l'utilisation d'une unité centrale adaptée.

4.2 Architecture visée

Les résultats de l'état de l'art mené en 95 (cf [2], [3]) ont montré qu'il n'existait pas de systèmes ou de structures d'accueil qui offrent des services pour l'extraction d'informations. La plupart des produits disposent d'un analyse syntaxique trop faible et/ou sont monolithiques (sans les API nécessaires). Pour offrir des outils de haut niveau et notamment des outils d'extraction, il est nécessaire de disposer d'une analyse syntaxique et d'une analyse sémantique (au moins partielle).

Cependant, il est possible d'élaborer une structure d'accueil ouverte et évolutive qui permette d'intégrer de manière incrémentale tous les services demandés en intégrant plusieurs produits. Dans ce cadre, les produits qui ont été retenus comme base de cette structure d'accueil sont Sylex et Topic.

Sylex fournit l'analyse syntaxique nécessaire. C'est le meilleur produit en terme de résultat et de rapidité. Sa modularité permet l'ajout de modules utilisateurs dans le processus d'analyse même. Son ouverture par l'intermédiaire de son API documentée rend son utilisation relativement simple. Les autres produits existants ne disposent pas d'une API linguistique aussi détaillée.

Topic fournit le moteur d'indexation et la gestion des documents. Topic est un produit à la technologie éprouvée, suffisamment ouvert pour être couplé avec Sylex.

4.3 Architecture système globale actuelle

L'architecture du prototype 96 est illustrée dans la figure 4.1. Il comprend les sous-systèmes suivants :

- réception,
- indexation,
- filtrage,
- recherche documentaire,
- gestion des noms propres,
- gestion des dossiers d'un exploitant.

Les documents arrivants sont réceptionnés dans un répertoire d'accueil puis indexés l'un après l'autre (on en extrait les mots ou expressions significatives appelés descripteurs). Une fois indexé, un document est ensuite filtré, c'est à dire comparé aux filtres contenus dans les dossiers. Si le document correspond au filtre du dossier, il est rangé dans celui-ci. Le filtre d'un dossier est un ensemble de descripteurs décrivant le dossier. Un filtre est obtenu à partir d'un descriptif en langage naturel du dossier, par une analyse de ce descriptif similaire à celle utilisée lors de l'indexation.

La recherche documentaire consiste à poser une requête en langage naturel. Celle-ci est analysée et traduite en un ensemble de descripteurs à partir desquels il est possible de rechercher, dans la base documentaire, les documents pertinents.

Le filtrage et la recherche documentaire peuvent faire appel à des ressources linguistiques tels que des dictionnaires. Une ressource particulière dans notre cas est la base de noms propres que connaît le système et qui peut être modifiée par l'intermédiaire du module de gestion de noms propres.

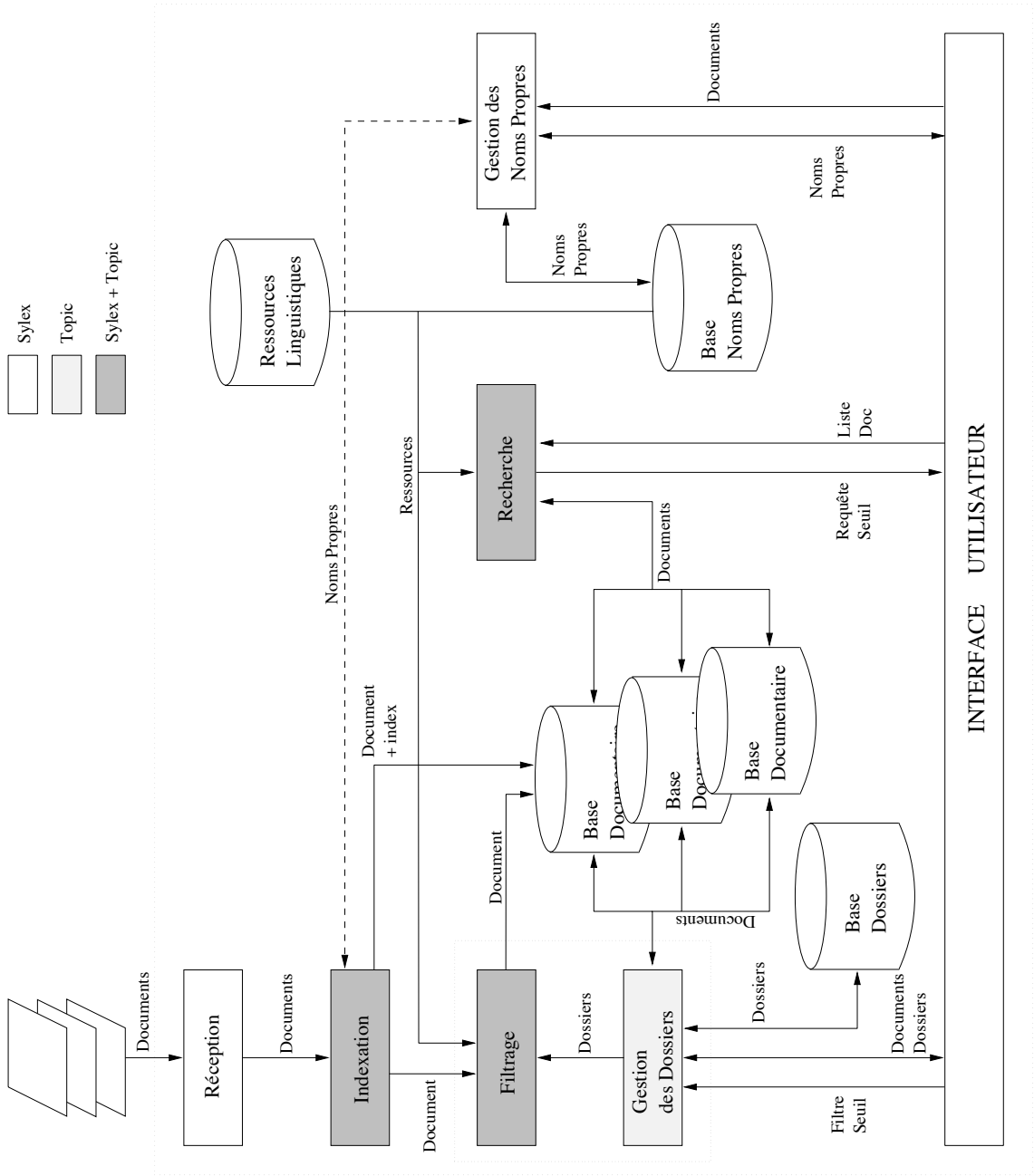


FIG. 4.1 - *Architecture globale du prototype 96*

4.4 Rappel des améliorations déjà effectuées

Un certain nombre d'améliorations et de nouvelles fonctionnalités ont déjà été présentées dans le premier et le second dossier de spécification 1996 (cf [5], [7]). Ces fonctionnalités sont les suivantes :

- la mise en évidence des éléments temporels tels que les dates, les expressions temporelles (après trois jours, la semaine dernière, ...),
- la gestion centralisée des fenêtres de texte,
- la prise en compte de la modification des ressources,
- le multilinguisme (traitement de documents de langues différentes, essentiellement français et anglais),
- le post-traitement : dans le prototype 95, le calcul des dépendances linguistiques (complément du nom, nom/adjectif, sujet/verbe, ...) était fait lors de l'indexation et les résultats mémorisés dans l'index. L'objectif est de ne plus mémoriser ces relations dans l'index, mais seulement les mots isolés et de vérifier après coup (d'où le terme de post-traitement) si les documents retrouvés concernant les mots recherchés contiennent également les dépendances recherchées. Ce traitement est indispensable pour des requêtes sur des bases documentaires déjà indexées (non par EXIST),
- la recherche documentaire sur une base Topic indexée par Topic. Topic seul indexe tous les mots d'un document sans aucune analyse linguistique (cheval et chevaux sont indexés différemment). Le problème est donc ici de retrouver lors d'une recherche, les formes fléchies sur lesquelles Topic a réalisé son indexation (par exemple, retrouver le terme chevaux à partir du mot cheval de la requête, et inversement) (cf reformulation et post-traitement dans [7]),
- l'indexation de textes français non accentués,
- la reformulation afin de pouvoir élargir la requête à d'autres mots que ceux exprimés dans la requête,
- l'interface de correction interactive de requête reformulée qui permet de contrôler la reformulation des requêtes avant de les envoyer au moteur de recherche documentaire. Cette interface permet en même temps de réaliser une reformulation manuelle des requêtes,
- la combinaison de requêtes afin de composer des requêtes en langage naturel entre elles et avec des requêtes Topic.

4.5 Améliorations et nouvelles fonctionnalités

Dans la version actuelle d'EXIST, après l'indexation, le filtrage consiste juste à faire correspondre le texte indexé avec les filtres des dossiers.

Mon stage consiste à reprendre et à améliorer les parties filtrage et gestion des documents, de manière à construire une interface plus "intelligente", une interface qui connaîtrait les préférences de l'utilisateur, et qui, en fonction de celles-ci, effectuerait les classements adéquats.

Pour pouvoir améliorer la partie filtrage, il est nécessaire de remplacer la gestion de dossier par une gestion plus globale : une gestion utilisateur. À chaque utilisateur correspondra un profil constitué d'un ensemble de centres d'intérêts (chaque centre d'intérêt correspondant à un dossier actuel, avec donc un filtre construit à partir d'une description en langage naturel).

Les nouvelles fonctionnalités pour la partie filtrage, après ce changement de gestion, sont les suivantes :

- reconnaissance de l'utilisateur d'où connaissance de son profil, de ses centres d'intérêts,
- conseils de rangement d'un document dans les différents centres d'intérêts de l'utilisateur (conseils calculés à partir du filtre présent dans chaque centre d'intérêt),
- interactivité avec l'utilisateur pour connaître ses choix de rangement,
- apprentissage à partir des choix faits par l'utilisateur, enrichissement du profil, des centres d'intérêts. Dans la version avec dossiers, le filtre ne changeait pas après l'ajout d'un document au dossier, ici, l'idée est de modifier le filtre en fonction du document que l'on ajoute,
- autonomie de classement donnée par l'utilisateur. L'utilisateur pourra interactivement donner une certaine autonomie de rangement s'il estime qu'à partir d'un certain seuil (par rapport au conseil) il n'a plus besoin de confirmer le rangement.

Ces nouvelles fonctionnalités sont résumées dans la figure 4.2 : après l'indexation, la phase de postindexation se charge d'extraire les éléments significatifs du document (comme le nombre d'occurences du mot par exemple). Ce postindex est ensuite transmis au module de conseil, qui, en fonction du profil de l'utilisateur, donne une note à ce document. Cette note passe dans le module d'autonomie, qui, toujours en fonction du profil utilisateur, choisit de soumettre cette note au choix de l'utilisateur, ou d'opérer l'apprentissage sans son approbation. Pour finir, l'apprentissage permet de classer le document dans l'intérêt approprié, de modifier le profil utilisateur.

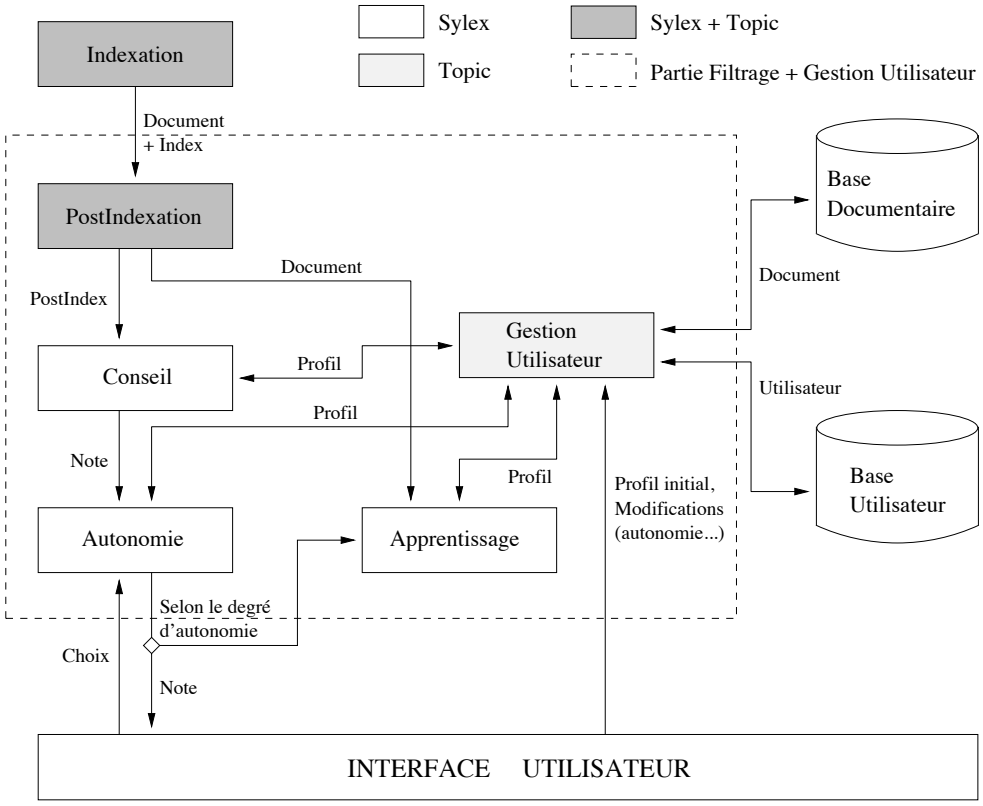


FIG. 4.2 - *Nouvelle partie Filtrage*

Après avoir examiné ces différentes fonctionnalités, on retrouve les trois composants de base d'un agent (cf [SAO]) :

- la coopération (ici avec l'utilisateur),
- l'apprentissage,
- l'autonomie.

Chapitre 5

État de l'Art

Le problème du filtrage de document peut être traité de façon indépendante vis à vis de la notion d'agent, mais il apparaît rapidement que si l'on veut étendre notre système de filtrage, la théorie sur les agents peut s'appliquer. Voici donc un état de l'art traitant, en premier lieu, des agents (avec un agent particulier : l'agent d'interface), et avec, en second lieu, le filtrage de documents.

5.1 Agents intelligents

5.1.1 Origines

Les agent intelligents se sont développés à partir des systèmes multi-agents, mais ils font maintenant partie d'une discipline plus vaste : l'intelligence artificielle distribuée *(DAI, Distributed Artificial Intelligence)* (cf [SAO]). Cette discipline touche à trois domaines :

- les systèmes multi-agents *(MAS, Multi-Agent System)*,
- la résolution de problèmes distribués *(DPS, Distributed Problem Solving)*,
- la parallélisation de l'intelligence artificielle *(PAI, Parallel Artificial Intelligence)*.

Les agents intelligents ont hérité des différentes motivations, buts, et bénéfices de ces domaines. Par exemple, grâce aux problèmes de distribution, les agents intelligents héritent de la modularité, de la rapidité (parallélisme), de la fiabilité; grâce

à l'intelligence artificielle, ils héritent de l'apprentissage, de la maintenance, et de l'indépendance vis à vis de la plateforme (cf [DAIIS]).

Dans le développement des agents, on peut distinguer deux grandes orientations :

- depuis 1977 jusqu'à aujourd'hui, développement de ce que l'on va plus loin appeler les agents collaborants *(collaborative agents)* avec des problèmes fondamentaux *(macro issues)* tels que l'interaction et la communication entre agents, la décomposition et la distribution des tâches, coordination, coopération, résolution de conflits, négociation ... Cette orientation se caractérise aussi par le développement de théories, d'architectures et de langages,
- depuis 1990 jusqu'à aujourd'hui, les recherches mettent en avant une diversification du type des agents. *"Everybody is now calling everything an agent"*.

Le tableau 5.1 résume ces deux orientations :

Orientation	Centre de recherche	Références majeures	
Orientation 1	*Macro issues*	Bond & Gasser	1988
		Gasser & Huhns	1989
		Chaib-draa *et al.*	1992
	Théories, architectures et langages	Gasser *et al.*	1995
		Wooldridge & Jennings	1995a, 1995b
		Wooldridge *et al.*	1996
Orientation 2	Diversification du type des agents	Hyacinth S. Nwana	1996

TAB. 5.1 - *Origine des agents intelligents*

5.1.2 Typologie

Il existe à l'heure actuelle différentes typologies, manières de classer les agents :

- selon leur mobilité, leur capacité à se déplacer sur les réseaux. Cela définit deux classes : les agents statiques et les agents mobiles,
- selon leur mode de réaction. Deux classes peuvent être ainsi définies : les agents délibératifs et les agents réactifs. Les agents délibératifs ont un état interne, un modèle de raisonnement, et interagissent avec les autres agents en vue

d'une décision. Au contraire, les agents réactifs utilisent un système de stimulus/réponses basé sur l'environnement immédiat,

- selon des attributs primaires que les agents devraient posséder. Trois attributs au minimum ont été identifiés : l'autonomie, l'apprentissage, et la coopération (cf figure 5.1). L'autonomie fait référence au degré de liberté que l'agent possède pour des actions sans l'accord de l'utilisateur. La coopération peut être de deux natures : la coopération avec d'autres agents, et la coopération avec l'utilisateur (interaction). Enfin, l'apprentissage consiste, pour l'agent, à voir comment agissent/réagissent les autres agents, ou bien l'utilisateur, pour pouvoir prévoir, reproduire ces mêmes schémas,

- selon leur rôle, par exemple les moteurs de recherche sur Internet,

- enfin, il y a aussi des agents hybrides qui regroupent plusieurs des points précédents.

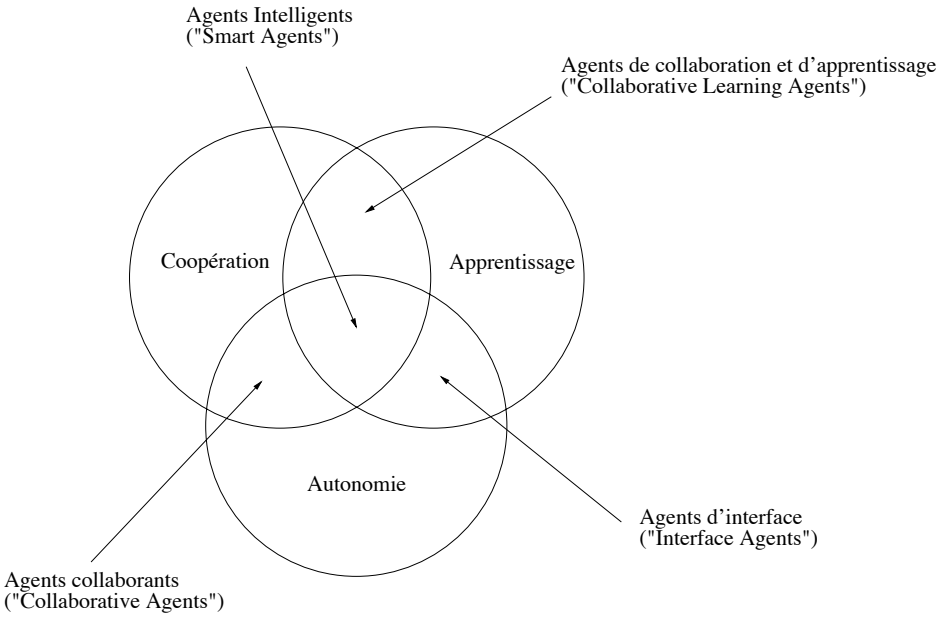


FIG. 5.1 - *Typologie d'un agent*

5.1.3 Agents d'interface

D'après la figure 5.1, on peut en déduire que notre agent de filtrage se trouve être un agent d'interface :

- la coopération se situe au niveau agent/utilisateur, c'est une interaction d'apprentissage (l'ordinateur conseille l'utilisateur, puis attend sa réaction pour en tirer des conséquences),
- l'apprentissage est bien réel puisqu'à partir d'un profil initial de l'utilisateur, l'agent doit apprendre des choix de l'utilisateur pour pouvoir modifier son profil,
- l'autonomie est le but recherché. En effet, après un apprentissage suffisamment conséquent, l'utilisateur voit que les suggestions faites par l'agent correspondent avec ses desiderata, et peut donc régler l'agent pour qu'il prenne ses décisions de lui-même (degré d'autonomie).

Les différentes actions de notre agent de filtrage se trouvent résumées dans la figure 5.2 (adaptation d'une figure de l'article de Pattie Maes [ARWIO]) :

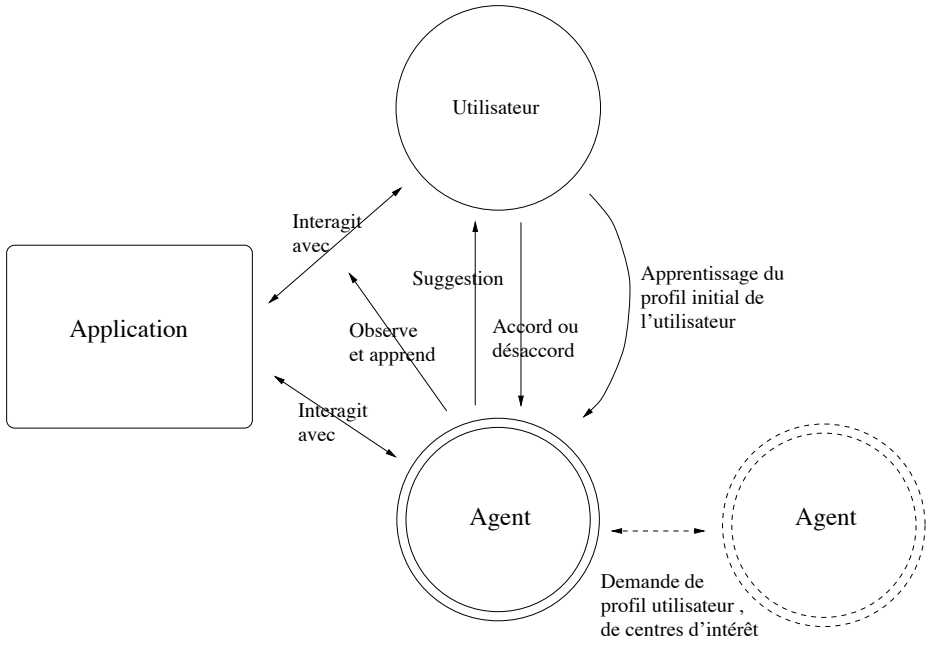


FIG. 5.2 - *Agent d'interface dans le cas du filtrage*

5.2 Filtrage de documents

5.2.1 Origines

En 1958, Luhn introduit, pour la première fois, l'idée d'un système de travail intelligent *(Business Intelligence System)* (cf [BIS]). À chaque utilisateur correspondrait un profil, utilisé pour produire une nouvelle liste de textes adaptés à chaque utilisateur. Dans ces travaux, Luhn a rapidement identifié tous les aspects du filtrage d'information actuel, notamment dans les descriptions du module de sélection *(selective dissemination of new information)*. En 1969, naît *the Special Interest Group on Selective Dissemination of Information (SIG-SDI)*. Une étude de Housman (cf [TRSIG]) pour ce groupe répertorie soixante systèmes opérationnels, ces systèmes suivants principalement le modèle de Luhn.

En 1982, Denning introduit le terme *"information filtering"* dans sa lettre du président parue dans *the Communications of the ACM* de mars (cf [EJ]). Son objectif est d'inclure dans la traditionnelle *génération* d'informations, la *réception* d'informations. Il décrit un besoin de filtrage d'information en provenance du courrier électronique *(e-mail)* et identifie six techniques possibles de filtrage de ce courrier.

Par la suite, de nombreux papiers sur le filtrage d'informations vont voir le jour, en élargissant le champ du filtrage aux *mailing list, news,* et à toutes les ressources disponibles sur les réseaux (cf [LSIIF], [EIOLN], [RBMFS]). Un des articles les plus important de cette période parait en 1987 dans *the Communications of the ACM: Intelligent information sharing systems* par Thomas Malone & co (cf [IISS]). Dans cet article, Thomas Malone introduit trois concepts de sélection d'information : *cognitive, economic, and social filtering*. Les dimensions *cogitive and economic* avaient déjà été abordées par Denning, mais la grande innovation, c'est ce qu'il appelle *the social filtering* (aussi appelé *the collaborative filtering*). Dans cette méthode de filtrage, il se base sur le fait qu'un document est représenté par les annotations des premiers lecteurs de ce document, et qu'en échangeant ces informations, les intérêts communs seraient facilement identifiables.

C'est à cette période que de nombreux projets de filtrage d'information sponsorisés par le gouvernement vont débuter. En 1989, *the United States Defense Advanced Research Projects Agency (DARPA)* sponsorise la première conférence de la série *Message Understanding Conferences (MUC)*, qui consiste en la sélection de messages au moyen d'extraction d'informations techniques. En 1990, DARPA lance le projet TIPSTER dans lequel on peut remarquer une présélection des documents grâce à des techniques statistiques *(document detection)*. En 1992, *the National Institute of Standards and Technology (NIST)*, avec le DARPA comme co-sponsor, lance une conférence annuelle *Text REtrieval Conference (TREC)* spécialisée sur le filtrage et

la recherche de texte (cf [OFTC]).

En novembre 1991, Bellcore et *the ACM Special Interest Group on Office Information Systems (SIGOIS)* co-sponsorisent un atelier de recherche sur *the High Performance Information Filtering*. Environ quarante personnes examinent les différents domaines de recherche : modélisation de l'utilisateur, sélection de l'information, domaines d'application, architectures matérielles et logicielles, systèmes d'étude, ... Toutes ces recherches aboutissent, en décembre 1992, à la publication d'une série d'articles dans *the Communications of the ACM* (cf [IFIRSC], [APDMI], [ACIF], [MUIIF], [CAIF], [PIDAM], [CFWIT], [DA], [NLUIFS]).

Dans un de ces articles : *"Information Filtering and Information Retrieval: Two Sides of the Same Coin?"* (cf [IFIRSC]), Belkin & Croft donnent une bonne définition du filtrage d'information : *"Information filtering is a name used to describe a variety of processes involving the delivery of information to people who need it."*, ainsi que de bonnes descriptions des similitudes et des différences par rapport à la recherche d'information (cf tableau 5.2). David Hull, dans sa conférence à la journée ATALA (cf [FATAS]), reprend aussi l'idée que le filtrage est l'inverse de la recherche d'informations (cf figure 5.3).

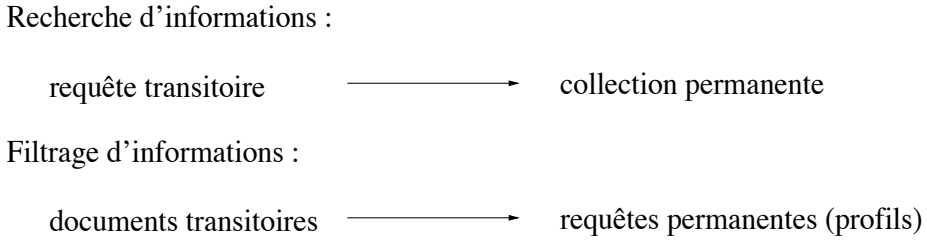


FIG. 5.3 - *Filtrage, inverse de la Recherche d'informations*

Filtrage	Recherche
- données dynamiques *(incoming data)*	- données statiques (bases)
- données non structurées, ou semi-structurées (textes)	- données structurées (fiche de paie ...)
- large flux de données	- quantité finie de données
- intérêts à long terme (profil)	- intérêts à court terme
- utilisations multiples	- utilisation unique

TAB. 5.2 - *Différences entre la recherche et le filtrage d'informations*

Pour résumer, un bon système personnalisé de filtrage doit être capable de fournir les informations dont l'utilisateur a besoin, de manière consistante et rapidement. Le système doit être capable de s'adapter aux changement de besoins de l'utilisateur, et de manière idéale, le système doit être capable d'explorer de nouveaux domaines pour trouver des informations potentiellement intéressantes pour l'utilisateur.

5.2.2 Cas d'études

Actuellement il existe de nombreux systèmes de filtrage notamment sur le courrier électronique (InfoScan [IS], Sift-Mail [SIM], MAXIMS [MX], Mailfilt [MF], Mailagent [MA], ...), sur les *news* (SMART [SM], InfoScan [IS], Browse [BR], GroupLens [GL], PEFNA [PE], Lurker [LU], SIFT [SI], NewsClip [NC], RAMA [RA], ...).

Fredrik Kilander fait une brève comparaison des différents systèmes de filtrage de *news* (cf [BCNFS]). Pour effectuer cette comparaison, Fredrik Kilander s'est basé sur plusieurs critères :

- le modèle,
- l'adaptation,
- la classification,
- les motivations,
- les opérations,
- l'implémentation,
- les résultats.

De cette comparaison résulte que :

- au niveau des modèles, les systèmes sont divisés en deux groupes : les systèmes de recherche *(retrieval system)*, et les systèmes à notation *(rating system)*. Les systèmes de recherche considèrent le flux de *news* comme une base de données dans laquelle des messages intéressants peuvent être trouvés. Les systèmes à notation essaient d'identifier les messages non pertinents pour l'utilisateur et les soustraient à ses yeux. Ils effectuent cela en identifiant les messages intéressants et en révoquant leurs messages "inverses", ceci au moyen d'une échelle de notation des messages,

- au niveau de l'adaptation, les filtres peuvent être instanciés de trois sortes :
  - le filtre est programmé par l'utilisateur,
  - l'utilisateur crée une requête à laquelle le filtre essaie de répondre,
  - le filtre est "entraîné" par des exemples donnés par l'utilisateur,
- au niveau de la classification, cinq méthodes sont appliquées :
  - *boolean expressions*,
  - *the cosine measure*,
  - *minimum description length (MDL)*,
  - *neural network modelling*,
  - *correlation of user profiles*,
- au niveau des motivations, les systèmes ne justifient pas leurs choix, ils mettent l'utilisateur face à leur sélection, avec un taux d'estimation pour certains,
- au niveau des opérations, les filtres opèrent selon trois modes :
  - en mode interactif avec l'utilisateur,
  - avec deux composants : un lecteur interactif de données, et un composant d'entraînement avec la liste de données sélectionnées,
  - avec un programme externe d'entraînement.

5.2.3 Théorie de conception

Belkin & Croft (cf [IFIRSC]), dans le cadre de la recherche d'information, avaient déjà extrait les techniques inhérentes à la sélection de documents textuels. Cette approche est reprise par Oard et Marchionini dans [CFTF], et est constituée de quatre composants de base :

- une technique de représentation des documents,
- une technique de représentation des informations désirées par l'utilisateur (profil utilisateur),
- une manière de comparer la représentation d'un document avec le profil utilisateur,
- une manière d'utiliser les résultats de cette comparaison.

Ces différents points vont être repris dans la conception.

Chapitre 6

Conception

Ce chapitre présente la partie conception de notre agent de filtrage, avec une analyse linguistique au niveau des mots uniquement. La première partie présente la conception générale de notre agent filtreur (6.1), ensuite viennent les étapes de la technologie de filtrage (cf chapitre 5.2.3)(6.2,6.3), et, pour finir, les différentes améliorations du système 96 (cf chapitre 4.5)(6.4,...6.8).

6.1 Conception générale de l'agent filtreur

À chaque nouvelle fonctionnalité du diagramme 4.2 correspond un objet. Le diagramme 4.2 montre aussi que le séquencement entre les différentes actions est important, un objet principal permettant cette gestion va donc être introduit : l'agent filtreur.

Le diagramme 6.1 présente les objets dont il va être question dans ce chapitre :

L'agent filtreur connaît les principaux objets et règle le séquencement des actions. Il prend en entrée le document postindexé, ensuite il fournit ce document ainsi que le profil utilisateur à l'objet conseil, qui lui retourne le pourcentage conseil pour chaque intérêt. L'agent filtreur fournit alors ces pourcentages ainsi que le profil utilisateur à l'objet autonomie, qui lui donne la démarche à suivre. Selon le résultat de cette démarche, l'agent filtreur demande conseil à l'utilisateur ou commence directement l'apprentissage. L'agent fournit le document, l'intérêt, et le profil à l'objet apprentissage qui modifie le profil utilisateur en conséquence.

Les objets conseil et apprentissage sont construits grâce à l'héritage, car il existe plusieurs façons de donner des conseils (cf chapitre 6.3), ou d'apprendre (cf chapitre 6.7).

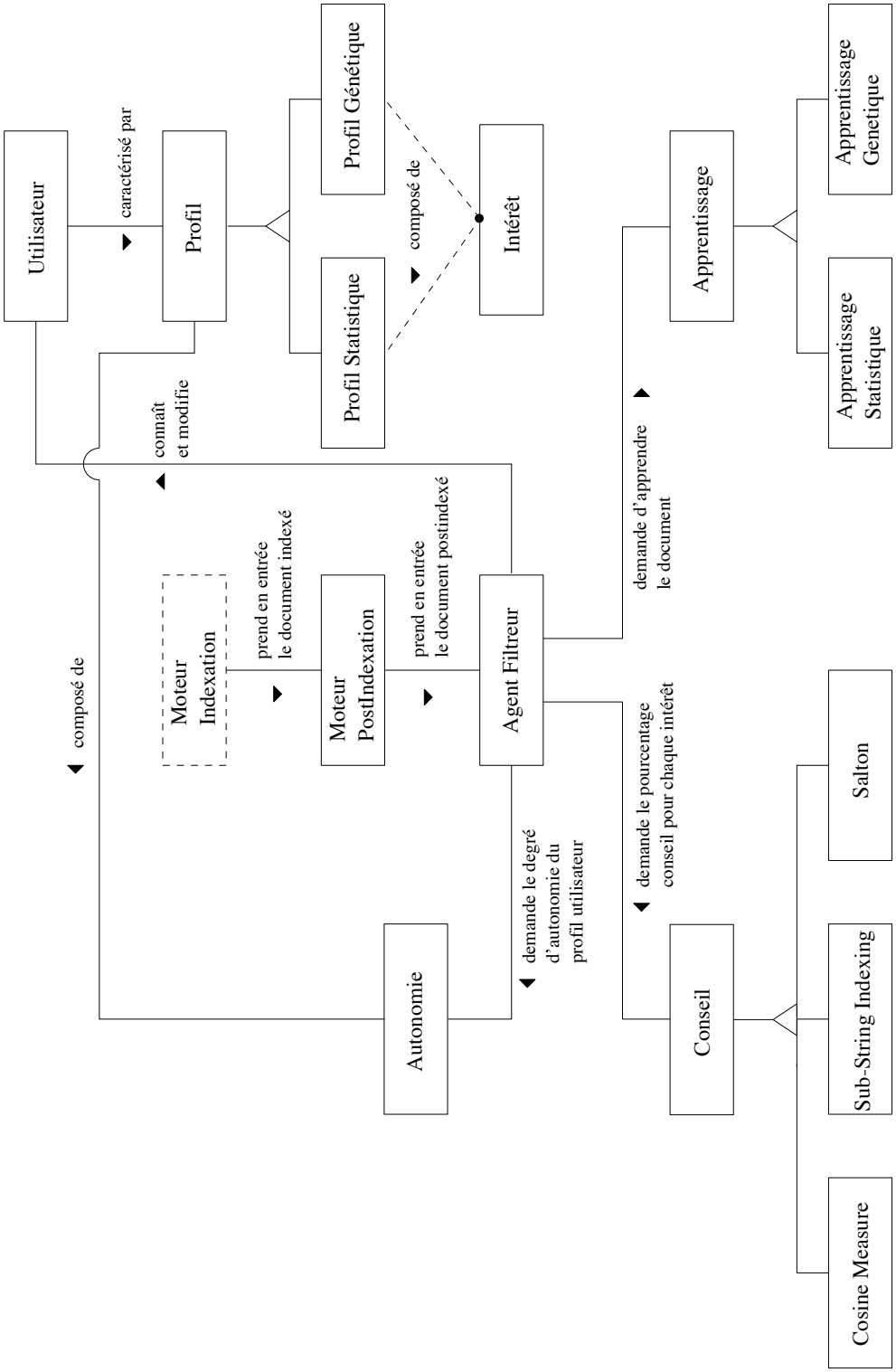


FIG. 6.1 - *Diagramme des objets*

6.2 Modélisations

6.2.1 Modélisation d'un document

Nous allons considérer une approche statistique du document, et puisque que notre première conception se situe au niveau des mots, un document sera représenté comme un liste de couples (mots,nombre d'occurences de ce mot). C'est le modéle vectoriel.

$$\forall m \in D \qquad D_r \quad = \quad \left\{ \quad (m, n_m) \quad | \quad n_m = card(\{m\} \cap D) \quad \right\}$$

avec

- D_r : modélisation du document D,
- D : document,
- m : mot (forme canonique),
- n_m : nombre d'occurences du mot m dans le document D.

De façon concrète, cette modélisation intervient dans la phase de PostIndexation (cf figures 4.2, 6.1) : en entrée de cette phase, le document arrive sous forme d'une liste d'index, en sortie le document est modélisé par une liste de PostIndex, structure { index, nombre d'occurences de cet index }.

6.2.2 Modélisation d'un utilisateur

L'utilisateur est modélisé avec deux attributs indispensables :

- un id (tout id sera unique), permettant de différencier les utilisateurs,
- un profil (qui peut être initialement vide), permettant de connaître les préférences de l'utilisateur.

D'autres attributs pourront être rajoutés pour compléter les connaissances sur l'utilisateur mais ils ne sont pas indispensables :

- mot de passe,
- nom,

- prénom,
- date de naissance,
- adresse, ...

6.2.3 Modélisation du profil utilisateur

Le profil utilisateur sera modélisé comme un ensemble de centres d'intérêts :

$$P_u \quad = \quad \left\{ \quad I_i \quad | \quad i \in [0,n] \quad \right\}$$

avec

- P_u : modélisation du profil de l'utilisateur u,
- I_i : i^{eme} centre d'intérêt du profil (il peut ne pas y en avoir).

6.2.4 Modélisation d'un centre d'intérêt

De manière à pouvoir comparer facilement un document avec un centre d'intérêt, les centres d'intérêt sont modélisés de façon presque similaire aux documents : comme une liste de triplets (mot,nombre d'occurences de ce mot, nombre de textes où se retrouve ce mot).

$$\forall m \in I_i \qquad I_{i,r} \quad = \quad \left\{ \quad (m, n_m, t_m) \quad | \quad \begin{array}{l} n_m = card(\{m\} \cap I_i), \\ t_m = \sum_{\forall D \in I_i} U_m(D) \end{array} \quad \right\}$$

avec

- I_i : $i^{ème}$ centre d'intérêt du profil,
- $I_{i,r}$: modélisation du $i^{ème}$ centre d'intérêt,
- m : mot,
- n_m : nombre d'occurences du mot m dans l'intérêt I_i,
- t_m : nombre de textes où se retrouve le mot m,
- $U_m(D)$: fonction unité $\left\{ \begin{array}{l} U_m(D) = 0 \; si \; m \notin D \\ U_m(D) = 1 \; si \; m \in D \end{array} \right.$

6.3 Comparaison d'un document avec un profil utilisateur

Pour estimer la similarité entre une requête et une collection de documents dans une base de données, de nombreuses métriques ont été appliquées notamment dans la recherche d'information. Dans le cadre de notre filtrage de documents, ces distances sont tout autant applicables : il s'agit de donner la similitude entre un document et un centre d'intérêt. Fredrik Kilander (cf [CCSUNA]) nous en donne deux : *The Sub-String Indexing* et *The Cosine Measure.*

Dans notre modèle vectoriel, chaque document est représenté par une liste de mots correspondant à un vecteur :

$V \quad = \quad \{wt_0, wt_1, \ldots, \quad wt_n\}$

avec :

- $V = D_r$ document représenté,
- $wt_i = (m, n_n)$, m mot appartenant n_m fois au document ($n_m = 0$ si m n'appartient pas au document).

6.3.1 *The Sub-String Indexing*

Cette mesure de similarité (cf [IFUATR]) donne simplement la proportion de termes communs entre le document D et le centre d'intérêt I :

$sim_{D,I} \quad = \quad \frac{2|V_D \cap V_I|}{|V_D|+|V_I|}$

avec $|V|$ cardinal du vecteur V : nombre de mots m du vecteur V tels que $n_m \neq 0$.

6.3.2 *The Cosine Measure*

The cosine measure vient d'une analogie avec la trigonométrie (cf [IMIR]) : un angle entre 2 vecteurs de dimension n est toujours mesurable par projection sur un plan (2D). Quand les vecteurs pointent dans la même direction, l'angle entre eux est $\varphi = 0$, quand ils sont perpendiculaires, l'angle est $\varphi = \frac{\pi}{2}$. Prendre le cosinus de cet angle normalise l'échelle entre ces deux extrêmes :

$\left\{ \quad cos(\varphi = \frac{\pi}{2}) = 0, \ldots \quad cos(\varphi = 0) = 1 \quad \right\}$

À partir d'un document D et d'un centre d'intérêt I, il est donc possible de calculer *the cosine similarity* :

$$cosine_{D,I} = \frac{\sum_{m \in D \cup I} n_{D,m} \cdot n_{I,m}}{\sqrt{\sum_{m \in D} n_{D,m}^2} \cdot \sqrt{\sum_{m \in I} n_{I,m}^2}}$$

avec $n_{x,m}$ le nombre d'occurences du mot m dans x (document D, ou intérêt I).

6.3.3 Métrique inspirée de celle de Salton

Gerard Salton a fait de nombreuses recherches sur la similitude entre deux documents notamment dans le cadre de la recherche d'information (cf [EBIR], [GTMIR], [DATR]).

Une des mesures proposées est de la forme :

$$salton_{D,I} = \frac{\sum_{m \in D \cup I} n_{D,m} \cdot n_{I,m} \cdot log^2\left(\frac{t_{I,m}}{N}\right)}{\sqrt{\sum_{m \in D} n_{D,m}^2 \cdot log^2\left(\frac{t_{I,m}}{N}\right)} \cdot \sqrt{\sum_{m \in I} n_{I,m}^2 \cdot log^2\left(\frac{t_{I,m}}{N}\right)}}$$

avec :

- $n_{x,m}$: nombre d'occurences du mot m dans x (document D, ou intérêt I),
- $t_{I,m}$: nombre de textes où se retrouve le mot m,
- N : nombre total de textes.

L'idée de Salton dans cette mesure est de reprendre le principe de la *cosine measure* avec pondération des mots les plus occurrents, mais aussi d'introduire un terme d'inverse de fréquence : $log\left(\frac{t}{N}\right)$. Ce terme pondère de manière plus importante les mots présents dans peu de texte, et diminue la pondération des mots présents dans tous les textes. Il y a une équirépartition des chances. Cette mesure donne de bon résultat dans la recherche d'information, mais dans le filtrage d'information, le problème est inverse : les mots présents dans tous les textes doivent être plus importants qu'un mot présent dans un seul texte.

La mesure devient donc :

$$salton_{D,I} = \frac{\sum_{m \in D \cup I} n_{D,m} \cdot n_{I,m} \cdot log^2(t_{I,m})}{\sqrt{\sum_{m \in D} n_{D,m}^2 \cdot log^2(t_{I,m})} \cdot \sqrt{\sum_{m \in I} n_{I,m}^2 \cdot log^2(t_{I,m})}}$$

6.4 Reconnaissance de l'utilisateur

La reconnaissance de l'utilisateur s'effectue lors du lancement de l'application : un login est demandé à l'utilisateur de manière à retrouver toutes les informations le concernant dans la base utilisateur, l'information la plus importante étant son profil. Si l'utilisateur n'est pas présent dans la base, on crée ce nouvel utilisateur avec un profil vierge (dans un système multi-agents, on pourrait imaginer une coopération entre agent pour savoir s'ils connaissent cet utilisateur).

6.5 Conseils de classement d'un document

Les conseils seront prodigués sous forme de note, de pourcentage de similitude entre le document et chacun des centres d'intérêt de l'utilisateur (calculés à partir des coefficients de similitude, chapitre 6.3).

6.6 Interactivité avec l'utilisateur

L'interactivité avec l'utilisateur s'effectue à plusieurs moments :

- choix du coefficient de similitude en vue du conseil,
- choix du centre d'intérêt où mettre le document après le conseil,
- réglage des paramètres de l'autonomie,
- choix de l'algorithme d'apprentissage, et de ses paramètres.

6.7 Apprentissage

6.7.1 Apprentissage vectoriel simple

L'apprentissage vectoriel consiste à ajouter les mots du document (avec leurs nombres d'occurences) dans le centre d'intérêt :

Avant apprentissage :

Intérêt $I : I \quad = \quad \{ \quad (m_I, n_{m_I}, t_{m_I}) \quad \}$

document $D : D_r \quad = \quad \{ \quad (m_D, n_{m_D}) \quad | \quad m_D \in D, n_{m_D} = card(\{m_D\} \cap D) \quad \}$

Après apprentissage :

centre d'intérêt $I : I \quad = \quad \{ \quad (m'_I, n_{m'_I}, tm'_I) \quad \}$

avec :

- $m'_I = \{m_D\} \cup \{m_I\}$,
- $n_{m'_I} = \begin{cases} n_{m_D} & si\ m_D \notin \{m_I\} \\ n_{m_D} + n_{m_I} & si\ m_D \in \{m_I\} \end{cases}$
- $t_{m'_I} = \begin{cases} 1 & si\ m_D \notin \{m_I\} \\ t_{m_I} + 1 & si\ m_D \in \{m_I\} \end{cases}$

6.7.2 Apprentissage statistique

L'apprentissage statistique consiste à ajouter les x mots les plus occurrents du document dans le centre d'intérêt. x est une valeur fixée par l'utilisateur (pouvant être fixe, ou relative au nombre de mots du document).

Avant apprentissage :

Intérêt $I : I \quad = \quad \{ \quad (m_I, n_{m_I}, t_{m_I}) \quad \}$

document $D : D_r \quad = \quad \{ \quad (m_D, n_{m_D}) \quad | \quad m_D \in D, n_{m_D} = card(\{m_D\} \cap D) \quad \}$

Après apprentissage :

soit $X \subseteq D$ tel que :

- $card(X) = x$,

- $\forall m \in X, \quad \forall m' \in D - X, \quad n_{\boldsymbol{m}} \geq n_{\boldsymbol{m'}}$

centre d'intérêt I: $I \quad = \quad \left\{ \quad (m'_I, n_{\boldsymbol{m'_I}}, t_{\boldsymbol{m'_I}}) \quad \right\}$

avec :

- $m'_I = \{m_X\} \cup \{m_I\}$,
- $n_{\boldsymbol{m'_I}} = \begin{cases} n_{\boldsymbol{m}_X} & si\ m_X \notin \{m_I\} \\ n_{\boldsymbol{m}_X} + n_{\boldsymbol{m}_I} & si\ m_X \in \{m_I\} \end{cases}$
- $t_{\boldsymbol{m'_I}} = \begin{cases} 1 & si\ m_X \notin \{m_I\} \\ t_{\boldsymbol{m}_I} + 1 & si\ m_X \in \{m_I\} \end{cases}$

6.7.3 Apprentissage génétique

Les algorithmes génétiques sont des procédures d'optimisation (cf [AGA]), c'est-à-dire des algorithmes résolvant des problèmes formulés comme la maximisation d'une fonction f, et sont inspirés des lois de l'évolution des êtres vivants selon Darwin.

En voici l'algorithme type :

1. Génération aléatoire et évaluation de n solutions pour former la population initiale $P(1)$, $t \leftarrow 1$,
2. Génération de $P(t+1)$:
   - Sélection d'une population $P_{\boldsymbol{s}}$ de n solutions à partir de $P(t)$ en utilisant la loi $P_{\boldsymbol{select}}(s) \quad = \quad \frac{\boldsymbol{f(s)}}{\sum_{s' \in P(t)} \boldsymbol{f(s')}}$
   - Recombinaison de $P_{\boldsymbol{s}}$ en $P_{\boldsymbol{r}}$ en utilisant le croisement à un point qui échange des sous-chaînes entre solutions,
   - Mutation de $P_{\boldsymbol{r}}$ en $P(t+1)$ en modifiant des bits avec une probabilité $p_{\boldsymbol{mut}}$ très faible,
3. Évaluation des nouvelles solutions de $P(t+1)$,
4. $t \leftarrow t+1$,
5. Aller en 2 ou Stop.

En regardant ces différentes étapes, deux conditions d'application des algorithmes génétiques apparaissent :

- représentation adéquate des solutions : généralement sous forme de chaîne de longueur fixe afin de permettre le bon fonctionnement du croisement. Cette hypothèse de "construction de blocs" suppose que le croisement peut effectivement combiner des blocs de solutions de manière à obtenir de meilleurs individus. Si cette hypothèse n'est pas vérifiée, le rôle et l'intérêt du croisement seront mineurs,
- conditions sur la fonction de qualité f : elle doit être calculable assez rapidement pour permettre l'évaluation de plusieurs générations de solutions. Notons que certaines fonctions sont en théorie difficiles à optimiser pour un algorithme génétique, ces fonctions sont telles que la combinaison de blocs utiles éloignent toujours l'algorithme de l'optimum de f.

On aimerait appliquer ces algorithmes génétiques à notre problème de filtrage, notre problème étant d'optimiser un profil. Les correspondances seraient les suivantes (tableau 6.1) :

Algorithme génétique type	Notre problème de filtrage
n chaînes solutions	n listes de couples (m, n_m)
sous-chaînes	sous-listes
f fonction qualité	fonctions de similitude
recombinaison de sous-chaînes	recombinaison de sous-listes
mutation de bits	changement de la valeur des occurrences n_m

TAB. 6.1 - *Application d'un algorithme génétique dans le cas du filtrage*

Cependant, en regardant les conditions d'applications d'un algorithme génétique, on voit apparaître quelques difficultés :

- les solutions, dans le cadre du filtrage, ne sont pas représentées avec une taille fixe, mais avec des listes,
- les coefficients de similitudes sont calculés en parcourant toute la liste de mots, et cette liste pouvant être de taille très variable, l'efficacité peut s'en trouver grandement réduite.

Pour résoudre ces problèmes de taille de liste et de complexité de calcul, on ne considère plus la totalité de la liste de mots, mais une sous-liste des mots les plus significatifs, les mots les plus occurrents : on utilise l'apprentissage statistique.

6.8 Autonomie

Un agent est autonome quand il opère de façon complètement autonome, c'est-à-dire qu'il décide de lui-même des opérations à effectuer, selon ses buts, et selon ce qu'il capte de son environnement (cf [MAAA]).

Cependant, ici, le problème est plus complexe. Notre agent est un agent d'interface, son but est d'aider l'utilisateur dans ces choix, et d'en déduire des préférences. Notre agent ne doit donc pas être complètement autonome, mais avoir un degré d'autonomie.

Dans notre agent filtreur, le degré d'autonomie est modélisé avec une approche similaire de celle de Pattie Maes dans [ARWIO]. En effet, les conseils prodigués par notre agent se situent sur une échelle de pourcentage, l'utilisateur va donc définir des seuils (en pourcentage) définissant le degré d'autonomie de cet agent. Ces différents seuils se trouvent dans la figure 6.2.

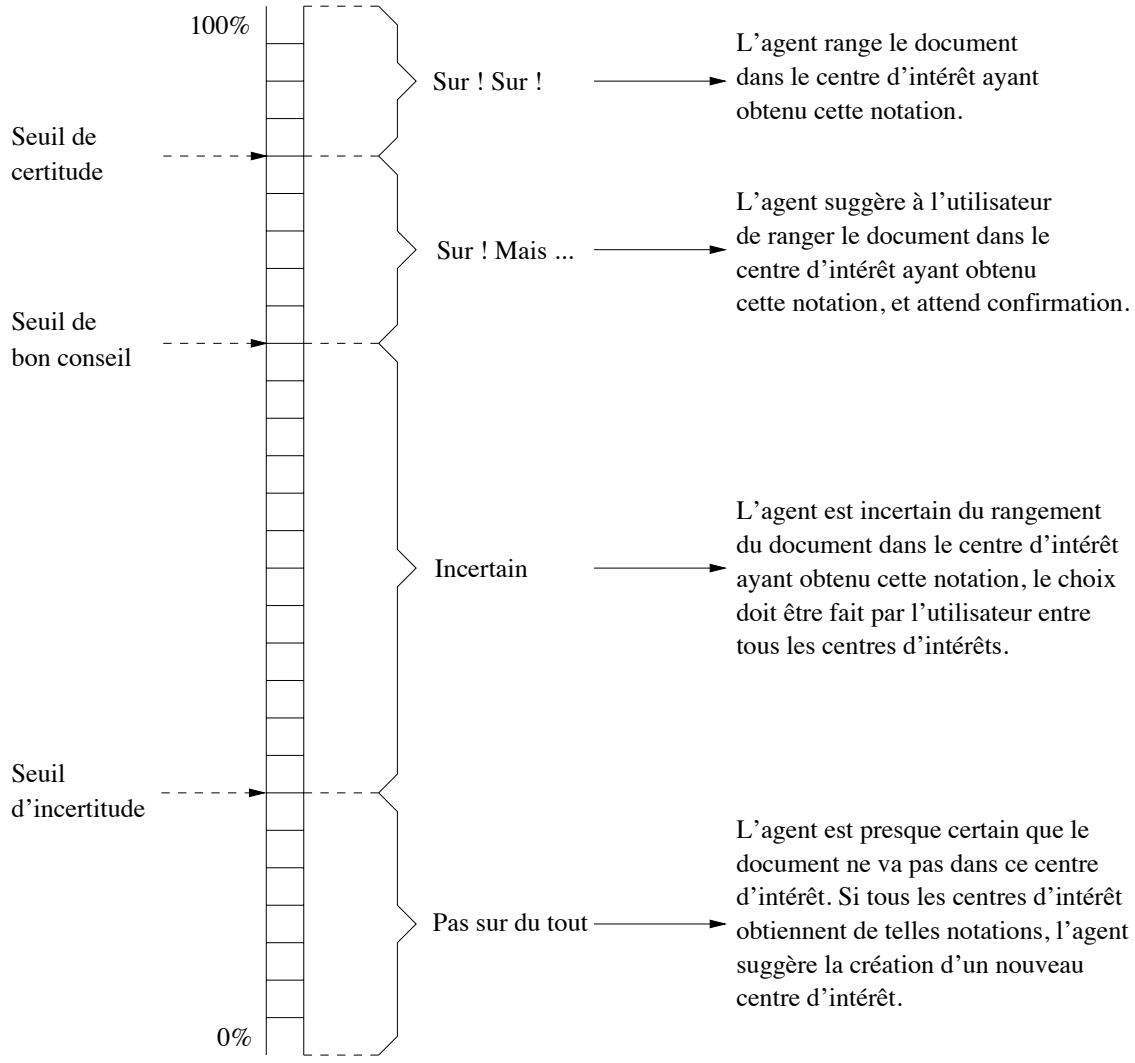


FIG. 6.2 - *Définition du degré d'autonomie de notre agent*

Chapitre 7

Résultats

Voici maintenant les différents résultats de tests opérés sur les différentes fonctionnalités de l'application lorsque l'on en fait varier les paramètres.

Par la suite, on considère :

- l'analyse linguistique se fait au niveau des mots,
- profil P : profil utilisateur avec un seul centre d'intérêt sur "La mort de François Mitterrand",
- texte t_1 : biographie de François Mitterrand,
- texte t_2 : texte sur la guerre en Arménie.

Les différentes mesures de performance ont été effectuées sur une SPARC 20.

7.1 Apprentissage vectoriel : Conseils / Nombre de textes appris

7.1.1 Texte t_1 : pertinent

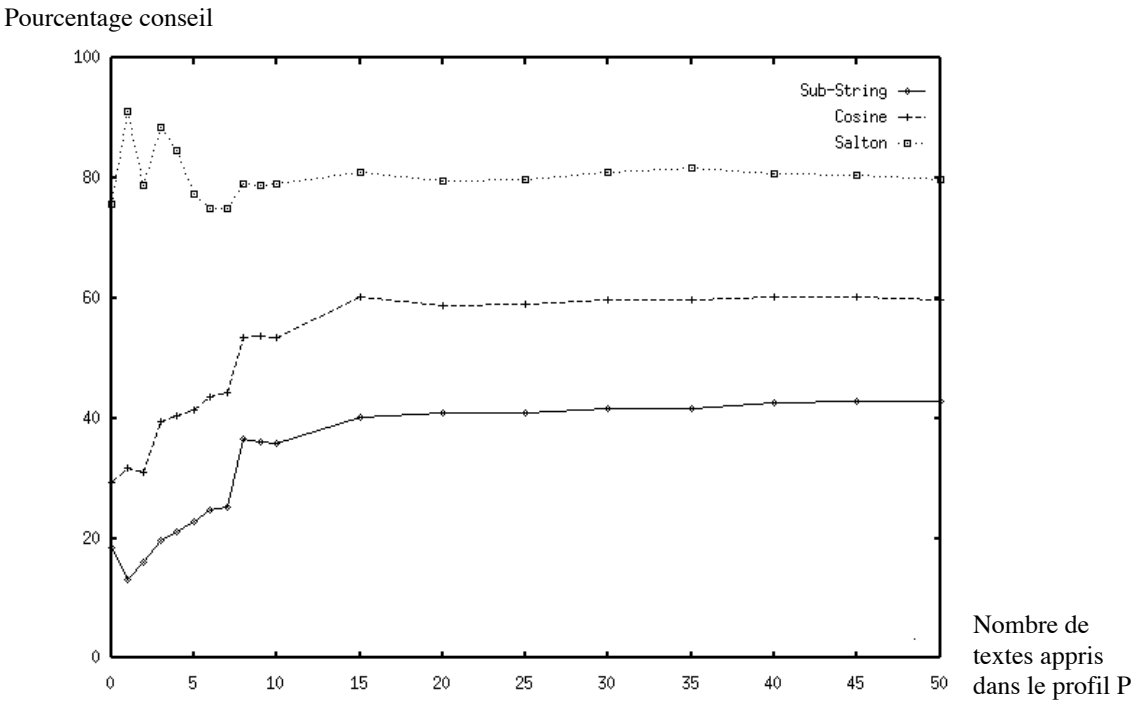


FIG. 7.1 - *Apprentissage vectoriel : texte pertinent*

Interprétation du graphique 7.1 :

- *Sub-string indexing measure* :
  - courbe croissante (sauf premier texte), assez continue,
  - écart entre plus mauvaise valeur et valeur stable important : ~30%,
  - stabilité à partir de 15 textes,
  - résultat pas tellement significatif : ~40% après stabilité pour un texte pertinent,
- *Cosine measure* :
  - courbe croissante, continue,

- écart entre plus mauvaise valeur et valeur stable important : ∼30%,
- stabilité à partir de 15 textes,
- résultat assez significatif : ∼60% après stabilité pour un texte pertinent,

– Mesure inspirée de Salton :

- démarrage discontinu,
- écart entre plus mauvaise valeur et valeur stable raisonnable : ∼10%,
- stabilité à partir de 10 textes,
- résultat significatif : ∼80% après stabilité pour un texte pertinent.

7.1.2 Texte t_2 : non pertinent

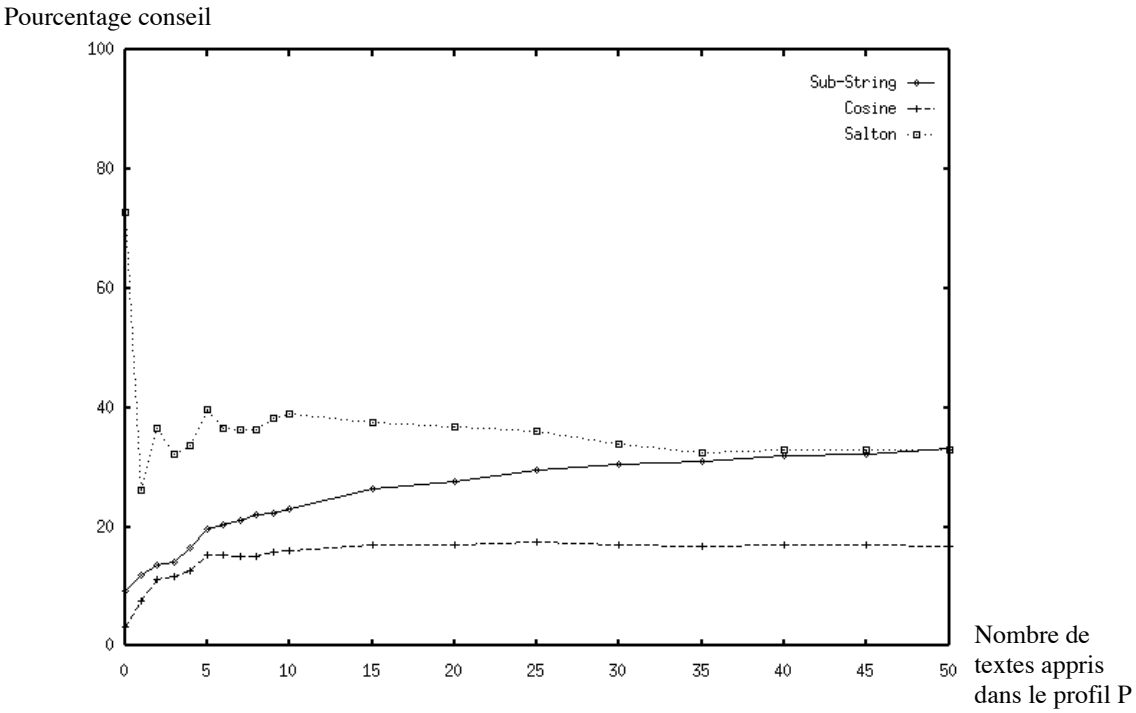


FIG. 7.2 - *Apprentissage vectoriel : texte non pertinent*

Interprétation du graphique 7.2 :

– *Sub-string indexing measure* :

- courbe croissante, continue,

  - écart entre plus mauvaise valeur et valeur stable assez important : ∼20%,
  - stabilité non confirmée avec 50 textes, courbe toujours légèrement croissante,
  - résultat moyennement significatif : ∼35% pour un texte non pertinent après 50 textes appris,

- *Cosine measure* :

  - courbe croissante, continue,
  - écart entre plus mauvaise valeur et valeur stable assez raisonnable : ∼15%,
  - stabilité à partir de 5 textes,
  - résultat significatif : ∼17% après stabilité pour un texte non pertinent,

- Mesure inspirée de Salton :

  - démarrage discontinu, puis décroissante,
  - écart entre plus mauvaise valeur et valeur stable très important si l'on tient compte de la discontinuité de départ : ∼40%, et si l'on n'en tient pas compte, cela redevient raisonnable : ∼5%,
  - stabilité à partir de 5-10 textes avec décroissance,
  - résultat moyennement significatif : ∼30-35% après stabilité pour un texte non pertinent.

7.1.3 Conclusion

À la suite de ces différents tests d'apprentissage, la mesure inspirée de Salton obtient les meilleurs résultats :

- conseils pour un texte pertinent et texte non pertinent significatifs (respectivement de l'ordre de 75-90% contre 25-35%),
- écart entre conseil pour un texte pertinent et conseil pour un texte non pertinent important (de l'ordre de 40 à 60%),
- stabilité très rapide (5 à 10 textes).

7.2 Apprentissage statistique : Conseils / Nombre de textes appris

7.2.1 Texte t_1 : pertinent

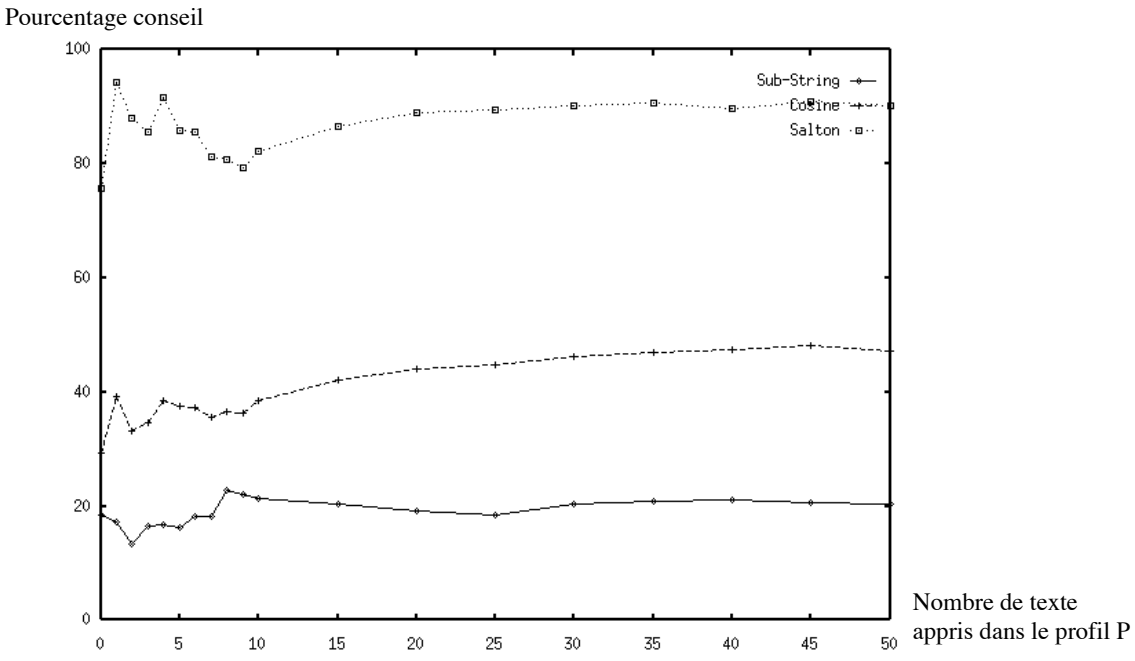


FIG. 7.3 - *Apprentissage statistique : texte pertinent*

Interprétation du graphique 7.3 :

- *Sub-string indexing measure* :
  - courbe assez continue, sauf au départ,
  - écart entre plus mauvaise valeur et valeur stable raisonnable : ~5%,
  - stabilité à partir de 10 textes,
  - résultat pas tellement significatif : ~20% après stabilité pour un texte pertinent,
- *Cosine measure* :
  - courbe croissante, continue, sauf au départ,

- écart entre plus mauvaise valeur et valeur stable assez raisonnable : ∼15%,
- stabilité à partir de 10-15 textes,
- résultat moyennement significatif : ∼40% après stabilité pour un texte pertinent,

– Mesure inspirée de Salton :

- démarrage discontinu,
- écart entre plus mauvaise valeur et valeur stable raisonnable : ∼10%,
- stabilité à partir de 10-15 textes,
- résultat significatif : ∼90% après stabilité pour un texte pertinent.

7.2.2 Texte t_2 : non pertinent

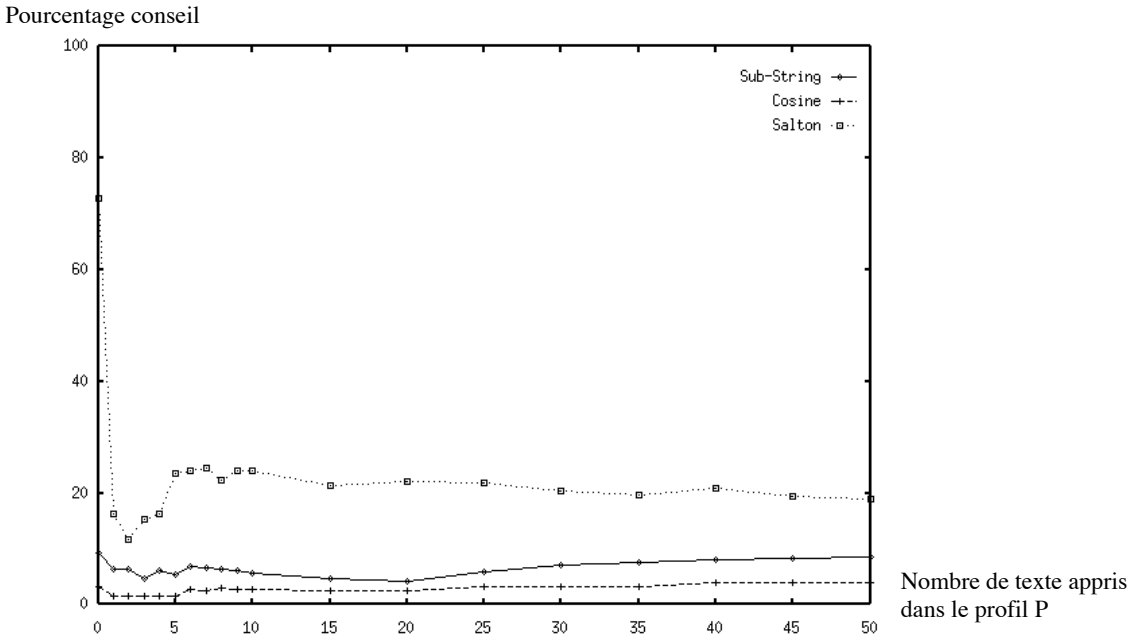


FIG. 7.4 - *Apprentissage statistique : texte non pertinent*

Interprétation du graphique 7.4 :

– *Sub-string indexing measure* :

- courbe continue, stable,

  - écart entre plus mauvaise valeur et valeur stable minime : ~1-5%,
  - stabilité à partir de 5 textes,
  - résultat significatif : ~2-5% pour un texte non pertinent après stabilité,
- *Cosine measure* :
  - courbe continue,
  - écart entre plus mauvaise valeur et valeur stable faible : ~5%,
  - stabilité à partir de 25 textes,
  - résultat significatif : ~7% après stabilité pour un texte non pertinent,
- Mesure inspirée de Salton :
  - démarrage discontinu, puis décroissante,
  - écart entre plus mauvaise valeur et valeur stable très important si l'on tient compte de la discontinuité de départ : ~50%, et si l'on n'en tient pas compte, cela redevient raisonnable : ~10%,
  - stabilité à partir de 5-10 textes avec décroissance,
  - résultat moyennement significatif : ~20% après stabilité pour un texte non pertinent.

7.2.3 Conclusion

À la suite de ces différents tests d'apprentissage, la mesure inspirée de Salton obtient les meilleurs résultats :

- conseils pour un texte pertinent et texte non pertinent significatifs (respectivement de l'ordre de 80-90% contre 20%),
- écart entre conseil pour un texte pertinent et conseil pour un texte non pertinent important (de l'ordre de 60%),
- stabilité très rapide (5 à 10 textes).

7.3 Apprentissage génétique : Conseils / Nombre de textes appris

7.3.1 Texte t_1 : pertinent

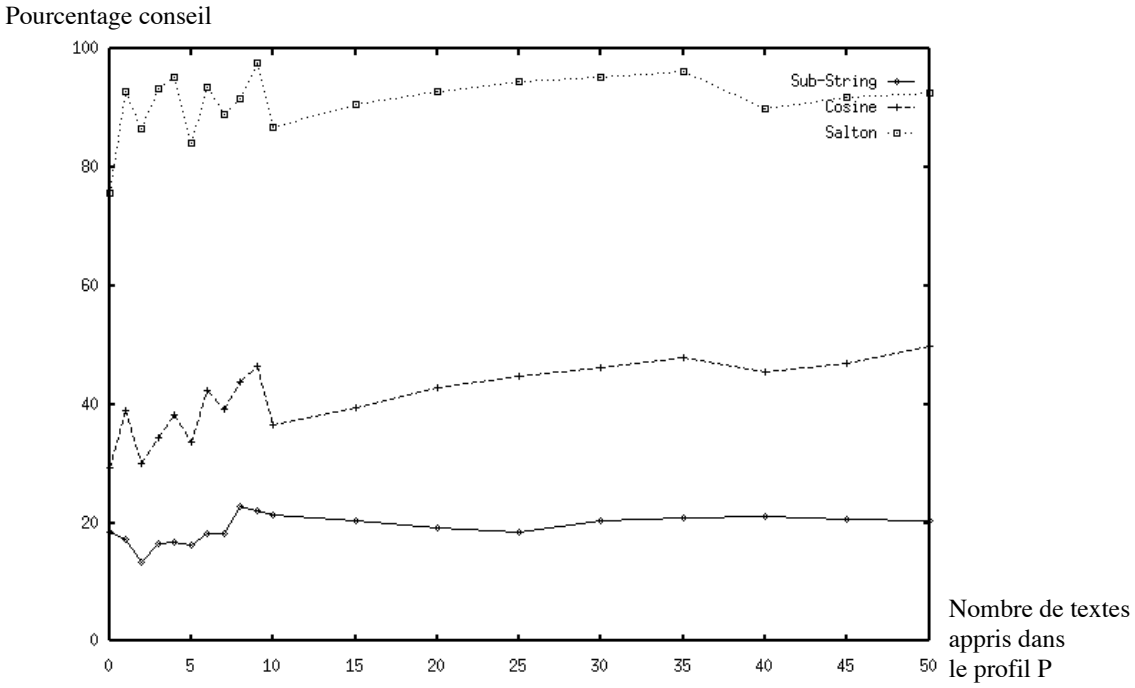


FIG. 7.5 - *Apprentissage génétique : texte pertinent*

Interprétation du graphique 7.5 :

- *Sub-string indexing measure* :
  - courbe assez continue, sauf au départ,
  - écart entre plus mauvaise valeur et valeur stable raisonnable : ~5%,
  - stabilité à partir de 10 textes,
  - résultat pas tellement significatif : ~20% après stabilité pour un texte pertinent,
- *Cosine measure* :
  - démarrage discontinu,

- écart entre plus mauvaise valeur et valeur stable important : ∼20%,
- stabilité à partir de 10-15 textes,
- résultat moyennement significatif : ∼50% après stabilité pour un texte pertinent,

– Mesure inspirée de Salton :

- démarrage discontinu,
- écart entre plus mauvaise valeur et valeur stable important : ∼20%,
- stabilité à partir de 10-15 textes,
- résultat significatif : ∼95% après stabilité pour un texte pertinent.

7.3.2 Texte t_2 : non pertinent

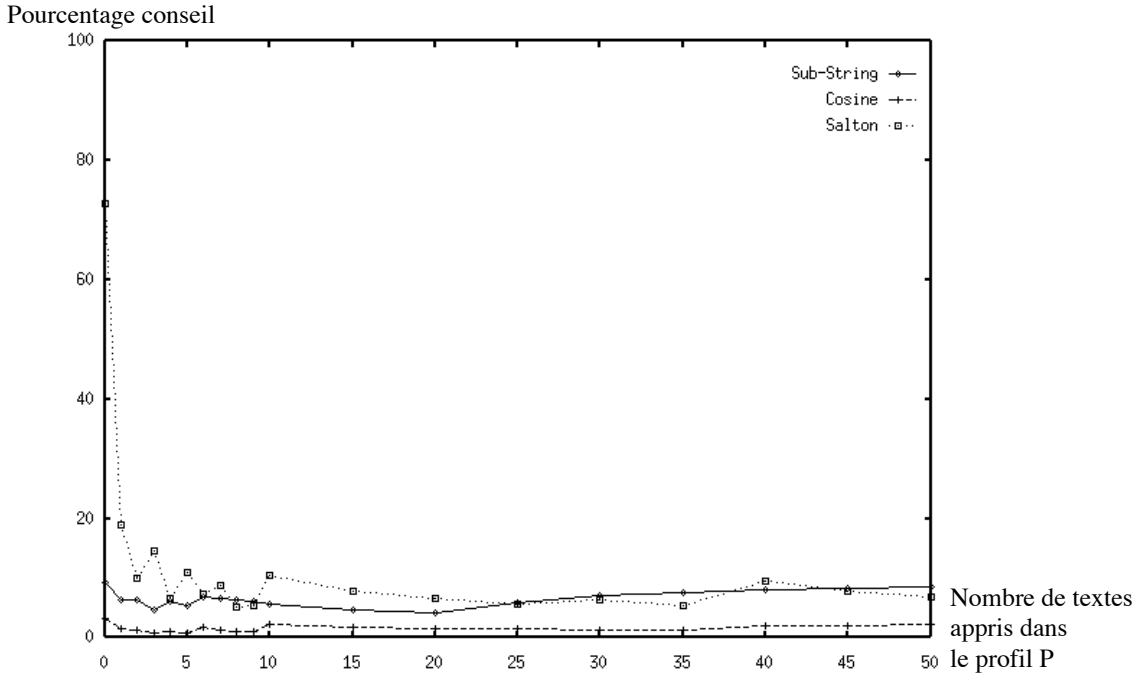


FIG. 7.6 - *Apprentissage Génétique : texte non pertinent*

Interprétation du graphique 7.6 :

- *Sub-string indexing measure* :
  - courbe continue, stable,
  - écart entre plus mauvaise valeur et valeur stable minime : ∼1-5%,
  - stabilité à partir de 10 textes,
  - résultat significatif : ∼2-5% pour un texte non pertinent après stabilité,
- *Cosine measure* :
  - courbe continue,
  - écart entre plus mauvaise valeur et valeur stable faible : ∼5%,
  - stabilité à partir de 25 textes,
  - résultat significatif : ∼7% après stabilité pour un texte non pertinent,
- Mesure inspirée de Salton :
  - démarrage discontinu, puis décroissante,
  - écart entre plus mauvaise valeur et valeur stable très important si l'on tient compte de la discontinuité de départ : ∼60%, et si l'on n'en tient pas compte, cela redevient raisonnable : ∼15%,
  - stabilité à partir de 5-10 textes avec décroissance,
  - résultat significatif : ∼10% après stabilité pour un texte non pertinent.

7.3.3 Conclusion

À la suite de ces différents tests d'apprentissage, la mesure inspirée de Salton obtient les meilleurs résultats :

- conseils pour un texte pertinent et texte non pertinent significatifs (respectivement de l'ordre de 90% contre 10%),
- écart entre conseil pour un texte pertinent et conseil pour un texte non pertinent très important (de l'ordre de 80%),
- stabilité rapide (10 à 15 textes).

7.4 Apprentissages : Performance en temps / Nombre de textes

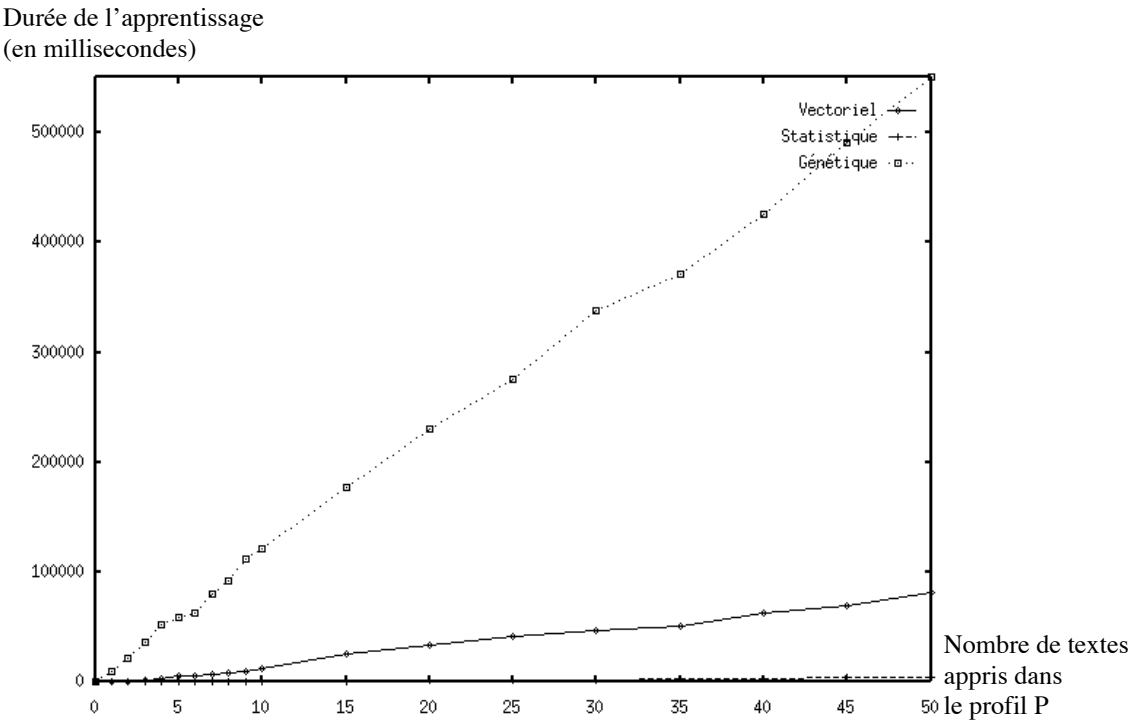


FIG. 7.7 - *Performances des différents apprentissages*

Dans le graphique 7.7, les différences de performances des apprentissages sont flagrantes :

- l'apprentissage statistique est le plus rapide : ~5 secondes pour apprendre 50 textes (préalablement indexés),
- ensuite, c'est l'apprentissage vectoriel : ~80 secondes pour 50 textes,
- et enfin, l'apprentissage génétique : ~9 minutes pour 50 textes.

En comparant les pourcentages conseil des différents apprentissages, il s'avère que les meilleurs résultats sont obtenus pour les apprentissages les plus longs. Néanmoins, le ratio (conseils / performance) est très largement en faveur de l'apprentissage statistique en raison de sa très grande rapidité et de la justesse de ses conseils.

7.5 Conseils / Nombre de mots dans le Profil P

Si dans l'apprentissage, on ne se limite pas sur le nombre de mots à apprendre, ce nombre augmente rapidement au détriment de la vitesse de calcul. Voici les différences de conseil que l'on observe si on fixe le nombre de mots les plus occurrents :

7.5.1 Texte t_1 : pertinent

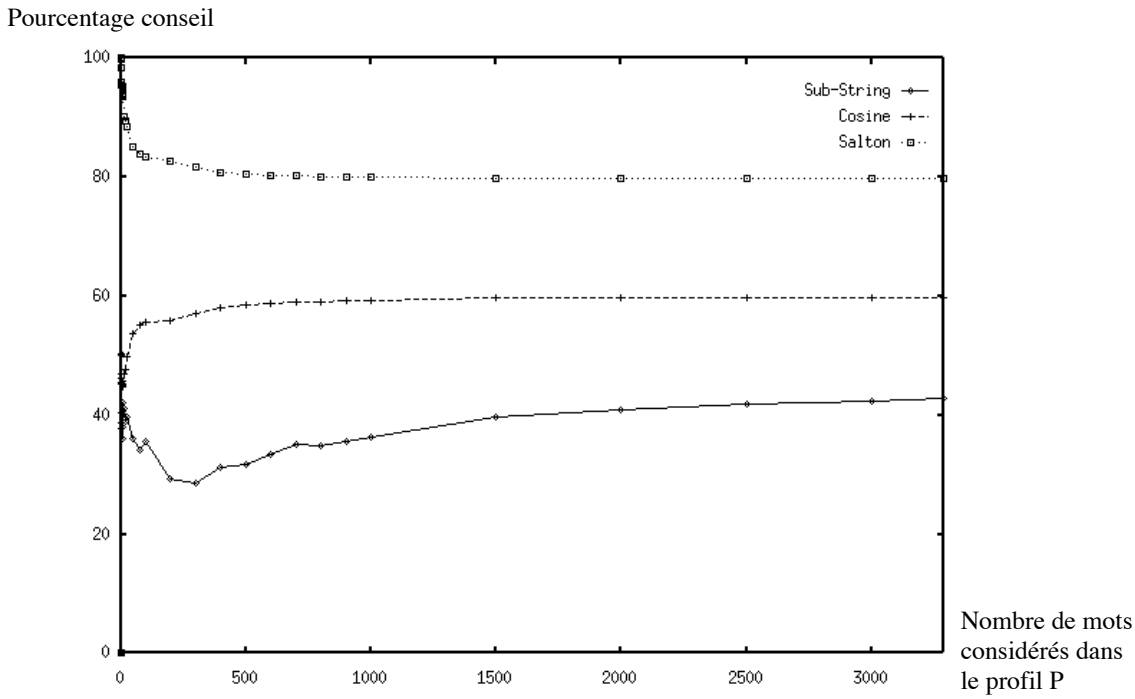


FIG. 7.8 - *Conseils pour le texte t_1 / Nombre de mots pris dans le profil P*

Des mesures du graphique 7.8, il ressort que :

- *the sub-string measure* est quand même bien affectée par la diminution de mots pris en compte,
- à l'inverse, *the cosine measure* et la mesure inspirée de Salton restent stables.

Mais ce qui intéressant se passe à l'origine (graphique 7.9) :

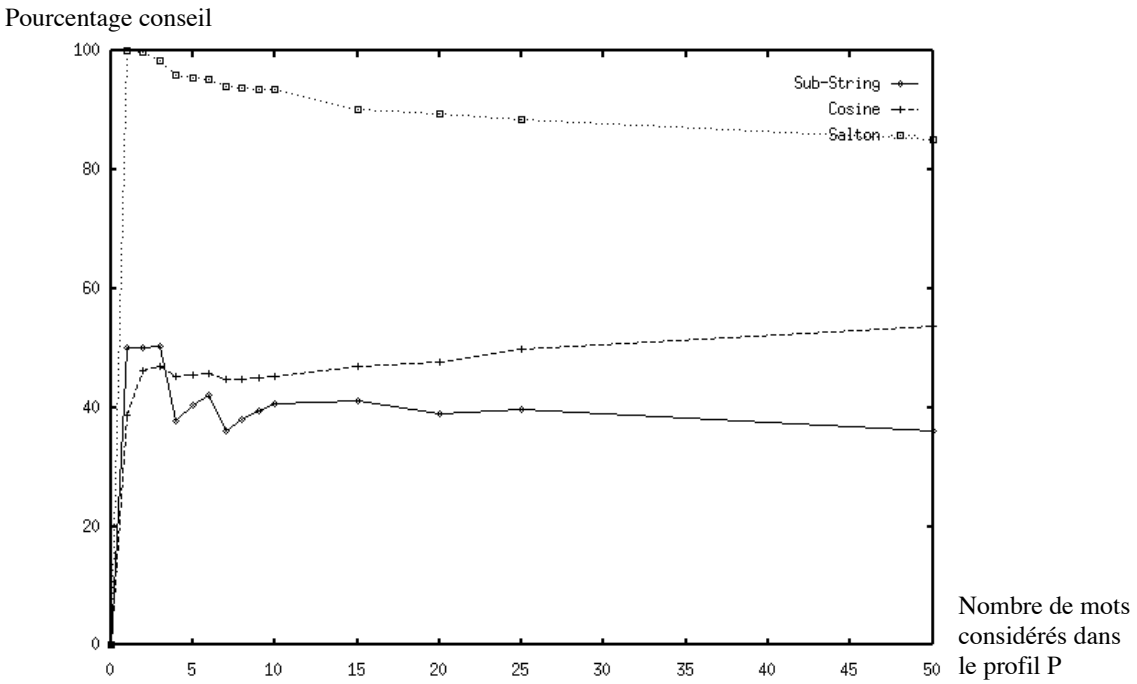


FIG. 7.9 - *Zoom sur le début du graphique précédent*

Lorsque le nombre de mots est grandement réduit, il résulte que :

- *the sub-string measure* retrouve son niveau de conseil comme avec beaucoup plus de mots, mais lorsque le nombre de mots est inférieur à 10, la courbe est discontinue, instable (dépend du texte conseillé),
- *the cosine measure* reste stable jusque ~5 mots,
- la mesure inspirée de Salton obtient de très bons résultats, puisqu'elle augmente significativement lorsque le nombre de mots diminue.

7.5.2 Texte t_2 : non pertinent

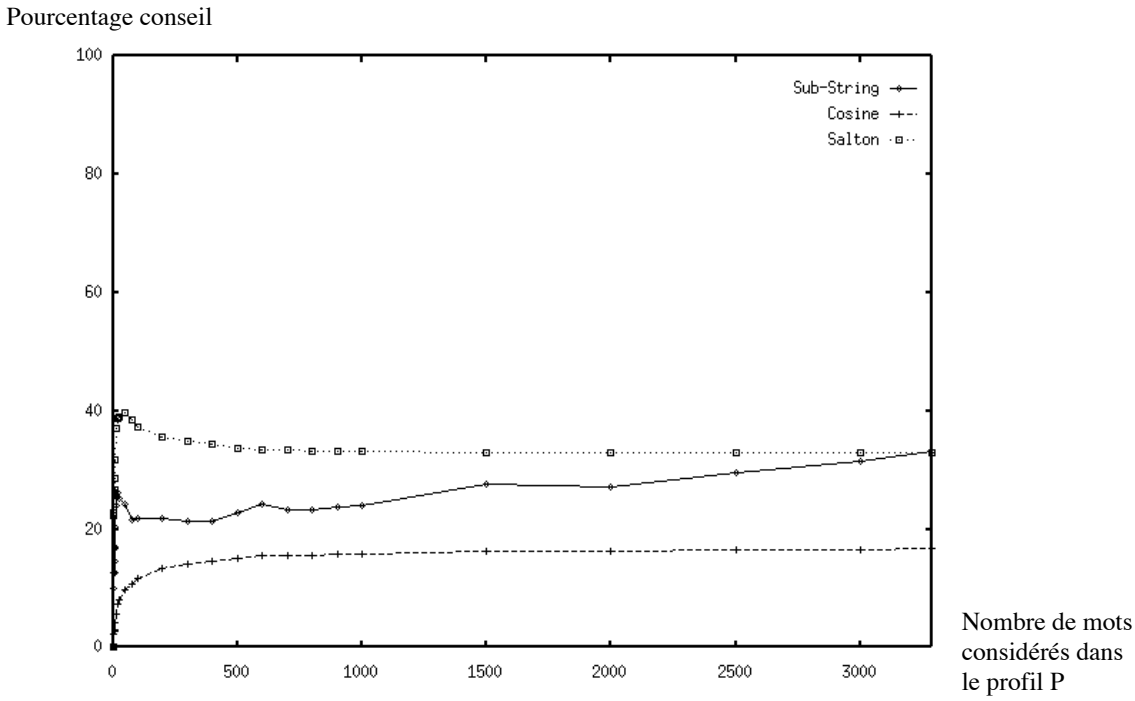


FIG. 7.10 - *Conseils pour le texte t_2 / Nombre de mots pris dans le profil P*

Des mesures du graphique 7.10, il ressort la même chose que pour un texte pertinent :

- *the sub-string measure* est bien affectée par la diminution de mots pris en compte,
- à l'inverse, *the cosine measure* et la mesure inspirée de Salton restent stables.

Mais ce qui intéressant se passe encore à l'origine (graphique 7.11) :

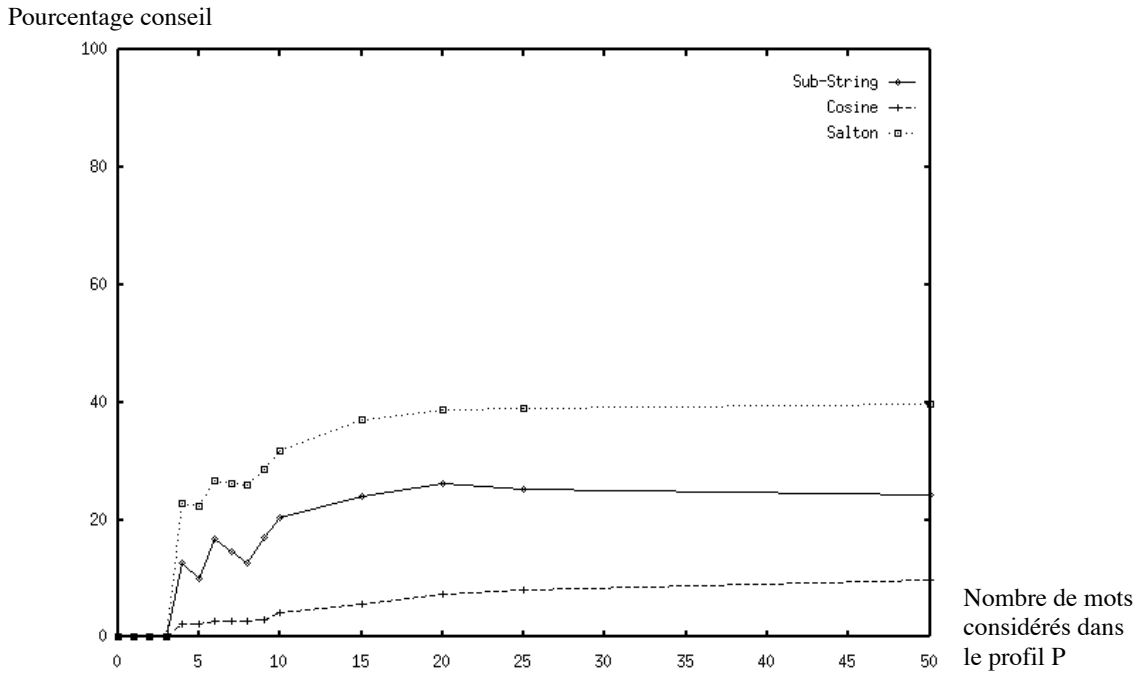


FIG. 7.11 - *Zoom sur le début du graphique précédent*

Lorsque le nombre de mots est grandement réduit, il résulte que :

- pour *the sub-string measure*, lorsque le nombre de mots est inférieur à 10-15, la courbe est discontinue, instable (dépend du texte conseillé),
- *the cosine measure* reste stable, légèrement décroissante jusque ~5 mots,
- la mesure inspirée de Salton obtient d'assez bons résultats, puisqu'après avoir augmentée, la courbe redescend à partir de 15 mots.

7.5.3 Conclusion

À la suite de ces différents tests de variation du nombre de mots dans le profil, on peut déduire que :

- le nombre de mots peut être réduit à 5 mots et obtenir de très bons résultats, mais ces résultats sont instables et dépendent du texte conseillé,

- le choix du nombre de mots le plus judicieux se situe entre 10 et 20 mots selon la stabilité recherchée,

- la mesure inspirée de Salton obtient encore les meilleurs résultats : pour un texte pertinent, le conseil augmente lorsqu'on diminue le nombre de mots; pour un texte non pertinent, le conseil diminue lorsqu'on diminue le nombre de mots (à partir de 20 mots).

7.6 Conseils / Nombre de mots dans les textes

La diminution du nombre de mots peut aussi s'effectuer sur les mots les plus occurrents du document :

7.6.1 Texte t_1 : pertinent

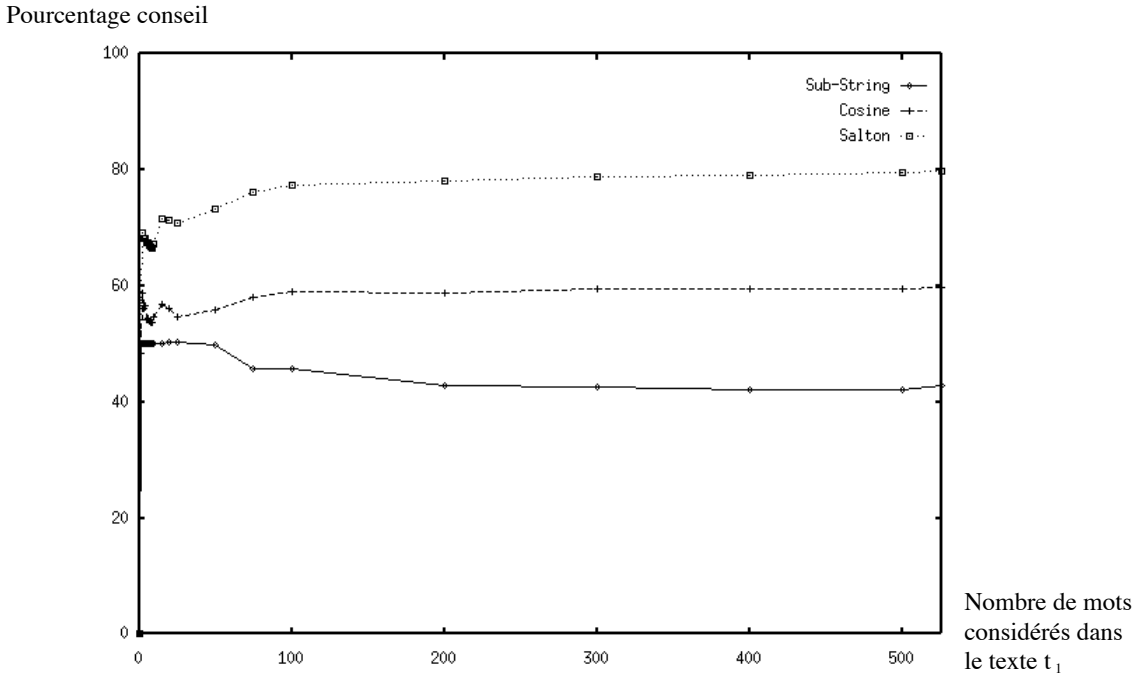


FIG. 7.12 - *Conseils pour le texte t_1 / Nombre de mots pris dans le texte t_1*

Les résultats observés sur le graphique 7.12 sont différents de ceux de la diminution de mots dans le profil P :

- toutes les courbes sont stables jusqu'à 100 mots,
- mais si l'on diminue encore, *the cosine measure* et la mesure inspirée de Salton enregistre une baisse, alors que *the sub-string measure* enregistre une hausse.

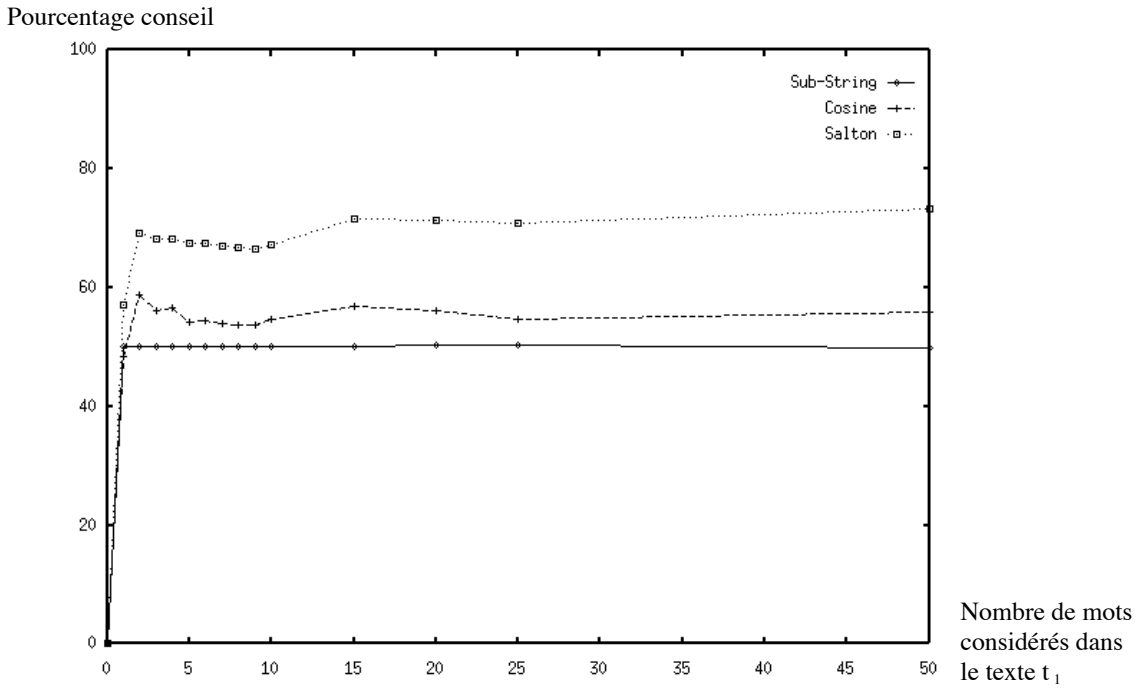


FIG. 7.13 - *Zoom sur le début du graphique précédent*

Et comme le montre le graphique 7.13, après ces baisses et cette hausse, les courbes sont stables.

7.6.2 Texte t_2 : non pertinent

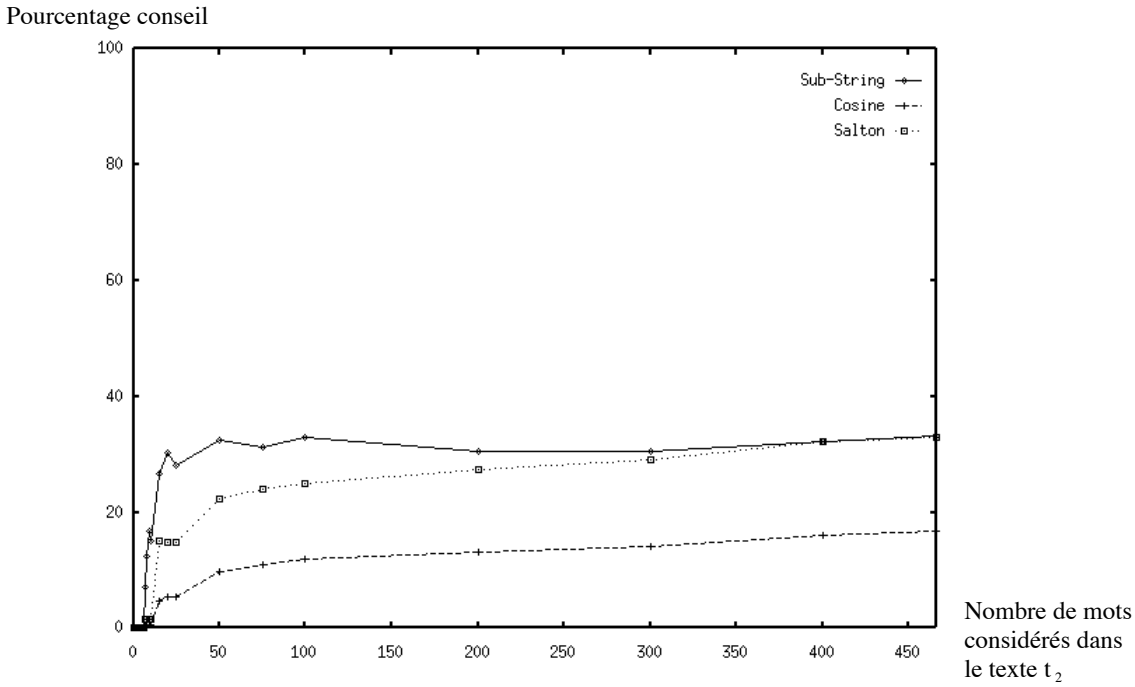


FIG. 7.14 - *Conseils pour le texte t_2 / Nombre de mots pris dans le texte t_2*

Pour un texte non pertinent, les conclusions sont presque identiques que pour le graphique 7.12 :

- légère diminution de *the cosine measure* et de la mesure inspirée de Salton jusqu'à 50 mots, ensuite la diminution s'accentue,
- stabilité de *the sub-string measure* jusqu'à 100 mots, augmentation entre 20 et 25 mots, puis diminution.

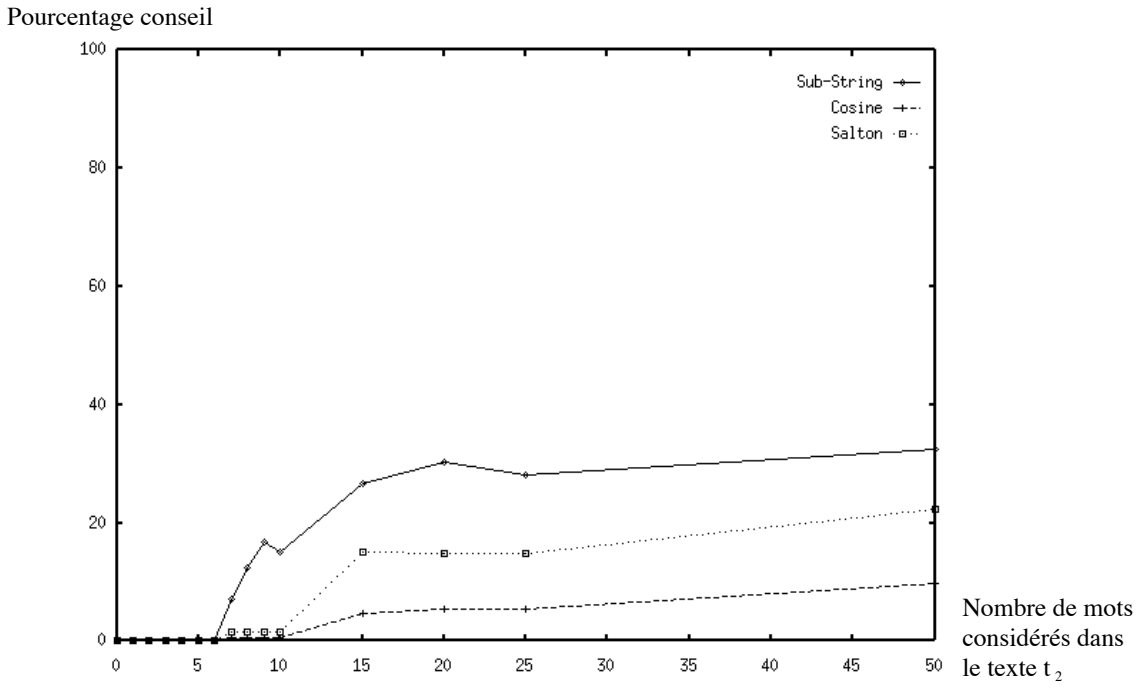


FIG. 7.15 - *Zoom sur le début du graphique précédent*

Par contre, à l'origine, on observe :

- une stabilité si l'on fait diminuer le nombre de mots jusqu'à 15,
- en deça, les résultats sont excellents (jusque 0%) mais sujet à l'instabilité.

7.6.3 Conclusion

À la suite de ces différents tests de variation du nombre de mots dans les textes, on peut déduire que :

- le nombre de mots peut être réduit à 5 mots et obtenir de très bons résultats, mais ces résultats sont instables et dépendent du texte conseillé,
- le choix du nombre de mots le plus judicieux se situe entre 10 et 20 mots selon la stabilité recherchée,
- la mesure inspirée de Salton obtient encore les meilleurs résultats : pour un texte pertinent, le conseil varie seulement de 5 à 10%; pour un texte non pertinent, le conseil peut diminuer jusque 0%.

7.7 Conseils : Performance en temps / Nombre de mots

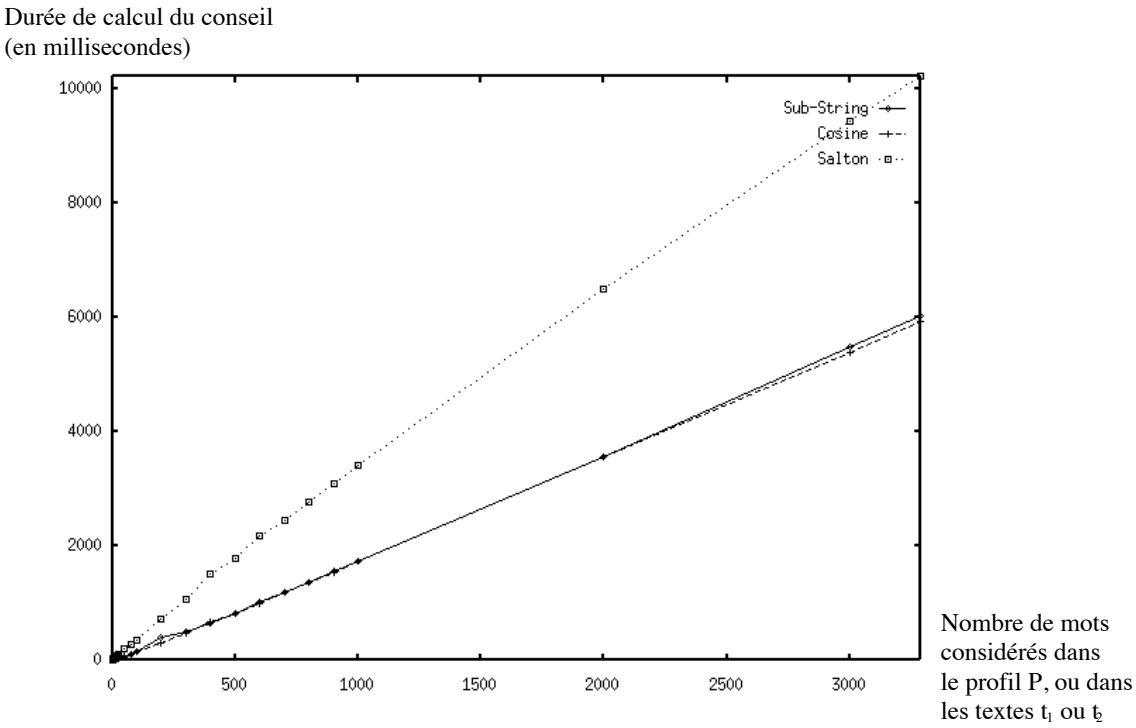


FIG. 7.16 - *Durée de calcul des différents conseils*

Le graphique 7.16 montre que la durée de calcul d'un conseil est linéaire avec le nombre de mots considérés. Comme dans les deux chapitres précédents, il a été démontré que le pourcentage conseil restait significatif malgré un baisse du nombre de mots, les différents calculs de conseil sont réalisés avec 25 mots pour une meilleure performance.

7.8 Conclusions des résultats

- Apprentissage :
  - l'apprentissage est stable en général à partir de 10-15 textes appris (quelque soit l'apprentissage considéré),
  - l'apprentissage génétique obtient les meilleurs conseils mais c'est aussi le plus lent,
  - l'apprentissage statistique obtient de très bon résultats et est surtout très rapide.
- Le nombre de mots considérés dans le profil ou dans le texte peut être grandement réduit :
  - avec 5 mots, très bon résultats ou très mauvais selon le texte : instabilité,
  - entre 10 et 25 mots, bon résultats, stables,
  - l'instabilité est d'autant plus grande si l'on considère en même temps 5 mots dans le profil et 5 mots dans le texte,
  - par contre en considérant 25 mots dans le profil et dans le texte, on obtient une bonne stabilité.
- La mesure inspirée de Salton donne les meilleurs résultats :
  - continue (sauf au démarrage),
  - le pourcentage reflète bien le conseil,
  - écart important entre les conseils pour un texte pertinent et un texte non pertinent,
  - petite variation de 5 à 10% sur le conseil selon le nombre de mots,
  - stable rapidement.

Chapitre 8

Conclusion

L'objectif de ce stage était de lancer les bases d'une application de filtrage qui soit adaptative, performante et évolutive. Au vu des résultats, les deux premiers critères sont parfaitement satisfaits.

Du fait de la méthodologie et du codage objet, l'évolutivité de l'application est aussi assurée, et de nombreuses extensions, utilisations peuvent être envisageables dans le futur :

- passage aisé à une analyse linguistique plus poussée : actuellement l'agent de filtrage travaille avec des textes analysés au niveau des mots, mais une analyse linguistique plus poussée met en évidence les relations syntaxiques entre les mots (sujet, verbe, complément d'objet, ...). Ces relations peuvent être considérées comme des mots mais avec une pondération plus importante correspondant aux mots qu'elle relie,
  exemple : chat_sujet_ronronne, considéré comme un mot de pondération = pondération de chat + pondération de ronronne = 4 + 3 = 7,
- passage aussi simple à une analyse sémantique : remplacement des mots par des concepts,
- passage au multimédia : indexation permettant la reconnaissance de sons, images dans les documents, et intégrant une extraction de signatures permettant des comparaisons et de l'apprentissage,
- utilisation de la conception objet pour ajouter des nouveaux modules de profils utilisateur et d'apprentissage basés sur les réseaux de neurones,
- utilisation de l'architecture en agent pour faire de la coopération entre agents, exemple : lorsque l'utilisateur trouve un texte intéressant mais qu'il n'a pas

de centre d'intérêt correspondant : il demande aux autres agents s'ils ont un centre d'intérêt qui est adéquat,

- utilisation du profil utilisateur et de l'apprentissage dans d'autres parties de l'étude EXIST, par exemple, au moment de la requête sur le moteur de recherche (Internet) :
  - apprentissage des requêtes de l'utilisateur pour déterminer ses nouveaux centres d'intérêts, ou affiner ses intérêts existants,
  - utilisation du profil utilisateur pour afficher des publicités pertinentes pendant la phase de recherche.

Toutes ces évolutions possibles montrent bien que le filtrage est un problème d'actualité, qui en s'adaptant au multimédia, saura trouver sa place sur le marché et dans la *Hi-Speed* stratégie d'Alcatel.

Annexe A

Exemple d'application d'un algorithme génétique

Considérons par exemple le problème "OneMax" où l'espace S des solutions est celui des chaînes binaires de $l = 6bits$, et où la fonction f à maximiser est définie par :

$f(s)$ = nombre de bit à 1 dans la chaîne s

Ainsi, on a $f(000000) = 0$, $f(010111) = 4$ et $f(111111) = 6$. La solution recherchée par l'algorithme est donc $s^* = 111111$.

Les algorithmes génétiques utilisent une population P de n solutions où n est un entier fixé. La population est initialisée aléatoirement au début de l'algorithme, en générant n points de S puis en évaluant leur qualité avec f. Dans l'exemple considéré, posons $n = 5$.

A.1 Tirage aléatoire

Aléatoirement, on obtient $P(1)$:

$P(1)$

s	$f(s)$
011011	4
100101	3
001010	2
001101	3
110100	3

Ensuite, l'algorithme génère la population $P(t+1)$ à partir de $P(t)$ en utilisant trois opérations :

1. la sélection, qui sélectionne une première population intermédiaire P_s de n solutions à partir de $P(t)$,

2. la recombinaison, qui recombine entre elles les solutions de $P(t)$ pour former une seconde population intermédiaire P_r de n solutions également,

3. la mutation, qui modifie aléatoirement les solutions de P_r pour donner $P(t+1)$.

A.2 Première étape : la sélection

Selon la loi

$$P_{select}(s_i) = \frac{f(s_i)}{\sum_{j=1}^{n} f(s_j)}$$

Dans notre exemple, les probabilités de sélection sont donc :

$P(1)$

s	$f(s)$	$P_{select}(s)$
011011	4	$\frac{4}{4+3+2+3+3} \simeq 27\%$
100101	3	20%
001010	2	13%
001101	3	20%
110100	3	20%

Un exemple de tirage aléatoire peut donner le résultat :

P_s

s	$f(s)$
011011	4
100101	3
011011	4
001101	3
110100	3

Notons que la sélection duplique les meilleurs et élimine les plus mauvais.

A.3 Seconde étape : la recombinaison

La deuxième population intermédiaire P_r est obtenue en recombinant deux à deux certaines solutions de P_s. Pour sélectionner des couples de solutions, la population P_s est parcourue et chaque solution a une probabilité p_{cross} (de l'ordre de 0.8) d'être sélectionnée pour la recombinaison. Les solutions sélectionnées sont couplées deux à deux dans l'ordre d'apparition dans P_s. Les couples ainsi formés subissent ensuite un opérateur de croisement à un point : cet opérateur reçoit en entrée un couple de solutions, sélectionne aléatoirement un point de coupure entre 2 et l, puis échange tous les bits à partir du point du coupure. Les descendants ainsi engendrés remplacent leurs parents dans la population, formant la population P_r.

Dans l'exemple, les solutions sélectionnées peuvent être les suivantes :

Solutions sélectionnées

s	$f(s)$
011011	4
100101	3
011011	4
110100	3

croisement :

$$\begin{array}{lcl} 0110|11 & \rightarrow & 0110|01 \\ 1001|01 & \rightarrow & 1001|11 \end{array}$$

$$\begin{array}{lcl} 01|1011 & \rightarrow & 01|0100 \\ 11|0100 & \rightarrow & 11|1011 \end{array}$$

On obtient donc la population intermédiaire P_r :

P_r

s	$f(s)$
011001	?
100111	?
010100	?
001101	3
111011	?

A.4 Troisième étape : la mutation

La population finale $P(t+1)$ est calculée en appliquant un opérateur de mutation à P_r. Cet opérateur considère toute la population P_r comme une seule chaîne binaire et inverse un bit avec une probabilité p_{mut} en générale très faible (de 0.001 à 0.01).

Dans notre exemple, avec $p_{mut} = 0.05$:

$P(2)$

s	$f(s)$
1 11001	4
1001 0 1	3
01010 1	3
001101	3
111011	5

Ensuite l'algorithme reprend à la première étape, ou s'arrête. Plusieurs critères de terminaisons peuvent être utilisés :

- lorsque la qualité de la meilleure solution dépasse un certain seuil,
- lorsqu'un certain nombre de générations se sont succédées.

Bibliographie

Documents internes de références :

[1] Jean-Paul Rossazza **Rapport technique :** *"Dossier d'Analyse des Besoins en Renseignement Documentaire"* UAR/RT/95/119/V1.2

[2] Jean-Paul Rossazza **Rapport technique :** *"Dossier d'État de l'Art"* UAR/RT/95/143/V1.2

[3] Jean-Paul Rossazza **Rapport technique :** *"Rapport complémentaire sur l'État de l'Art"* UAR/RT/95/234/V1.1

[4] Jean-Paul Rossazza **Rapport technique :** *"Dossier de Spécifications Logicielles"* UAR/RT/95/206/V1.1

[5] Jean-Paul Rossazza & Eric Colaviti **Rapport technique :** *"Première spécifications EXIST 96"* UAR/C/96/0110/V1

[6] Cyrille Herbon **Rapport de stage :** *"Outil de recherche documentaire avancée sur Internet"* UAR/C/96/0136/V1

[7] Jean-Paul Rossazza & Eric Colaviti **Rapport technique :** *"Deuxièmes spécifications EXIST 96"* UAR/C/96/0334/V1

Documents externes :

Agents :

[ARWIO] Pattie Maes *"Agents that Reduce Work and Information Overload"* Communications of the ACM, Vol.37, N°7, pp.31-40, 146, ACM Press July 1994 http://pattie.www.media.mit.edu/people/pattie/CACM-94/CACM-94.p1.html

[DAIIS] Michael N. Huhns & Munindar P. Singh *"Distributed Artificial Intelligence for Information Systems"* CKBS-94 Tutorial, June 15, University of Keele, UK 1994

[MAAA] Pattie Maes *"Modeling Adaptive Autonomous Agent"* Artificial Life Journal, edited by C. Langton, Vol.1, N°1& 2, pp.135-162, MIT Press 1994 http://pattie.www.media.mit.edu/people/pattie/alife-journal.ps

[MAIMRM] Anne-Claire Boury-Brisset & Bernard Moulin *"Un modèle d'agent intégrant des mécanismes de raisonnement multiples"* Revue d'intelligence artificielle, Vol.11, N°1, pp.73-107 1997

[SAO] Hyacinth S. Nwana *"Software Agents: An Overview"* Knowledge Engineering Review, Vol.11, N°3, pp.1-40 September 1996 http://www.sce.carleton.ca/netmanage/docs/AgentsOverview/ao.html

Recherche documentaire :

[DATR] G. Salton *"Developments in Automatic Text Retrieval"* Science, Vol. 253, Article pp.974-980 30 August 1991

[EANDRF] Mark D. Dunlop *"The Effect of Accessing Nonmatching Documents on Relevance Feedback"* ACM Transactions on Information Systems, Vol.15, N°2, pp.137-153 April 1997

[EBIR] G. Salton & E. A. Fox & H. Wu *"Extended Boolean Information Retrieval"* Communication of the ACM, Vol.26, N°11, pp.1022-1036 November 1983

[GTMIR] G. Salton *"Global Text Matching for Information Retrieval"* Science, Vol. 253, Reports pp.1012-1015 30 August 1991

[IMIR] G. Salton & M. J. McGill *"Introduction to Modern Information"* McGraw-Hill 1983

[IRUSC] David A. Hull **Doctor of Philosophy Report :** *"Information Retrieval Using Statistical Classification"* Standford University November 1994

[RIW] François Jacquenet *"Recherche d'Informations sur le Web : Un Nouveau Challenge pour l'I.A."* LLIA N°123, pp.111-114, Interfaces 97 Mai 1997

[URFRIS] Nicholas J. Belkin & Colleen Cool & Jürgen Koenemann & Kwong Bor Ng & Soyeon Park *"Using Relevance Feedback and Ranking in Interactive Searching"* School of Communication, Information & Library Studies, Rutgers University

Filtrage documentaire :

[ACIF] Curt Stevens *"Automating the Creation of Information Filters"* Communication of the ACM, Vol.35, N°12, pp.48 December 1992

[APDMI] Shoshana Loeb *"Architecting Personalized Delivery of Multimedia Information"* Communication of the ACM, Vol.35, N°12, pp.39-48 December 1992

[BCNFS] Fredrik Kilander *"A Brief Comparison of News Filtering Software"* IntFilter, K2Lab, Stockholm University and the Royal Institue of technology, Electrum 230, S-164 40 KISTA, Sweden June 19, 1996 http://www.cs.kun.nl/is/research/filter/literature/Comparions.ps

[BIS] H. P. Luhn *"A Business Intelligence System"* IBM Journal of Research and Development, Vol.2, N°4, pp.314-319 October 1958

[CAIF] Paul E. Baclace *"Competitive Agents for Information Filtering"* Communication of the ACM, Vol.35, N°12, pp.50 December 1992

[CCSUNA] Fredrik Kilander *"Comparisons of the Cosine Measure and Sub-String Indexing on Usenet News Articles"* IntFilter, K2Lab, Stockholm University and the Royal Institue of technology, Electrum 230, S-164 40 KISTA, Sweden January 4, 1995 http://www.cs.kun.nl/is/research/filter/literature/T.ps

[CFTF] Douglas W. Oard & Gary Marchionini *"A Conceptual Framework for Text Filtering"* EE-TR-96-25 CAR-TR-830 CLIS-TR-96-02 CS-TR-3643 May 1996 http://www.ee.umd.edu/medlab/filter/

[CFWIT] David Goldberg, David Nichols, Brian M. Oki & Douglas Terry *"Using Collaborative Filtering to Weave an Information Tapestry"* Communication of the ACM, Vol.35, N°12, pp.61-70 December 1992

[DA] T. F. Bowen, G. Gopal, G. Herman, T. Hickey, K. C. Lee, W. H. Mansfield, J. Raitz & A. Weinrib *"The Datacycle Architecture"* Communication of the ACM, Vol.35, N°12, pp.71-81 December 1992

[EIOLN] Paul S. Jacobs & Lisa F. Rau *"SCISOR: Extracting Information from On-Line News"* Communication of the ACM, Vol.33, N°11, pp.88-97 November 1990

[EJ] Peter J. Denning *"Electronic Junk"* Communication of the ACM, Vol.25, N°3, pp.163-165 March 1982

[FATAS] David Hull (XEROX) *"Filtrage Adaptatif des Textes par l'Apprentissage Statistique"* Journée ATALA, Résumé et Filtrage automatique de textes 24 Mai 1997

[GBUMF] Jussi Karlgren & Kristina Höök & Ann Lantz & Jacob Palme & Daniel Pargman *"The glass box user model for filtering"* Departments of Computer and Systems Sciences, Computational Linguistics, and Psychology, Stockholm University February 1994 http://www.cs.kun.nl/is/research/filter/literature/Glassbox1.1.ps

[IACPSIS] Gerhard Fischer & Curt Stevens *"Information Access in Complex, Poorly Structured Information Spaces"* Appeared in Human Factors in Computing Systems, CHI'91 Conference Proceedings @1991 ACM

[IDFMS] Jacob Palme & Jussi Karlgren & Daniel Pargman *"Issues when designing filters in messaging systems"* Department for Computer and Systems Sciences, Stockholm University http://www.cs.kun.nl/is/research/filter/literature/IssuesDesFilter.ps

[IFBKO] Ann Lantz & Fredrik Kilander *"Intelligent Filtering; Based on Keywords Only?"* IntFilter, K2Lab, Stockholm University and the Royal Institue of technology, Electrum 230, S-164 40 KISTA, Sweden http://www.cs.kun.nl/is/research/filter/literature/chi95.ps

[IFIRSC] Nicholas J. Belkin & W. Bruce Croft *"Information Filtering and Information Retrieval: Two Sides of the Same Coin?"* Communication of the ACM, Vol.35, N°12, pp.29-38 December 1992

[IFUATR] M. Morita & Y. Shinoda *"Information Filtering based on User behavior Analysis and best match Text Retrieval"* In Proceedings of the Seventeenth Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval, pp.272-281, Springer-Verlag 1994

[IIF] Fredrik Kilander & Eva Fåhræus & Jacob Palme *"Intelligent Information Filtering"* IntFilter, K2Lab, Stockholm University and the Royal Institue of technology, Electrum 230, S-164 40 KISTA, Sweden February 17, 1997

[IISS] Thomas W. Malone, Kenneth R. Grant, Franklyn A. Turbak, Steven A. Brobst & Mickael D. Cohen *"Intelligent Information Sharing Systems* Communication of the ACM, Vol.30, N°5, pp.390-402 May 1990

[LAPIF] Beerud Dilip Sheth **Master Degree Report :** *"A Learning Approach to Personalized Information Filtering"* Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology January 14, 1994

[LSIIF] Peter W. Foltz *"Using Latent Semantic Indexing for Information Filtering"* In R. B. Allen (Ed.) Proceedings of the Conference on Office Information Systems, Cambridge, MA, pp.40-47 April 1990 http://www-psych.nmsu.edu/~pfoltz/cois/filtering-cois.html

[MCF] Fredrik Kilander *"Message Classification and Filtering"* IntFilter, K2Lab, Stockholm University and the Royal Institue of technology, Electrum 230, S-164 40 KISTA, Sweden January 4, 1995 http://www.cs.kun.nl/is/research/filter/literature/aifiltr.ps

[MUIIF] Irene Stadnyk & Robert Kass *"Modeling User's Interets in Information Filters"* Communication of the ACM, Vol.35, N°12, pp.49-50 December 1992

[NLUIFS] Ashwin Ram *"Natural Language understanding for Information-Filtering Systems"* Communication of the ACM, Vol.35, N°12, pp.80-81 December 1992

[OFTC] Donna Harman *"Overview of the First TREC Conference"* In Robert Korfhage, Edie Rasmussen, & Peter Willet, editors, Proceedings of the Sixteenth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp.36-47, ACM June 1993

[PEFNA] Fredrik Kilander & Eva Fåhræus & Jacob Palme *"PEFNA - The Private Filtering News Agent"* IntFilter, K2Lab, Stockholm University and the Royal Institue of technology, Electrum 230, S-164 40 KISTA, Sweden February 17, 1997

[PIDAM] Peter W. Foltz & Susan T. Dumais *"Personalized Information Delivery: An Analysis of Information Filtering Methods"* Communication of the ACM, Vol.35, N°12, pp.51-60 December 1992

[RBMFS] Stephen Pollock *"A Ruled-Based Message Filtering System"* ACM Transactions on Office Information Systems, Vol.6, N°3, pp.232-254 July 1988

[STWAID]
Tak W. Yan & Hector Garcia-Molina *"SIFT - A Tool for Wide-Area Information Dissemination"* Department of Computer Science, Stanford University February 16, 1995 http://www.cs.kun.nl/is/research/filter/literature/sift.ps

[TRSIG] Edward M. Housman **Technical Report SIG/SDI-1 :** *"Survey of current systems for selective dissemination of information"* American Society for Information Science Special Interest Group on SDI, Washington, DC June 1969

Apprentissage :

[AGA] Gilles Venturini *"Algorithmes génétiques et apprentissage"* Revue d'intelligence artificielle, Vol.10, N°2-3, pp.345-387 1996

[GACL] Kenneth A. De Jong & William Spears & Diana F. Gordon *"Using Genetic Algorithms for Concept Learning"* In J. Grefenstette, editor, special issue

on genetic algorithms for Machine Learning Journal, v13, pp.161-188. Kluwer Academic 1993 ftp://ftp.aic.nrl.navy.mil/pub/spears/mlj93.ps.Z

[GASCL] William Spears & Kenneth A. De Jong *"Using Genetic Algorithms for Supervised Concept Learning"* In N.G. Bourbakis, editor, Artificial Intelligence Methods and Applications, World Scientific. [Based on previous IEEE90 AI Tools paper] 1992 ftp://ftp.aic.nrl.navy.mil/pub/spears/world92.ps.Z

[GCCLA] Mitchell A. Potter *"A Genetic Cascade-Correlation Learning Algorithm"* In Proceedings of COGANN-92 International Workshop on Combinations of Genetic Algorithms and Neural Networks, pp.123-133, Baltimore, MD. IEEE Computer Society Press 1992 http://www.cs.gmu.edu/research/gag/papers/gencascor.ps

[OEC] William M. Spears & Kenneth A. De Jong & Thomas Bäck & David B. Fogel & Hugo de Garis *"An Overview of Evolutionary Computation"* In the proceedings of the European Conference on Machine Learning, pp.442-449 1993 ftp://ftp.aic.nrl.navy.mil/pub/spears/ecml93.ps.Z

[SAGARN] Patrick Gallinari & Olivier Gascuel *"Statistiques, apprentissage et généralisation. Application aux réseaux de neurones"* Revue d'intelligence artificielle, Vol.10, N°2-3, pp.285-343 1996

URL :

Agents :

[AI] *"Agents info." :* http://www.cs.bham.ac.uk/~amw/agents/index.html

[IARL] *"Intelligent Agent Resource Links" :*
http://www.cs.mu.oz.au/~leon/agentlinks.html

[IATP] *"Intelligent Agents: Theory and Practice" :*
http://www.doc.mmu.ac.uk/STAFF/M.Wooldridge/ker95/ker95.html

[ISA] *"Intelligent Software Agents" :* http://www.sics.se/ps/abc/survey.html

[MITML] *"MIT Media Lab, Software Agents Group" :*
http://agents.www.media.mit.edu/groups/agents/
"Pattie Maes'Home Page" :
http://pattie.www.media.mit.edu/people/pattie/

[WJIAP] *"Willi jamison's Intelligent Agent Page" :*
http://www.cat.syr.edu/~wcjamiso/ia/source.html

Filtrage :

[IFRF] *"Information Filtering Refs" :*
http://www.cs.kun.nl/is/research/filter/references.html

[IFRS] *"Information Filtering Resources" :* http://www.ee.umd.edu/medlab/filter/

[JMRTF] *"JMs Research Text Filtering":* http://129.79.33.62/jmdocs/restext.html

[SGHP] *"Steve Gant's Home Page" :* http://ils.unc.edu/gants/

[TREC] *"TREC Home Page" :* http://www-nlpir.nist.gov/TREC/

Applications de filtrage de l'e-mail ou des news :

[BR] *"Browse" :* http://www.ee.umd.edu/medlab/filter/browse.tar.Z

[GL] *"GroupLens" :* http://www.cs.umn.edu/Research/GroupLens/

[IS] *"InfoScan" :*
http://www.machinasapiens.com/english/products/infoscan/infoscanang.html

[LU] *"Lurker" :* ftp://ftp.ph.utexas.edu/pub/perl/lurker.1.1.tar

[MA] *"Mailagent" :* ftp://ftp.foretune.co.jp/pub/network/mail/mailagent/

[MF] *"Mailfilt" :* http://www.nmt.edu/~mfisk/style.cgi?unixtools/mailfilt.html

[MX] *"MAXIMS" :*
ftp://ftp.media.mit.edu/pub/agents/interface-agents/MAXIMS

[NC] *"NewsClip" :* http://www.clarinet.com/newsclip.html

[PE] *"PEFNA" :* http://www.dsv.su.se/~fk/if_Doc/IntFilter.html

[RA] *"RAMA" :* ftp://ftp.cs.pdx.edu/pub/faculty/jrb/rama/

[SI] *"SIFT" :* ftp://db.stanford.edu/pub/sift/

[SIM] *"Sift-Mail" :* http://www.island-resort.com/sm.htm

[SM] *"SMART" :* ftp://ftp.cs.cornell.edu/pub/smart

Apprentissage :

[EEAAX] *"Équipe Évolution Artificielle et Apprentissage de l'X" :*
http://blanche.polytechnique.fr/www.eeaax/eeaax.html
http://blanche.polytechnique.fr/www.eeaax/eeaax_french.html

[GAGPL] *"GA Group Publications List" :*
http://www.cs.gmu.edu/research/gag/pubs.htmlb

[RNTM] *"Les réseaux de neurones Table des matières" :*
http://www.worldnet.fr/~willy/neural/neutm.html