République Algérienne Démocratique et Populaire.
وزارة التعليم العالي والبحث العلمي
Ministère de l'Enseignement Supérieur et de la Recherche Scientifique.
Université des Sciences et de la Technologie d'Oran –USTO Mohamed Boudiaf.

Faculté des Sciences – Département d'Informatique.

Spécialité : Informatique ***Option :*** Reconnaissance de Formes et Intelligence Artificielle

Thèse

Présentée par :

Mr Abderrezak BRAHMI

Pour l'obtention du diplôme de

Doctorat en Sciences en Informatique

Thème

Contribution à la Recherche Intelligente sur le Web :

Indexation Sémantique des Textes Non-Structurés

Soutenu le : 17 / 04 / 2013

Devant le jury composé de :

Qualité	Nom et prénoms	Grade	Etablissement
Président	Mohamed BENYETTOU	Professeur	USTO - Oran
Rapporteur	Abdelkader BENYETTOU	Professeur	USTO - Oran
Examinateur	Bouziane BELDJILALI	Professeur	Univ. Oran
Examinateur	Okba KAZAR	Professeur	Univ. Biskra
Examinateur	Ahmed LEHIRECHE	Maître de conf."A"	Univ. Sidi-Belabes
Examinateur	Lynda ZAOUI	Maître de conf."A"	USTO - Oran

Année universitaire 2012/2013

République Algérienne Démocratique et Populaire.
وزارة التعليم العالي والبحث العلمي
Ministère de l'Enseignement Supérieur et de la Recherche Scientifique.
Université des Sciences et de la Technologie d'Oran –USTO Mohamed Boudiaf.

Faculté des Sciences – Département d'Informatique.

Spécialité : Informatique ***Option :*** Reconnaissance de Formes et Intelligence Artificielle

Thèse

Présentée par :

Mr Abderrezak BRAHMI

Pour l'obtention du diplôme de

Doctorat en Sciences en Informatique

Thème

Contribution à la Recherche Intelligente sur le Web :

Indexation Sémantique des Textes Non-Structurés

Soutenu le : 17 / 04 / 2013

Devant le jury composé de :

Qualité	Nom et prénoms	Grade	Etablissement
Président	Mohamed BENYETTOU	Professeur	USTO - Oran
Rapporteur	Abdelkader BENYETTOU	Professeur	USTO - Oran
Examinateur	Bouziane BELDJILALI	Professeur	Univ. Oran
Examinateur	Okba KAZAR	Professeur	Univ. Biskra
Examinateur	Ahmed LEHIRECHE	Maître de conf."A"	Univ. Sidi-Belabes
Examinateur	Lynda ZAOUI	Maître de conf."A"	USTO - Oran

Année universitaire 2012/2013

Dédicace

A mes parents, que Dieu les protège,

A ma très chère femme avec qui je continue de partager les joies et les chagrins,

A tous ceux qui ont prié Dieu pour ma réussite.

Abderrezak.

Remerciements

Je suis très reconnaissant au Professeur Abdelkader BENYETTOU pour m'avoir guidé dans la réalisation et la finalisation de cette thèse. Je tiens à lui exprimer ma profonde gratitude pour son soutien et sa patience. Je remercie également le Docteur Ahmed ECH-CHERIF avec qui j'ai réalisé une partie de ce travail.

Mes remerciements et mes sincères salutations s'adressent aux membres du jury qui m'ont honoré en prenant soin d'évaluer ce travail. En particulier, je remercie le Professeur Mohamed BENYETTOU pour avoir accepté de présider le jury de cette thèse. Je remercie également le Professeur Bouziane BELDJILALI, le Professeur Okba KAZAR, le Docteur Ahmed LAHIRECHE et le Docteur Lynda ZAOUI qui m'ont honoré en acceptant d'examiner le présent travail.

Je remercie également le Professeur Patrick GALLINARI, directeur du laboratoire d'informatique de Paris-6 en France, pour l'aide et l'accueil qu'il m'a réservés pendant mon séjour au LIP6.

Je tiens à remercier le Docteur Belkacem ATHAMENA, enseignant à l'université d'Al-Zaytoona en Jordanie, pour son aide précieuse et son accueil chaleureux pendant mon séjour à Amman.

Sans oublier ma très chère femme Hakima qui m'a soutenu tout au long de cette expérience aussi difficile qu'enseignante, mes remerciements s'adressent à tous ceux qui ont contribué de près ou de loin dans la réalisation de ce travail.

Résumé

Contribution à la Recherche Intelligente sur le Web :
Indexation Sémantique des Textes Non-Structurés.

Depuis sa promotion au grand public au début des années 1990, le Web a connu une croissance extraordinaire aussi bien dans son contenu que dans son utilisation. Malheureusement, la nature non-structurée, des larges volumes d'information disponibles sur la toile mondiale, a rendu de plus en plus difficile de cibler et retrouver l'information pertinente. Dans les systèmes classiques de recherche d'information, basés sur les mots-clés, les utilisateurs trouvent souvent des difficultés à exprimer leur besoin d'information. Parmi les nouvelles approches, qui ont été proposés pour promouvoir la recherche intelligente d'information, celle introduisant la dimension sémantique dans la modélisation des documents.

La recherche sémantique sur le Web peut être réalisée selon trois approches principales : (i) Organiser la recherche (indexation de documents et/ou analyse de requêtes) autour de connaissances conceptuelles (thésaurus ou ontologie), (ii) Utiliser un système d'annotation documenté par des experts ou une masse d'utilisateurs pour promouvoir la recherche collaborative, (iii) Développer des méthodes d'indexation sémantique des textes non-structurés. C'est dans cette dernière approche que la présente étude s'inscrit en essayant d'analyser les modèles de thèmes suivant trois axes d'investigation :

1. Quelle est la faisabilité d'utiliser un modèle de thème comme approche d'indexation sémantique des textes pour les tâches de recherche d'information ?
2. Comment évaluer et interpréter le modèle de thème pour l'analyse sémantique du contenu d'une collection ?
3. Dans quelle mesure peut-on appliquer les modèles de thème dans le texte non-structuré non-anglais (l'arabe comme exemple d'étude) ?

Comme contribution majeure dans cette étude, il intéressant de citer :

1. L'analyse et l'évaluation du modèle d'allocation latente de Dirichlet dans les tâches de recherche et de catégorisation des textes sur des corpus réels.
2. La proposition d'une nouvelle mesure, à base de la divergence de Kullback-Leibler, pour le paramétrage de l'apprentissage des thèmes dans une collection donnée.
3. Le développement d'un nouvel algorithme de stemming à base de lemme pour l'analyse et l'indexation du texte arabe.
4. L'élaboration de trois collections arabes, à base d'articles de presse relatifs à la période 2007-2010, pour les expérimentations de tâches de la recherche d'information.

Par ailleurs, les modèles de documents générés, par l'allocation latente de Dirichlet dans des espaces réduits de thèmes, ont été utilisés efficacement dans la catégorisation des textes et la recherche ad-hoc. En plus, nos travaux ont montré l'efficacité de considérer les aspects morphologiques et les variations typographiques dans l'indexation sémantique des langues hautement flexionnelles telles que l'arabe.

Mots clé :

Recherche d'information, indexation sémantique, modèle de thème, catégorisation des textes, analyse du texte arabe, mesures d'évaluation, collections de test.

Abstract

Contribution to Intelligent Web Search:
Semantic Indexing of Unstructured Texts.

Since its promotion to public in early 1990, the World Wide Web has experienced an extraordinary growth in both its content and its use. Nevertheless, the information available is mostly unstructured so that it is increasingly hard to find the relevant object. In classical information systems, the keyword-based search is not convenient to express a particular information need for a wide public of users. Introducing the semantic dimension in document modeling may improve efficiency intelligent retrieval systems.

Three main approaches can be followed for allowing semantic search in the Web: (i) organizing retrieval systems around conceptual knowledge, (ii) using an annotation system with information collected from experts or a large number of users, (iii) developing efficient methods for unstructured texts semantic indexing. The purpose of this study concerns this third approach by attempting to analyze topic models three investigation ways:

1. What is the feasibility of applying the topic models as texts semantic indexing for information retrieval tasks?
2. How to evaluate and interpret the topic model for semantic analysis of collection content?
3. How one can apply topic models on non-English unstructured text (Arabic as a study case)?

The main contribution in this study can be summarized as follows:

1. Analyzing and evaluating of the Latent Dirichlet Allocation model in text categorization and search tasks on several real word corpora.
2. Introducing a novel measure based on the *Kullback-Leibler* divergence for setting latent topics learning from a given collection.
3. Developing a new lemma-based stemming for Arabic text analysis and indexing.
4. Building three Arabic texts corpora from press articles published during 2007-2010 period for retrieval tasks evaluation.

Moreover, the generated document models, by the latent Dirichlet allocation in reduced topic spaces, have been successfully applied in search and text categorization. Furthermore, we have shown that semantic indexing, in highly inflected language such Arabic, is more effective when considering morphological features and typographical variations.

Keywords:

Information retrieval, semantic indexing, topic model, text categorization, Arabic text analysis, evaluation measures, test corpora.

ملخص

مساهمة في البحث الذكي على الويب:
الفهرسة الدلالية للنصوص غير المهيكلة

شهد الويب، منذ تعميمه للجمهور مع بداية التسعينات، تطورا مذهلا إن في محتواه أو في استخداماته. إلا أن الطبيعة غير المهيكلة لغالبية المعلومات المتوفرة على الشبكة العالمية جعلت من الصعوبة بمكان الوصول إلى المفيد منها. كما أن الأنظمة التقليدية للبحث عن المعلومات والقائمة على الكلمات المفتاحية، غالبا ما تعيق تعبير المستخدمين عن احتياجاتهم الاستعلامية. إن إدراج البعد الدلالي في نمذجة الوثائق النصية يعد أحد المقاربات الحديثة المقترحة لتحسين أنظمة البحث الذكي عن المعلومة.

يمكن تجسيد البحث الدلالي على الويب من خلال ثلاث مقاربات رئيسية: (أ) تنظيم البحث (فهرسةً للوثائق وتحليلا للاستعلامات) حول معارف مفاهيمية (مكنز أو أنطلوجيا)، (ب) استخدام نظام ترميزي مُوَثق من طرف خبراء أو كمّ هائل من المستخدمين في سياق بحث تعاوني، (ج) تطوير طرق الفهرسة الدلالية للنصوص غير المهيكلة. تندرج دراستنا في إطار هذه المقاربة الأخيرة من خلال تحليل نماذج المواضيع وفق ثلاثة محاور استقصائية:

1. ما جدوى استخدام نموذج المواضيع كطريقة للفهرسة الدلالية للنصوص في وظائف البحث عن المعلومات؟
2. كيف يمكن تقييم نموذج المواضيع بغرض التحليل الدلالي لمحتوى ذخيرة نصوص؟
3. كيف يتسنى لنا تطبيق نماذج المواضيع في النصوص غير المهيكلة غير الإنجليزية (العربية كمثال للدراسة)؟

اهتمت هذه الدراسة بالإجابة على عناصر الإشكالية المطروحة من خلال الإسهامات الرئيسية التالية:

1. تحليل نموذج التخصيص الكامن لديريكلي وتقييمه في وظائف البحث وتصنيف النصوص في ذخائر حقيقية.
2. اقتراح قياس جديد، على أساس تباعد *كولباك-ليبلير*، لضبط التلقين الآلي للمواضيع من ذخيرة معينة.
3. تطوير خوارزمية تشذيب جديدة على أساس المفردة المعجمية من أجل تحليل النص العربي وفهرسته.
4. إنشاء ثلاث ذخائر عربية للمقالات الصحفية المنشورة خلال الفترة 2007-2010، لاستخدامها في تقييم وظائف البحث عن المعلومات.

إضافة إلى أنه أمكن استخدام نماذج الوثائق، المولدة بواسطة التخصيص الكامن لديريكلي في فضاء مختصر من المواضيع، بفعالية في وظائف البحث وتصنيف النصوص. إن الدراسة الحالية تُظهر أهمية اعتبار جوانب البناء اللغوي وتغيرات الكتابة في الفهرسة الدلالية للغات ثرية الاشتقاق والصرف كالعربية.

الكلمات المفتاحية:

البحث عن المعلومات، الفهرسة الدلالية، نموذج المواضيع، التصنيف الآلي للنصوص، تحليل النص العربي، قياسات التقييم، ذخائر الاختبار.

Sommaire

Liste des Figures

Liste des Tableaux

Liste des Abréviations

AFP-22k	Corpus de (22.135) article de agence France-Presse	
BBw	Analyseur du texte arabe à base des ressources linguistiques de Buckwalter	
BKL	Mesure combinée Brahmi à base de la divergence de Kullback-Leibler	
CACM	Corpus de (3.204) résumés des articles publiés dans le journal *ACM*	*Corpus from the Association of Computing Machinery journal*
CF	Corpus de (1.239) résumés scientifiques sur la fibrose kystique	*Cystic-Fibrosis corpus*
Ech-11k	Corpus de (11.313) articles arabes du quotidien algérien Echorouk	
ICF	Facteur de compression d'index	*Index compression factor*
IDF	Fréquence inversé des documents	*Inverted document frequency*
ISRI	Algorithme de stemming léger du texte arabe	*Arabic stemmer from The Information Science Research Institute, University of Nevada, Las Vegas, USA*
Khoja	Algorithme de racinisation du texte arabe de Shereen Khoja	
LDA	Allocation latente de Dirichlet	*Latent Dirichlet allocation*
LSI	Indexation sémantique latente	*Latent semantic indexing*
pLSI	Indexation sémantique latente probabiliste	*Probabilistic latent semantic indexing*
OI	Index de sur-stemming	*Over-stemming index*
QE	Extension de la requête	*Query expansion*
RI (IR)	Recherche d'information	*Information retrieval*
RSV	Valeur (ou score) de pertinence	*Retrieval Status Value*
Rtr-41k	Corpus de (41.251) articles arabes de l'agence de presse internationale Reuters	
SRI	Système de recherche d'information	*Information retrieval system*
SVM	Séparateur à vaste marge	*Support vectors machine*
TALN (NLP)	Traitement automatique du langage naturel	*Natural language prcessing*
TF	Pondération par fréquence des termes	*Term frquency*
TREC	Conférence de recherche d'information	*Text retrieval conference*
UI	Index de sous-stemming	*Under-stemming index*
Xnh-36k	Corpus de (36.696) articles arabes de l'agence de presse chinoise Xinhua	

Chapitre 1 :
INTRODUCTION GENERALE

1 Motivation

Depuis sa promotion au grand public au début des années 1990, le Web a connu une croissance impressionnante aussi bien dans son contenu que dans son utilisation. En 2010, le volume d'information sur Internet dépassait déjà les 5 millions de Téraoctets[1]. Si le cerveau humain est estimé à contenir une moyenne de 5 Téraoctets d'information, nous aurions besoin d'un million de personnes pour stoker l'Internet. En termes de support physique, il nous faudrait plus d'un milliard de DVD pour la même fin. L'humanité n'a jamais connu autant d'information facilement accessible et partagée.

Malheureusement, la nature non-structurée, des larges volumes d'information disponibles sur la toile mondiale, a rendu de plus en plus difficile aux utilisateurs de cibler et retrouver l'information pertinente. A titre d'indication, la production mondiale d'information (en volume) dans les bases structurées n'a augmenté que de 4% en 2006 et ne représentait qu'environ 10% des données, a contrario, le reste (90%) était des données non structurées et dont l'augmentation enregistrait 6400% par an [Bonny et Garnier, 2008]. Cependant, nous sommes loin de satisfaire l'accès utile à l'information disponible. En 2010, le plus large index prétendu être détenu par le géant de la recherche sur le Web (Google), ne répertoriait que 200 Téraoctets de son contenu soit 0,004% de tout l'Internet.

D'un point de vue organisationnel, l'information numérique sur le Web peut être classée en trois catégories :

[1] Source : *Internet World Stats* http://www.internetworldstats.com/stats.htm

- Information structurée : respectant un modèle de donnée et une organisation formelle en table et généralement associée à des bases de données relationnelles.
- Information semi-structurée : respectant une certaine organisation sans aucune structure formelle de tables ou de modèle de base de données classiques. Les données semi-structurées sont basées sur un système de balisage (ou de marquage) pour séparer les entités sémantiques et appuyer leur hiérarchie. La famille des langages XML est la forme la plus appropriée pour ce type de données.
- *Information non-structurée : référant à toute information n'ayant aucun modèle de données prédéfini structurant sa composition. Un contenu textuel brut d'un document ou une image bitmap constituent des exemples types.*

Nombreuses sont les techniques de recherche d'information qui ont été développées pour traiter le problème de recherche d'information (RI) dans des collections de textes non-structurés. Les modèles classiques de recherche d'information sont basés sur les mots-clés. Ils utilisent une liste de mots comme descripteurs de documents à partir de laquelle une mesure de similarité sera calculée pour chaque requête. L'un des problèmes de cette approche réside dans le fait qu'elle néglige la sémantique des mots-clés descripteurs et garde un silence pénalisant sur leur éventuelle interdépendance. Par ailleurs, les utilisateurs trouvent souvent des difficultés pour exprimer leur besoin d'information et le formuler dans une requête composée de mots-clés. Ceci est dû en partie à l'impossibilité d'exprimer le besoin d'information par les seuls termes d'indexation utilisés par le système. Ainsi, de nouvelles approches sont proposées pour améliorer la satisfaction des utilisateurs envers les systèmes de recherche d'information (SRI).

2 Recherche intelligente

Ce qu'on peut dénommer par "recherche intelligente" constitue l'ensemble des approches et des techniques développées dans le cadre de la recherche d'information moderne pour faciliter davantage l'accès à l'information. Améliorer la recherche d'information et la rendre plus "intelligente" peut suivre trois axes d'études :

1. Concevoir des interfaces plus conviviales et plus intuitives,
2. Adapter la recherche à certains contextes spécifiques (profil, domaines, …etc.),
3. *Introduire l'aspect sémantique dans les SRI.*

Dans le premier axe, l'exploitation de l'interaction homme-machine dans le processus de recherche est devenue l'articulation centrale dans tout SRI. La visualisation des résultats n'est qu'une phase préliminaire dans une session interactive de recherche afin d'aider l'utilisateur à reformuler et satisfaire son besoin. La visualisation graphique, des pages dans un réseau de pertinence enrichi par les associations les plus fortes, constitue l'une des techniques utiles dans cette catégorie.

Dans le deuxième axe, il est devenu de plus en plus avantageux d'adapter le traitement d'une requête à l'environnement et au contexte de l'utilisateur d'un SRI. En essayant de sortir de la coquille des systèmes classiques orientés-requête, cette approche vise à développer des modèles de recherche adaptables au contexte de la session de recherche elle-même. Plusieurs aspects de ce contexte peuvent être recensés (personnel, social, professionnel, spatial, temporel, …etc.). Par ailleurs, une quantité impressionnante de connaissances pourrait être accumulée d'autres sessions de recherches et efficacement exploitée pour satisfaire au mieux le besoin en information. Des sujets, tels que la modélisation du contexte et le développement

de méthodes efficaces de filtrage collaboratif, restent un domaine de recherche en pleine expansion.

Dans la troisième direction, introduire l'intelligence dans la recherche d'information revient à répondre à la question :

Comment prendre en charge la sémantique imbriquée dans un texte ?

Cette tâche, qui représente une faculté humaine par excellence, représente le noyau de tout système prétendant analyser le langage naturel de façon automatique en vue d'indexer le contenu textuel des Téraoctets stockés dans les pages Web et les bibliothèques électroniques.

Nous pouvons résumer les techniques proposées pour permettre une recherche sémantique sur le Web dans trois approches principales :

1. Organiser la recherche (indexation de documents et/ou analyse de requêtes) autour de connaissances conceptuelles (thésaurus ou ontologie),
2. Utiliser un système d'annotation (documenté par des experts ou une masse d'utilisateurs) pour promouvoir la recherche collaborative,
3. *Développer des méthodes d'indexation sémantique des textes non-structurés.*

3 Ontologie ou annotation pour la recherche sémantique

Dans la première approche, un thésaurus peut être incorporé dans le SRI pour aider les utilisateurs à formuler les principaux concepts et relations sémantiques dans un domaine particulier. La connaissance conceptuelle peut être exploitée autrement comme propriété intrinsèque dans le SRI lors du processus de calcul de correspondance. A la fin des années 1990, Tim Berners-Lee avait introduit l'idée du Web sémantique en vue de structurer le contenu en-ligne autour d'une ontologie préconstruite [Berners-Lee, 1998]. L'approche visait à créer des entités semi-structurées facilitant la communication intelligente entre machines pour récupérer des informations ou offrir des services. La représentation de la connaissance dans le Web sémantique est réalisée grâce à une famille de langages de balisage spécifique (XML, RDF, RDFS, OWL). Une architecture assez consistante a été développée puis standardisée par le consortium W3C afin de définir les règles de conception, de déploiement et de validation des ressources en-ligne sous forme sémantique. Néanmoins, la construction d'ontologies nécessite un effort humain considérable et, par conséquent, un coût trop élevé.

Les chercheurs se sont intéressés alors à l'automatisation du processus d'apprentissage des ontologies, soit en partant du seul contenu non structuré des documents, soit en exploitant des ressources externes linguistiques ou organisationnelles comme méta-connaissance pré-structurée (taxonomies, ontologies, WordNet, articles Wikipedia, …etc.) [Paliouras, 2005] [Ciravegna et Chapman, 2005] [Dellschaft et Staab, 2006]. D'autres études se sont penchées sur la faisabilité d'exploiter ces ressources dans différentes tâches en RI [Hu et al., 2008] [Wang et Domeniconi, 2008]. La construction automatique, ou du moins semi automatique, des ontologies à partir de textes non-structurés constitue un champ d'étude intéressant mais qui reste loin de satisfaire les idéalistes du Web sémantique.

En termes de recherche d'information, cette approche offre une alternative utile pour passer de la recherche lexicale à base de mots-clés vers une recherche sémantique à base de concepts. L'idée d'intégrer la construction d'ontologie a été proposée dans [Hwang et al. 2007] et [Wei et al. 2008], mais ces travaux restent encore théoriques et se contentent de proposer des cadres de références génériques. Il s'avère difficile de concevoir et de gérer des ontologies relatives à tous les domaines pour être exploitées par les moteurs de recherche sémantiques. L'ambition de "structurer" le contenu du Web autour d'ontologie s'avère

irréalisable avec la croissance impressionnante des ressources en-ligne non-structurées. L'explosion du Web-2.0, où le contenu des utilisateurs (systèmes collaboratifs, réseaux sociaux), réduit considérablement la possibilité de couvrir la toile mondiale par un système de recherche à base d'ontologie. L'indexation par vocabulaire contrôlé s'avère coûteuse et non-évolutive surtout lorsqu'il s'agit des informations numériques aussi diversifiées que les pages Web [Shirky, 2005].

La deuxième approche repose sur l'ajout d'annotations aux documents en vue de décrire leur contenu sémantique [Popov 2003]. Plusieurs travaux ont proposé et étudié des systèmes d'annotation offrant des architectures spécifiques, pour le stockage, la recherche des documents, tels que SHOE [Heflin et Hendler, 2000], CREAM [Handschuh et Staab, 2002] et KIM [Popov 2003]. L'idée d'étiquetage des documents semble trouver ses arguments dans l'aspect non-structuré du Web mais surtout dans l'impossibilité de conceptualiser tout le Web de façon générale, stable et non ambiguë. La solution de catégorisation par annotation semble plus pratique surtout en tirant profit de l'approche collaborative dans le partage des signets. Les partisans de cette approche suggèrent souvent que les techniques d'annotation fournissent, au pire, un complément aux systèmes de classification traditionnels et, au mieux, un remplacement complet de tels systèmes [Shirky, 2005].

L'étude de la recherche d'information dans le contexte des folksonomies (étiquetage collaboratif ou indexation sociale) est généralement vue comme le chevauchement entre le vocabulaire contrôlé et celui d'étiquetage [Kipp et Campell, 2010]. Néanmoins, la réussite d'une telle approche dans sa dimension sociale nécessite un cumul considérable, aussi équilibré que diversifié, d'annotations par des masses d'utilisateurs. Dans une vision plus large, cette approche favorise le monopole des sites sociaux les plus populaires puisqu'elles détiennent la grande masse des jugements de pertinence ou d'annotations personnelles.

4 Indexation sémantique des textes non structurés

La troisième approche, pour prendre en charge l'aspect sémantique en RI, repose sur le développement et l'application de méthodes algébriques ou statistiques sur les textes non-structurés. Le modèle vectoriel d'indexation en sémantique latente (LSI) proposa de réduire, par une technique de l'algèbre linéaire, la matrice de cooccurrences termes×documents, pour capturer les descripteurs les plus pertinents dans une collection [Deerwester et al., 1990] [Dumais, 1993].

Par ailleurs, les modèles de thèmes partent d'une approche probabiliste plus solide et proposent une solution pratique et complètement automatique afin de prendre en charge certains aspects sémantiques des textes d'une collection. Sans faire recours à des ressources linguistiques externes, ces modèles sont basés sur l'idée que les documents constituent des mélanges de thèmes où chaque thème est une distribution de probabilités sur les mots [Hofmann et al., 1999] [Blei et al., 2003]. En particulier, le modèle d'allocation latente de Dirichlet (LDA) propose une solution plus générale, par rapport aux autres, puisqu'il produit non seulement une indexation sémantique des documents utilisés dans l'apprentissage mais, en plus, un modèle génératif calculant à la fois la distribution de chaque thème sur les mots et la distribution de nouveaux documents sur ces thèmes.

4.1 Problématique

Depuis son introduction originale, le modèle LDA a connu des développements diversifiés relatifs à l'amélioration de l'algorithme d'apprentissage, au traitement des collections volumineuses, au développement de variantes de modélisation particulières ou à la visualisation et l'interprétation des résultats. Néanmoins, il reste beaucoup à faire dans ce

domaine afin d'analyser la faisabilité d'utiliser le modèle LDA dans l'indexation sémantique de collections réelles de textes et de l'exploiter dans les tâches de RI [Yi et Allan, 2009].

Sur le plan du paramétrage et l'évaluation, le critère de choix du nombre approprié de thèmes pour apprendre le modèle LDA, à partir d'une collection donnée, reste subjectif et nécessite davantage d'études. Quant aux aspects multi-langages, peu de travaux sur le modèle LDA s'y sont intéressés, notamment aux applications dans le texte non-anglais. Ce n'est que récemment, que certaines études ont commencé à établir un cadre compréhensif pour la modélisation par thèmes multi-langages ou pour des langues autres que l'anglais telles que le chinois, l'espagnol et l'allemand [Zhang et al., 2010] [Jagaralamudi et Daumé, 2010].

Néanmoins, nous n'avons pu identifier aucune étude du modèle LDA pour le texte arabe. En fait, les développements relatifs à la RI dans le texte arabe sont insuffisantes vis-à-vis de l'importance grandissante que la langue arabe gagne à l'échelle internationale. A titre d'indication, l'arabe fait partie des six langues officielles de l'organisation des nations unies et enregistre la plus haute croissance en utilisateurs d'Internet durant la dernière décennie à 2,501.2%[2].

Dans un cadre plus général, l'analyse des thèmes dans le texte arabe avait fait l'objet de quelques travaux dans un contexte multilingue [Brants et al., 2002] [Oard and Gey, 2002] [Larkey et al., 2004]. Le constat dominant était qu'il est préférable de concevoir des modèles de thèmes arabes dans un contexte monolingue. Cependant, la majorité des études ont été réalisées sans connaissance préalable de la langue ni l'aide d'experts en langue arabe. Il est difficile d'assimiler qu'il soit possible d'apprécier les aspects sémantiques d'un contenu rédigé dans une langue aussi riche que l'arabe par des étrangers à la langue du texte analysé.

Ainsi, nous arrivons à définir la problématique de notre étude sous la forme suivante :

1. *Quelle est la faisabilité d'utiliser un modèle de thème comme approche d'indexation sémantique des textes pour les tâches de RI ?*
2. *Comment évaluer et interpréter le modèle de thème pour l'analyse sémantique du contenu d'une collection ?*
3. *Dans quelle mesure peut-on appliquer les modèles de thème dans le texte non-structuré non-anglais (l'arabe comme exemple d'étude) ?*

4.2 Contribution

Suite à une large exploration des modèles de recherche d'information pour les textes non-structurés et les approches d'indexation sémantique du contenu textuel, nous avons conçu des outils d'évaluation des modèles de thèmes. En particulier, le modèle LDA a été analysé et évalué sur une panoplie de collections. En plus des expérimentations menées sur les documents anglais, nous avons jugé nécessaire de développer des outils linguistiques pour l'analyse morphologique et l'indexation sémantique des textes arabes. Des collections d'articles de presse (anglais et arabes) ont été automatiquement extraites du Web et utilisés comme corpus de test en plus des standards librement disponibles. Nous pouvons, ainsi, résumer l'essentiel de notre contribution dans les travaux suivants :

[2] Internet World Stats, Usage and population statistics (Miniwatts Marketing Group), accessible sur : http://www.internetworldstats.com/

1. *L'analyse et l'évaluation du modèle LDA dans les tâches de recherche et de catégorisation des textes sur des corpus réels.*
2. *La proposition d'une nouvelle mesure, à base de la divergence de Kullback-Leibler, pour le paramétrage de l'apprentissage des thèmes dans une collection donnée.*
3. *Le développement d'un nouvel algorithme de stemming à base de lemme pour l'analyse et l'indexation du texte arabe.*
4. *L'élaboration de trois collections linguistiques arabes, à base d'articles de presse relatifs à la période 2007-2010, et leur mise à la disposition des chercheurs dans le domaine de la RI et l'analyse du texte arabe*.

5 Plan de la thèse

En plus de ce chapitre d'introduction générale, nous présentons dans le **chapitre 2** l'essentiel des notions de base en RI. Nous dressons le contexte historique de l'évolution des pratiques d'organisation et d'accès à l'information. Nous décrivons les principaux systèmes de classification de la connaissance que l'homme a développés pour faciliter l'accès aux documents et satisfaire son besoin en information. Les deux tâches de RI qui nous importent le plus dans notre étude, la recherche et la catégorisation, sont ensuite décrites avant d'en expliquer les principales méthodes d'évaluation.

Le **chapitre 3** est consacré aux modèles classiques en RI pour l'indexation et la recherche des documents textuels. Les fondements théoriques et les aspects pratiques, des approches de modélisation de textes et du calcul de pertinence, sont développés. Nous terminons le chapitre par une analyse comparative des principaux modèles, booléens, vectoriels et probabilistes, en RI.

Dans le **chapitre 4**, nous décrivons l'approche de modélisation par thème des textes non structurés. En particulier, nous développons différents aspects relatifs au modèle LDA avec des exemples illustratifs. Différentes techniques d'évaluation son présentées avant d'introduire notre proposition de mesure empirique à base de la divergence de Kullback-Leibler. Par ailleurs, nous analysons certaines approches basées sur le modèle LDA dans la recherche ad-hoc.

Nous revenons dans le **chapitre 5** au prétraitement linguistique dans le processus d'indexation du texte non-structuré. Les principales notions de morphologie dans le langage naturel sont décrites avant d'analyser de près les techniques de stemming. Nous abordons les approches d'évaluation des différentes méthodes de stemming que ce soit avec ou sans prise en compte des performances des tâches de RI. La deuxième partie du chapitre est consacrée à l'analyse du texte arabe en vue de son indexation dans un SRI. Nous expliquons les principales caractéristiques dérivationnelles et morphologiques qui accentuent la complexité d'analyse automatique dans le texte arabe. Nous proposons un nouvel algorithme *BBw* de stemming à base de lemme pour une indexation sémantique efficace des documents arabes.

Le **chapitre 6** décrit les aspects pratiques de notre contribution relative à l'analyse sémantique dans les textes non-structurés. Nous décrivons d'abord les corpus automatiquement collectés du Web. En particulier, les trois collections d'articles de presse arabe extraits des sites (*Echorouk*, *Reuters* et *Xinhua*) sont présentées et analysées. Plusieurs algorithmes de stemming du texte arabe sont comparés aux variantes de notre analyseur *BBw* sur différentes collections arabes. Nous présentons nos expérimentations d'évaluation du modèle de thème LDA dans les textes non-structurés arabes.

En conclusion, nous dressons un récapitulatif de nos travaux et nous discutons les différents résultats obtenus dans notre étude de la faisabilité d'utiliser les modèles de thèmes pour l'indexation sémantique dans les tâches de RI. Nous traçons les perspectives qu'on doit explorer et raffiner afin de mieux comprendre et appliquer les méthodes d'indexation sémantiques dans les textes non-structurés.

Chapitre 2 :
RECHERCHE D'INFORMATION : ETAT DE L'ART

1 Introduction

Les moteurs de recherche sur le Web sont largement utilisés par des millions d'utilisateurs dont l'activité en dépend quotidiennement. Les gens s'en servent de plus en plus pour faciliter leurs travaux dans l'éducation, le commerce, l'administration …etc. Des moteurs de recherche tels que Google, Yahoo ou Bing, qui sont devenus des stars du Web, épuisent pleinement des modèles fondamentaux et des nouvelles techniques de la recherche d'information et des services Web afin d'offrir les dernières nouveautés technologiques, de localiser les groupes et les organisations, de résumer les dépêches de presse ou de simplifier le shopping. Les chercheurs académiques et les étudiants du primaire à l'université considèrent la recherche dans les bibliothèques électroniques un moyen incontournable pour élaborer leur étude avec les dernières mises à jour. Bien qu'un moteur de recherche ne couvre qu'une partie du Web public, Google enregistre plus qu'un milliard de recherches par jour [Go-Globe, 2011].

Par ailleurs, les applications de recherche ont fait preuve d'utilité et d'efficacité dans deux autres environnements aussi importants que le Web : d'un côté, les systèmes de recherche de bureau permettent aux individus d'accéder rapidement à leurs courriers, documents et fichiers personnels. D'un autre côté, les solutions de recherche et de mining pour l'entreprise constituent le cœur de tout système d'information professionnel des grandes compagnies internationales aussi bien pour leurs besoins internes en production, gestion et management que pour leurs activités externes de marketing et de veille stratégique. La recherche est devenue la deuxième fonction la plus utilisée dans l'entreprise après le mail [Bonny et Garnier, 2008].

Ce chapitre présente l'essentiel des concepts et des domaines reliés à la recherche d'information. Il trace un aperçu historique sur l'évolution des pratiques et des fondements de l'indexation documentaire et de l'analyse automatique des contenus textuels.

2 Définitions et Objectifs

La recherche d'information (*RI*) (*en angl. Information retrieval*) est un domaine qui s'intéresse à la représentation, le stockage, l'organisation et l'accès aux éléments d'information. Elle étudie la manière de répondre pertinemment à une requête pour retrouver de l'information utile dans une collection de données. Cette dernière est constituée de documents, d'une ou de plusieurs bases de données, décrits par un contenu ou des métadonnées associées. Le contenu des documents peut être un texte, une page Web, une image, un son, une vidéo ou même un élément spatial d'une carte géographique. On parle ainsi de plusieurs sous-disciplines plus fines telles que la recherche du texte, la recherche sur le Web, la recherche d'image, la recherche multimédia ou la recherche d'information géographique. Notre étude s'intéresse seulement à la recherche dans le contenu textuel.

2.1 Accès à l'information

L'objectif principal de la *RI* est de faciliter à l'utilisateur l'accès à l'information qui l'intéresse. Ceci nécessite, d'une part, une meilleure représentation des éléments d'information et d'une autre part, une caractérisation efficace du besoin de l'utilisateur.

Bien que la recherche d'information puisse être subdivisée sous plusieurs façons, il semble qu'il y ait trois principaux domaines de recherche qui composent entre eux une partie considérable du domaine de la RI. Ils sont [Rijsbergen, 1979]:

- l'analyse du contenu,
- les structures d'information et,
- l'évaluation.

En bref, le premier concerne la description du contenu des documents sous une forme appropriée pour le traitement informatique ; le second s'intéresse à l'exploitation des relations entre les documents afin d'améliorer les stratégies de recherche ; cependant le troisième concerne les mesures d'évaluation des méthodes de recherche.

2.2 Besoin d'information

Un système de recherche d'information (SRI) manipule une ou plusieurs collections de données. Le document, ou parfois une partie seulement, constitue l'information élémentaire qui devrait être incluse, ou non, dans l'ensemble des réponses au besoin de l'utilisateur. Ce besoin en information se réfère à la quantité manquante d'information nécessaire pour qu'un utilisateur puisse atteindre ses objectifs dans une situation particulière. Cette définition sous-entend trois hypothèses :

- L'utilisateur peut ne pas savoir exactement son besoin d'information.
- L'utilisateur peut ne pas être capable de formuler son besoin d'information.
- Le besoin en information d'un utilisateur peut changer durant une session de recherche.

Ce besoin en information prend trois formes possibles Selon [Ingwersen, 1992]:

- Un besoin vérificatif où l'utilisateur cherche à examiner ou localiser certains documents par rapport à des informations préalablement connues par l'utilisateur.
- Un besoin thématique connu, lorsque l'utilisateur cherche à clarifier, revoir ou enrichir ces connaissances sur un sujet connu.
- Un besoin thématique inconnu, lorsque l'utilisateur cherche à explorer de nouveaux concepts en dehors de ces connaissances.

Par conséquent, la formulation de la requête de recherche peut être plus ou moins complète. Les termes utilisés expriment d'une certaine manière le besoin en information du point de vue de l'utilisateur. Cette formulation peut changer pour des situations et des contextes différents pour le même utilisateur et peut diverger davantage d'un utilisateur à un autre pour le même besoin. De ce fait, l'appréciation de la qualité d'un SRI sur la base de la satisfaction de l'utilisateur constitue un défi aussi cognitif que technique (quel sens porte sa requête et comment déterminer la réponse idéale).

Alors que l'objectif d'un SRI est de retrouver des documents ayant une certaine signification (par rapport au besoin), son implémentation est réalisée de façon pour chercher des documents contenant certains termes (par rapport à la requête). Un besoin d'information est défini par le sujet sur lequel un utilisateur désire savoir plus alors qu'une requête est ce que l'utilisateur communique à l'ordinateur pour exprimer ce besoin [Manning et al., 2008].

2.3 Bases de données et RI

La collection, souvent appelée corpus, peut être structurée comme dans une base de données relationnelle ; elle peut être semi-structurée en associant aux documents une métadonnée organisationnelle, comme elle peut être non-structurée en se limitant au seul contenu des documents. Il en découle la distinction entre la recherche des données (*en angl. Data retrieval*) et la RI. Nous récapitulons dans Tableau 2-1 les principales caractéristiques recensées en littérature pour tracer des frontières claires entre les deux domaines [Rijsbergen, 1979] [Baeza-Yates et Ribeiro-Neto, 1999].

	Recherche de données	**Recherche d'information**
Donnée	Structurée	Non structurée
Champ	Sémantique claire	Pas de champs (texte brut)
Requête	Complète, langage artificiel	Incomplète, langage naturel
Correspondance	Exacte	Partielle, pertinente
Modèle	Déterministe	Probabiliste
Inférence	Déductive	Inductive

Tableau 2-1. Comparaison entre la recherche de données et la RI.

Cette distinction revient en premier lieu au niveau de structuration des objets recherchés. Alors que dans une base de données "structurée", nous recherchons des informations ciblées avec une correspondance exacte ; la recherche d'information traite généralement des documents textuels "non structurés". La pertinence des résultats est plus visée en RI que la complétude et l'exactitude.

Mais avant d'en arriver à l'aire des ordinateurs et du Web, l'humanité a dû réaliser plusieurs progrès aussi théoriques que pratiques dans les domaines de représentation, de stockage et d'accès à l'information.

3 Contexte historique

La pratique d'archivage des informations écrites remonte à environ 3000 avant J.-C., lorsque les Sumériens eurent réservé des endroits pour stocker les tablettes d'argile portant des inscriptions cunéiformes. Réalisant que la bonne organisation et l'accès aux archives étaient essentiels pour l'utilisation efficace de l'information, les Sumériens eurent développé un système de classification spécifique [Singhal, 2001].

3.1 Quatre inventions historiques

Avant d'arriver à la recherche documentaire sous ses formes modernes du 21ème siècle, il est nécessaire de citer les quatre inventions fondamentales qu'avait marquées l'histoire de l'organisation de l'information :

- le papier,
- l'imprimerie,
- l'ordinateur et,
- le Web.

En effet, Les besoins en archivage et en recherche de l'information écrite devinèrent de plus en plus importants à travers les siècles, en particulier avec le développement de la fabrication du papier par les chinois en l'an 8 avant J.C. L'invention de l'imprimerie par (Gutenberg, 1440), avait réduit considérablement le coût du livre et avait permis ainsi d'en élargir la diffusion.

L'apparition de l'ordinateur, à la fin de la deuxième guerre mondiale, représentait le troisième progrès significatif. La première machine entièrement électronique (ENIAC) a été achevée en 1946. Dès lors, des avancées considérables, tant conceptuelle que technologiques, avaient bouleversés le mode de vie de l'humanité tout au long de la deuxième moitié du 20 siècle. Nous verrons plus tard comment l'organisation, le stockage et l'accès à la documentation électronique avaient pris un essor révolutionnaire.

Au début des années 90 du siècle passé, Tim Berners-Lee, inventa le World-Wide-Web. Le Web combinait trois technologies principales : l'adresse universelle d'une ressource Web (URI), le protocole de communication (HTTP) et le langage d'écriture des pages Web (HTML). Bien qu'il ne représente qu'une application du grand réseau mondial Internet (officiellement créé en 1983), c'est le Web qui a déclenché le Boom d'Internet au grand public. Le Web public, souvent confondu à l'Internet, représente actuellement la plus grande mine d'information libre et est librement accessible à travers le monde. Les méthodes d'indexation, d'accès et de représentation de l'information sur le Web ne cessent d'augmenter et de murir pour tirer plein profit des ressources gigantesques de la toile mondiale.

3.2 Evolution de la classification documentaire

Sur le plan de l'organisation documentaire, l'histoire retient le nom de l'Irakien (Ibn An-Nadim, 987) qui proposa dans son ouvrage «Al-Fihrist[3]» une classification bibliographique des livres existants et connus à son époque. La Figure 2-1 présente un extrait de la hiérarchie de l'index qui comprend 1303 notices (résumé d'un livre avec biographie de l'auteur) regroupées dans 10 classes et 33 sous-classes [Rebhi et Odoura, 1990].

[3] Mot d'origine persane qui veut dire liste méthodique des livres ou catalogue.

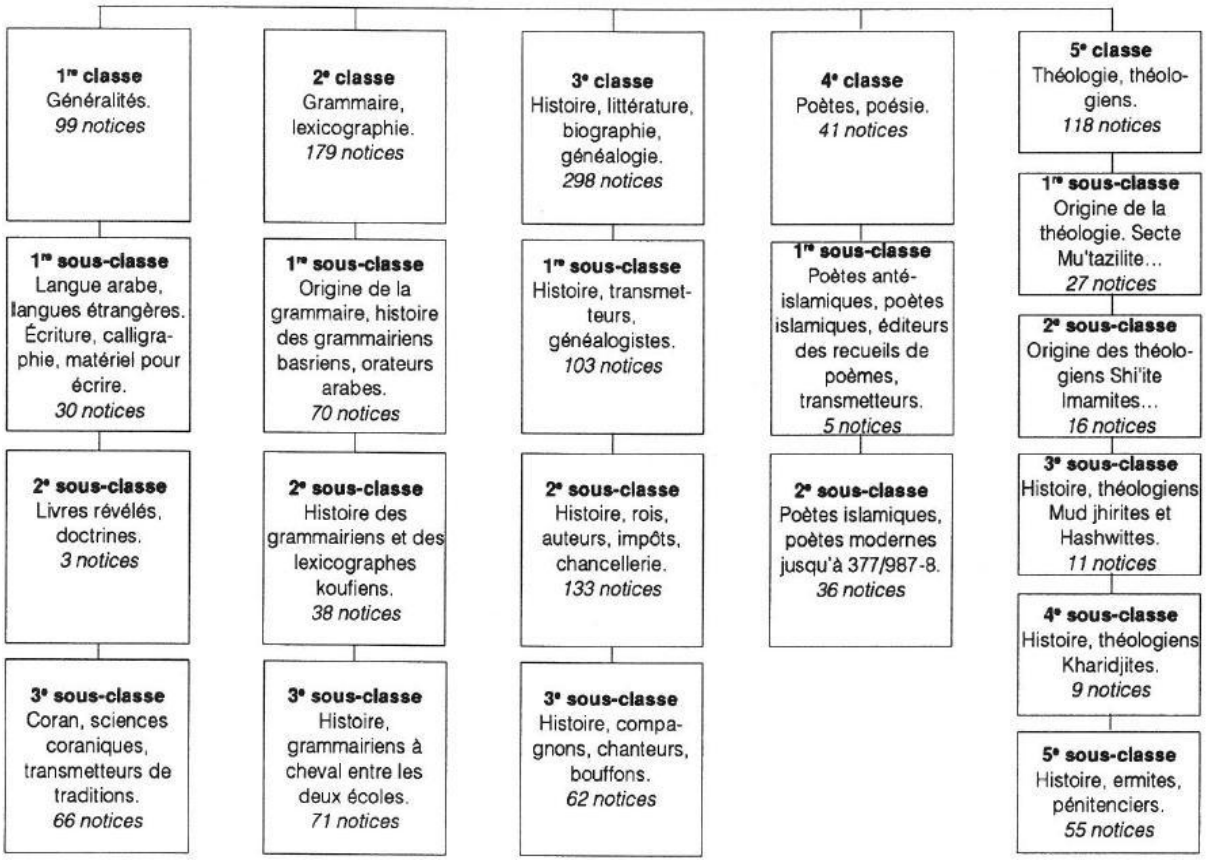


Figure 2-1. Extrait du classement de l'index «Al-Fihrist» de Ibn-An-Nadim.

Plus tard au 17ème siècle, le turc Haji Khalifa (connu aussi par Katip Celebi) avait produit son dictionnaire bibliographique répertoriant près de 15.000 entrées dans l'ordre alphabétique et selon un catalogue bibliographique qui était le précurseur direct des méthodes d'indexation modernes.

Ce n'est que deux siècles après que Melvil Dewey développa en 1876 un système pour la classification hiérarchique du savoir humain dans une bibliothèque. Basé sur 10 classes générales notées de 000 à 900, le système «Classification Décimale de Dewey» compte plus de 20.000 indices répartis sur 100 divisions et 1000 sections. Le Tableau 2-2 présente un exemple d'un code Dewey.

Code 521.1	
500 Science	Classe
520 Astronomie	Division
521 Mécanique céleste	Section
521.1 Gravitation	Sous-section

Tableau 2-2. Description hiérarchique d'un code Dewey "521.1"

Bien qu'ayant été considérablement amélioré, le CDD reflète toujours l'organisation générale du savoir aux USA lors de la fin du 19 siècle. En effet, deux juristes Belges Paul Otlet et Henri La Fontaine, proposèrent, en 1905, la CDU «Classification Décimale Universelle» (voir le Tableau 2-3).

Classement de Dewey (CDD)	Classement universel (CDU)
000 Généralités, Informatique, information et ouvrages généraux	**0** Généralités, informatique, information, documentation
100 Philosophie et psychologie	**1** Philosophie, psychologie
200 Religion	**2** Religion, théologie
300 Sciences sociales	**3** Sciences sociales.
400 Langues	**4** inoccupée
500 Science	**5** Sciences pures, sciences exactes
600 Technologie	**6** Sciences appliquées, médecine, technologie
700 Arts et divertissement	**7** Arts, divertissements, sports
800 Littérature	**8** Langue, Linguistique, philologie, littérature
900 Histoire et géographie	**9** Géographie, biographie, histoire

Tableau 2-3. Description des deux classifications décimales (CDD et CDU)

Mais ceci n'a pas empêché de voir naître d'autres classifications spécifiques et plus appropriées à certains contextes culturels et sociopolitiques. La classification de la bibliothèque du Congrès (LCC) en est un exemple type. Initié par Herbert Putnam en 1897, le système LCC a été développé et raffiné au long d'un demi siècle en vue d'organiser la bibliothèque du Congrès Américain. Le système reste d'actualité puisqu'il est toujours utilisé par la plupart des bibliothèques académiques et universités aux USA et même ailleurs. Le codage du classement et alphabétique (A-Z) avec des subdivisions hiérarchiques (BC, BD, …etc.)

Classement du Congrès Américain (LCC)	Classement Soviétique (BBK)
A : Généralités	**1.** Marxisme-léninisme (A-A)
B : Philosophie. Psychologie. Religion	**2.** Sciences naturelles en général (B-Б)
C : Sciences auxiliaires de l'histoire	**3.** Sciences physiques et mathématiques (V-B)
D : Histoire du monde et histoire de l'Europe, l'Asie, l'Afrique, l'Australie, la Nouvelle- Zélande, etc.	**4.** Sciences chimiques (G-Г)
E : Histoire des Amériques (généralités et Etats-Unis)	**5.** Sciences de la terre (D-Д)
F : Histoire des Amériques (autres pays d'Amérique)	**6.** Sciences biologiques (E-E)
G : Géographie. Anthropologie. Loisir	**7.** Techniques et sciences techniques en général (Ž-Ж)
H : Sciences sociales	**8.** Energétique et radio-électronique (Z-З)
J : Sciences politiques	**9.** Industrie minière (I-И)
K : Droit	**10.** Technologies des métaux, construction mécanique, construction d'appareils (K-К)

Tableau 2-4. Extrait (les 10 premières classes) des deux classifications (LLC et BKK)

Non convaincus par la modélisation "libérale" du savoir humain, les russes publièrent en 1968 la BBK «Classification Bibliothéco-Bibliographique» ; un système basé sur la CDU et reflétant l'organisation de la connaissance de l'ère de l'Union Soviétique. Le codage du système BBK était alphanumérique à base de 28 classes générales en intégrant le marxisme-léninisme et en favorisant les sciences objectives.

La classification BBK, comme le montre le Tableau 2-4, reflète clairement l'organisation de la connaissance du point de vue de l'ère Soviétique. On constate immédiatement la place privilégiée attribuée au "marxisme-léninisme" ; viendra après les sciences pures puis les techniques …etc. On trouve une dichotomie distinctive entre l'URSS et le reste du monde ou bien encore entre marxiste d'un côté et pré-marxiste ou non marxiste d'un autre côté.

Ceci nous conduit à conclure que l'organisation de la connaissance humaine est souvent soumise à différents facteurs (temporel, géographique et idéologique) relatifs à l'environnement socioculturel des peuples. Malgré les variations descriptives, minimes qu'elles soient ou majeures, les systèmes de classification cités (CDD, CDU, BBK et LCC) sont toujours d'actualité et représentent jusqu'à nos jours la conceptualisation de la connaissance humaines pour différentes bibliothèques et universités à travers le monde.

3.3 Systèmes automatiques pour la recherche d'information

Dès l'invention des premiers ordinateurs, les scientifiques réalisèrent qu'ils peuvent être utilisés pour le stockage et la recherche automatiques de l'information. Vannevar Bush publia en 1945 un article révolutionnaire sur un système futuriste automatisant la recherche et l'accès aux connaissances humaines. Intitulé «As We May Think», l'article est reconnu à nos jours comme le texte fondateur de la bibliothèque électronique et du cyberspace. L'idée du Memex, proposée dans cet article, fut développée pour décrire comment des archives de textes peuvent être classés et recherchés automatiquement (voir Figure 2-2).

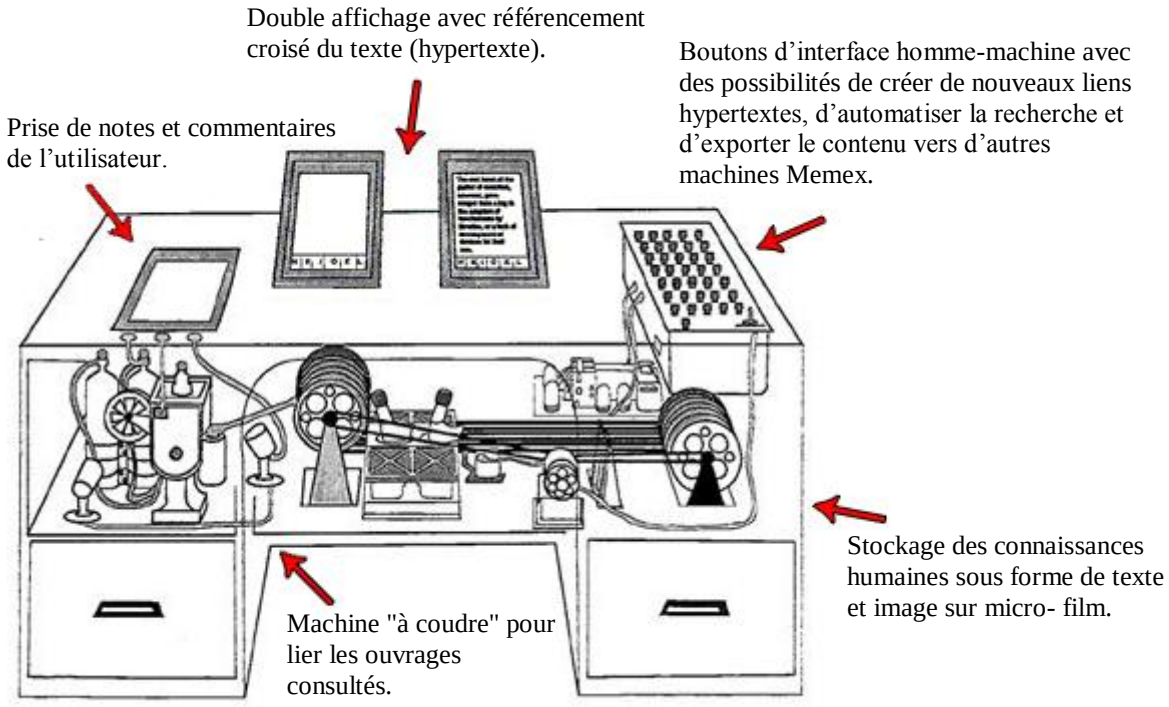


Figure 2-2. Schéma descriptif du Memex imaginé par V. Bush.

Peu après, plusieurs travaux furent apparus en se basant sur le principe de recherche textuelle avec un ordinateur. L'utilisation des mots, comme unités d'indexation des documents, fût proposée pour la première fois par [Luhn, 1957]. Pour caractériser et rechercher les textes stockés dans un ordinateur, il détermina les mots significatifs en

calculant leur fréquence d'apparition. Ainsi, une liste de ce qu'on pourrait appeler «mots clés» était générée pour chaque document.

L'approche expérimentale en RI émergea avec les travaux de [Cleverdon, 1967] qui développa une méthode d'évaluation des systèmes d'indexation textuelle avec des tests sur la collection Cranfield relative aux publications des revues scientifiques d'aéronautique. Mais le plus marquant de cette période, était le système SMART (système pour l'analyse mécanique et la recherche de texte) développé par G. Salton et ses étudiant à l'université de Harvard puis à Cornell [Salton, 1971]. Le regroupement de plusieurs modèles d'indexation et de méthodes d'évaluation dans SMART permettait aux chercheurs de perfectionner plusieurs algorithmes de recherche sur une variété de corpus.

Les années 1970 et 1980 ont connu plusieurs développements couvrant différents aspects du processus de recherche d'information. Nous décrivons dans les sections suivantes les principaux modèles et méthodes dans ce domaine. Néanmoins les expérimentations menées à cette époque n'étaient validées que sur des collections modérées (quelques milliers d'articles). L'efficacité des nouveaux modèles sur des corpus réels à grande échelle restait sans réponse [Saigal, 2001].

3.4 Annuaires et moteurs de recherche Web

Ce n'ai qu'au début des années 1990 que la RI prenait un nouvel élan tout à fait révolutionnaire. D'un côté, la conférence de recherche de texte (TREC) lança une série de tests d'évaluation et de concours pour différentes tâches RI sur de large corpus [TREC, 1992]. Plusieurs algorithmes et modèles ont été révisés et perfectionnés pour s'adapter à la recherche dans de grandes collections. D'un autre côté, l'explosion spectaculaire du World Wide Web a fait rapidement intervenir les techniques de la RI dans la recherche sur le Web à partir de 1996. A la veille du troisième millénaire, la RI moderne était déjà concrétisée par des bibliothèques électroniques uniformément accessibles à travers le globe un demi siècle après l'apparition de l'idée du Memex[4].

Avec la croissance impressionnante de la toile mondiale, plusieurs systèmes en ligne sont mis en service pour faciliter l'accès au contenu grandissant et diversifié du Web. D'un côté, certains annuaires proposent des catalogues thématiques élaborés manuellement. Parmi les plus connus, nous citons les catégories Yahoo et le projet du répertoire ouvert (pus connu sous le nom de dmoz). D'un autre côté, sont apparus des moteurs de recherche (Google, Yahoo!, Altavista … etc.) qui représentent de nos jours l'entrée indispensable pour tout utilisateur Web.

Dans le contexte du Web, le besoin d'information n'est pas toujours d'un aspect informationnel. Une étude avait proposé trois catégories pour la classification de la recherche Web [Andrei, 2002] :

- Recherche navigationnelle : lorsque l'objectif immédiat est d'atteindre un site particulier.
- Recherche informationnelle : lorsque l'intention est d'acquérir certaines informations supposées être présentes sur une ou plusieurs pages.
- Recherche transactionnelle : lorsque l'objectif est d'effectuer certaines activités de médiation sur le Web.

[4] Vannevar Bush mourra en 1974 avant de jouir de Google ou de l'encyclopédie libre Wikipedia.

Lancaster (l'un des pionniers de la RI) disait qu' "un système de recherche d'information n'informe pas (c'est à dire changer la connaissance de) l'utilisateur sur l'objet de son enquête. Il informe sur la seule existence (ou la non-existence) et la localisation des documents relatifs à sa demande" [Lancaster, 1968]. Cette définition semble être révisée avec l'évolution des besoins en documentation électronique et le développement des objectifs des systèmes de recherche sur le Web. De nos jours, un moteur de recherche n'est pas interrogé pour la simple recherche d'un contenu Web mais, bien plus, il permet d'évaluer la popularité du sujet associé aux termes de recherche, d'identifier leur contexte et d'accéder aux contenus pertinents et leur sources officielles.

3.5 Tendances actuelles

La croissance incessante de la quantité du texte en ligne et de la demande d'accès aux différents types d'information, ont, toutefois, ressuscité l'intérêt pour une panoplie de domaines liés étroitement à la recherche d'information mais allant au-delà de la simple recherche documentaire [Allan et al., 2003].

En effet, la vision classique est rectifiée en RI moderne avec l'évolution des technologies du Web et des bases données distribuées. La propriété de structuration, qui traçait une frontière claire entre la recherche de données et la RI, est devenue moins évidente dans les textes balisés accessibles sur le Web. Les contenus sous des formats tels que *xml* ou *rdf*, sont considérés comme semi-structurés et leur manipulation relève, à priori, du domaine des bases de données.

Munie des techniques avancées d'intelligence artificielle, la recherche sur le Web a fait donc converger RI et Bases de données vers des intérêts communs tels que la recherche distribuée et la fouille de données. Ce constat a été évoqué dans un rapport collectif élaboré par plus qu'une trentaine d'auteurs des plus renommées dans la spécialité [Allan et al., 2003]. Nous présentons dans ce qui suit, les principaux domaines d'application de la RI moderne.

4 La recherche Web

La recherche sur le Web représente, de nos jours, l'application la plus importante pour la RI. Du point de vue de l'utilisateur, un système de recherche doit réaliser trois tâches :

- L'acquisition et l'analyse de la requête,
- Filtrage des documents pertinents,
- Visualisation des résultats (voir Figure 2-3).

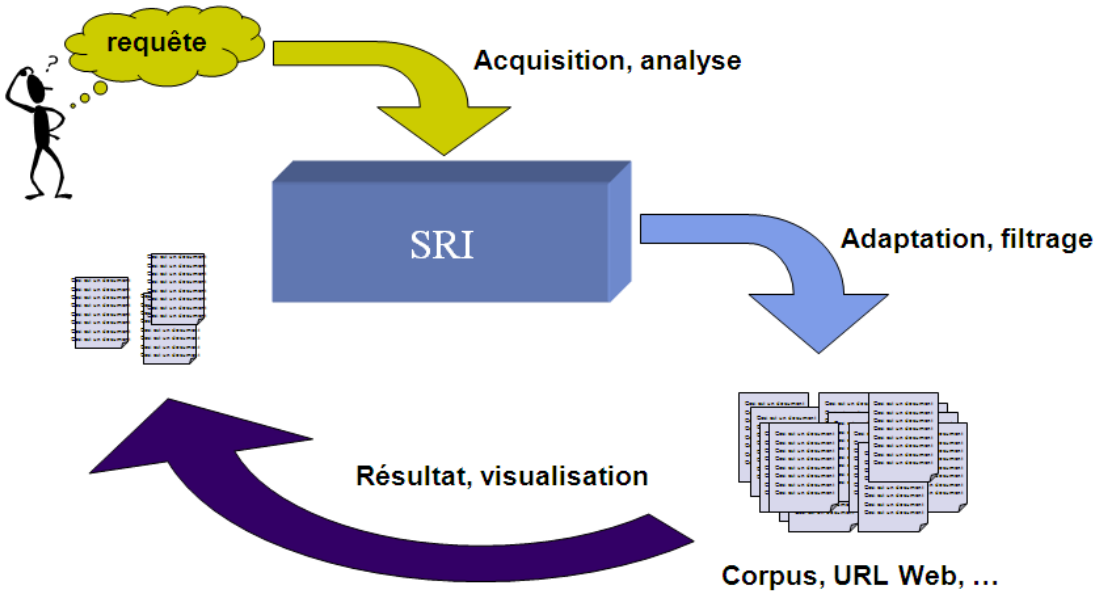


Figure 2-3. Système de recherche d'information (vue de l'utilisateur).

4.1 Architecture d'un système de recherche

Afin de satisfaire les besoins grandissants et spécifiques des utilisateurs dans un environnement aussi riche que le Web, un moteur de recherche doit considérer d'autres fonctionnalités aussi indispensables que complexes. Nous considérons dans le présent contexte, l'architecture plus généralisées d'un moteur de recherche Web mais qui comprend les fonctionnalités de base d'un système classique. La collecte, la représentation et le calcul des correspondances constituent le cœur de tout système de recherche en-ligne. On peut résumer ces fonctions comme suit :

- La collecte des documents et pages Web accessibles (crawling),
- L'analyse du contenu documentaire
- L'indexation et la représentation,
- L'analyse et la représentation de la requête utilisateur,
- L'appariement et l'ordonnancement (ranking),
- La visualisation.

L'architecture appropriée à ce genre de système est en "U" (Figure 2-4). Ceci implique que les techniques d'analyse linguistique, ainsi que les modèles de représentation, doivent être équivalents aussi bien pour les documents que pour les requêtes.

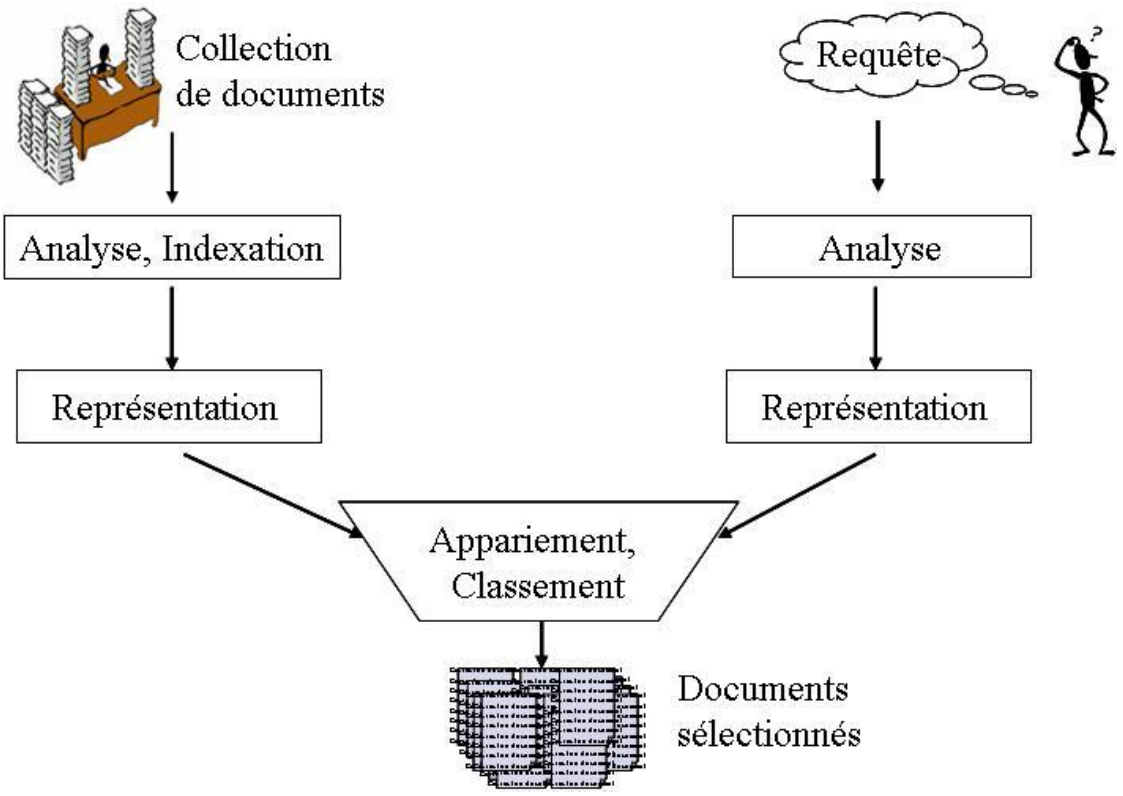


Figure 2-4. Architecture globale d'un système de recherche d'information.

4.2 Fonctionnement

Le processus de recherche dans un système de recherche en ligne peut être décrit par les fonctions suivantes :

4.2.1 La collecte des documents

Elle est réalisée par exploration continue du Web (Web-crawler). Un robot (agent parcourant de façon régulière les adresses d'Internet) est chargé de répertorier les nouveaux contenus mis el ligne. En suivant récursivement les liens hypertextes, un crawler construit graduellement l'index brut du moteur de recherche.

4.2.2 L'analyse du contenu

L'analyse du contenu documentaire déploie des techniques de prétraitement linguistique pour sélectionner les contenus significatifs lexicalement. La détection préalable du codage des caractères et de la langue utilisée décide de la manière à suivre pour cette fonction. Les méthodes d'analyse automatique du langage naturel sont très sollicitées et la performance des systèmes de recherche en dépende fortement. Nous décrirons l'essentiel des approches de prétraitement linguistique dans (Chapitre 5 :1).

4.2.3 La représentation

La représentation du texte respecte un modèle de document qui affectera directement la performance et l'efficacité de tout système de recherche. La fonction d'indexation a pour rôle d'extraire d'un document ou d'une requête, une représentation paramétrée (descripteurs) qui couvre au mieux son contenu significatif. Les descripteurs représentent généralement les termes reconnus par le système et sont rangés dans un dictionnaire constituant le langage d'indexation. Le chapitre suivant décrira les différents modèles d'indexation documentaire.

4.2.4 L'appariement

L'appariement est basé sur une fonction d'inférence qui permet d'associer à une requête les documents pertinents. Etroitement liée au modèle de représentation, l'objectif de cette tâche est de sélectionner parmi des millions de documents seulement quelques dizaines ou centaines les plus appropriés.

La valeur de pertinence du système est définie comme mesure de similarité entre une requête (*Q*) et un document (*d*) du corpus et dénotée généralement par *RSV*(*d*,*Q*) (*en angl. Retrieval Status Value*). La fonction de calcul de la *RSV* dépend du modèle de document utilisé notamment de la façon avec laquelle les termes sont pondérés. Elle permet dans la majorité des modèles, non seulement de sélectionner les documents pertinents à une requête, mais aussi de les ordonner par degrés de correspondance. L'astuce d'attribuer un score de pertinence, pour classer l'ensemble des documents par ordre décroissant de pertinence à une requête, a fait sa meilleure concrétisation avec le *PageRank* de Google.

4.2.5 La reformulation de la requête

Cette étape peut paraître facultative mais elle peut améliorer considérablement la qualité d'un système de recherche. Celui-ci, qui se base essentiellement sur la requête exprimée par l'utilisateur, doit répondre efficacement au besoin d'information. Néanmoins, ce besoin n'est pas toujours clairement et explicitement formulé. Par conséquent, les documents retournés par le système de recherche peuvent appartenir à des domaines tout à fait divergents du centre d'intérêt de l'utilisateur.

La reformulation des requêtes consiste généralement à enrichir la requête de l'utilisateur en ajoutant des termes permettant de mieux exprimer son besoin [Efthimiadis, 2000]. L'un des principes de reformulation consiste à modifier la requête pour ressembler davantage aux documents jugés pertinents et s'éloigner des documents non pertinents. La reformulation peut être soit interactive (assistées par l'utilisateur), soit automatique. Parmi les techniques les plus répandues nous citons la rétroaction de pertinence (*en angl. Relevance feedbak*) et l'expansion de la requête (*en angl. Query expansion*) [Rocchio, 1971] [Efthimiadis, 2000] [Manning et al., 2008]. Néanmoins l'amélioration des résultats de recherche dépend du corpus lui-même et du nombre et de la façon avec laquelle les termes sont rajoutés.

4.2.6 La visualisation

La forme la plus simple, pour afficher les résultats d'une recherche Web, consiste à inclure une liste d'hyperliens vers les adresses des documents jugés pertinents. Cette liste est souvent ordonnée selon la distance de chaque document par rapport à la requête. Toutefois, l'aspect d'interactivité avec un poste informatique connecté à Internet, offre la possibilité d'exploiter l'appréciation préliminaire de l'utilisateur en vue de relancer une nouvelle recherche plus adéquate. Ainsi, la visualisation des résultats de recherche n'est plus considérée comme dernière phase du processus de recherche mais plutôt comme articulation centrale dans un processus interactif mettant en évidence la collaboration du demandeur.

L'objectif étant d'exploiter les caractéristiques du système visuel humain pour faciliter la manipulation et l'interprétation des données, les tâches de la visualisation d'information peuvent être décrites comme suit :

- L'exploration rapide d'ensembles d'informations inconnues,
- La mise en évidence de relations et de structures dans les informations,
- La mise en évidence de chemins d'accès à des informations pertinentes,
- La classification interactive des informations.

La plupart des informations que l'on souhaite visualiser dans ce contexte peuvent être classées en quatre grandes catégories selon leur nature [Salton et McGill, 1983] :

- Des ensembles qui peuvent être vus comme des listes,
- Des ensembles qui peuvent être structurés de manière arborescente,
- Des ensembles dont on peut extraire des structures de graphe,
- Des ensembles pour lesquels on peut exploiter un indice de similarité.

En plus de la référence en hyperlien, un moteur de recherche présente pour chaque document trouvé, son titre et son résumé (ou un extrait contenant les termes de recherche). C'est le cas de la majorité des moteurs de recherche qui affichent les résultats dans une liste ordonnée (Google, Yahoo, Altavista, …etc.)

Cependant, certains outils de recherche Web proposent une visualisation graphique et interactive des résultats de recherche en exploitant les hyperliens contenus les pages Web et les similarités calculées entre elles. Plusieurs techniques, peuvent être déployées pour cette fin, telles que le coloriage, la cartographie thématique et les graphes sémantiques. Un exemple d'une recherche *Google* des termes "information retrieval" sous *TouchGraph* est présenté dans Figure 2-5.

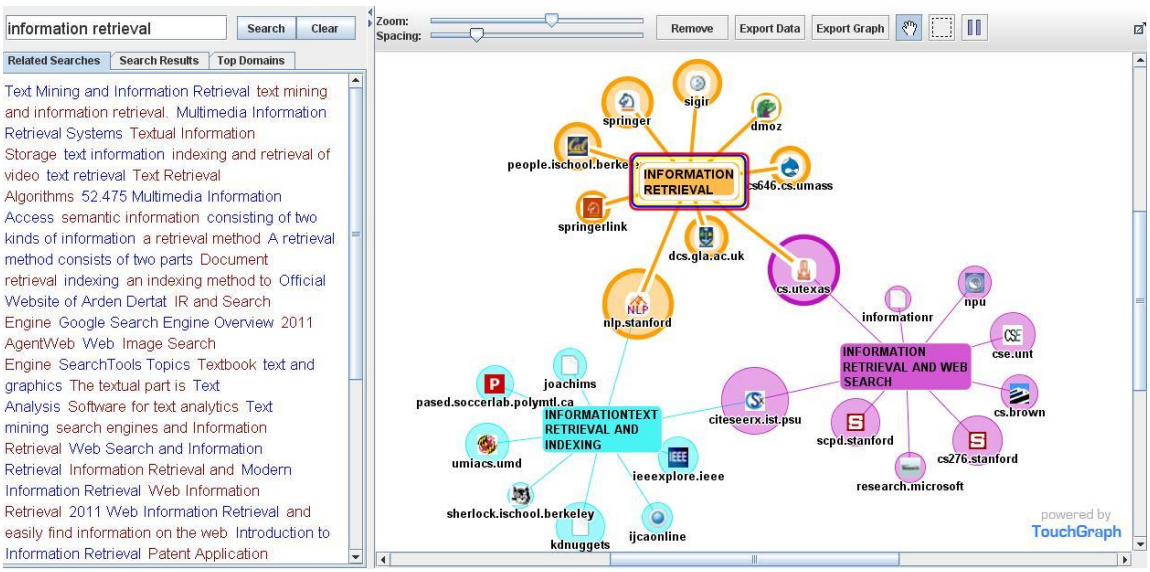


Figure 2-5. Visualisation graphique des résultats d'une recherche par *TouchGraph*.

4.3 Autres formes de recherche

Bien que la recherche sur le Web représente, de nos jours, la forme dominante des applications de recherche d'information, deux autres formes de recherche sont aussi importants et partagent la majorité des fonctionnalités précédentes. Il s'agit de la recherche locale (ou recherche de bureau) et la recherche d'entreprise dans les bibliothèques électroniques de son réseau Intranet [Büttcher et al., 2010].

Pour la première (*desktop search*), nous trouvons des logiciels installés sur un ordinateur personnel qui combinent la recherche dans les fichiers locaux sur le PC et la recherche dans les sites Web. Nous citons à titre d'exemple *Google Desktop* et *Copernic Desktop Search*.

Pour la deuxième forme, les applications de recherche, dans le fond documentaire de l'entreprise, ont pris leur place parmi les outils vitaux de gestion et d'analyse. Ces applications se sont dotées des atouts des nouvelles technologies du Web et des méthodes modernes de recherche d'information afin d'élever le niveau de concurrence de l'entreprise. Des systèmes intelligents de fouilles de données, notamment de text-mining, leur sont associées dans un cadre de solution intégrée allant de la gestion classique à la veille stratégique et passant par le marketing et le commerce électronique. A titre d'exemple, nous citons dans cette catégorie *Fast* et *Google Search Appliance*.

Par ailleurs, On trouve également des méta-moteurs, c'est-à-dire des sites Web où une même recherche est lancée simultanément sur plusieurs moteurs de recherche (les résultats étant ensuite fusionnés pour être présentés à l'internaute) — on peut citer *Mamma*, *Seek* et *Kartoo*. Ce dernier (*Kartoo*) présentait ses résultats de recherche sous forme de carte graphique avant de s'éteindre en 2010.

5 La catégorisation des textes

La tâche de catégorisation des textes consiste à classer automatiquement un ensemble de documents selon des catégories prédéfinies. Cette approche est attractive du moment où elle décharge les organisations des tâches fastidieuses de classification manuelle des documents [Sebastiani, 2005]. Elle représente une alternative efficace, à titre de pré-organisation, des catalogues de recherche dans les bibliothèques électroniques.

La classification automatique des documents fait intervenir les méthodes d'apprentissage automatique avec les modèles de la RI en vue d'offrir à l'utilisateur une navigation thématique guidée dans les bibliothèques électroniques spécialisées. Les méthodes d'apprentissage supervisé (binaire, multi-classe ou multi-étiquette) trouvent dans les applications de catégorisation de texte l'un des premiers challenges pour évaluer leur performance. La classification Bayésienne naïve, les réseaux de neurones ou les machines à vecteurs de support (SVM) constituent des techniques très répandues pour cette tâche [Kaufmann, 1997].

Pour résoudre le problème de classification supervisée, plusieurs approches ont été proposées et appliqués parmi lesquels nous citons [Brahmi, 2005] :

- Les méthodes à base de modèle statistique tel que les réseaux Bayésiens (RB) et le modèle de Markov (HMMs),
- Les méthodes d'apprentissage par approche connexionniste telles que les réseaux de neurones (RNs),
- Les méthodes d'apprentissage à base de noyau telles que les séparateurs[5] à vaste marge (SVM) et le classifieur à moindre carrés régularisés (RLSC).

Nous dressons dans Tableau 2-5 un résumé comparatif, des méthodes de classification de la recherche d'information, extrait d'une étude réalisée dans le cadre de notre projet en magister [Brahmi, 2005].

[5] Transformation sémantique de l'anglais du nom de la méthode "Support Vector Machine"

	Taille / Dimension (des données)	**Garantie théorique**	**Paramètres d'ajustement**	**Utilisation des outils externes**	**Aspect temps réel, implémentations**
Réseaux de Bayes	- / -	+	+	+	-
HMMs	- / -	+-	++	+	-
Réseaux de neurones	+ / -	-	+++	-	+
SVM et RLSC	- / +	++	+	-	-

Tableau 2-5. Comparatif des méthodes de classification.

Sans rentrer dans le détail des justifications théoriques et les applications de chaque modèle, nous nous contentons de résumer ce comparatif par les remarques suivantes [Brahmi, 2005] :

- Les RNs et les méthodes à base de noyau (SVM et RLSC) n'utilisent aucun outil supplémentaire externe, à la différence des autres approches.
- Les méthodes à base de noyaux prennent mieux en charge la haute dimensionnalité des exemples d'apprentissage. Ceci les favorise dans la tâche de catégorisation des textes où les documents sont généralement représentés dans l'espace des mots dont la dimension atteint des dizaines de milliers.
- Les RNs sont les moins sensibles à la taille élevée (nombre des exemples). Les SVMs peuvent traiter les collections volumineuses dans la mesure où l'apprentissage est formulé par la résolution d'un problème quadratique généralement creux. Le RLSC par contre, est légèrement moins pratique sur ce point puisqu'il doit résoudre un problème linéaire avec une matrice de Gram généralement dense. Le choix convenable de certains noyaux avec des techniques adaptées de résolution ont été proposées pour l'application efficace des classifieurs régularisés dans les problèmes de la RI [Brahmi et Ech-Cherif, 2005].
- Bien qu'on puisse minimiser l'erreur empirique sur les exemples d'apprentissage, aucune garantie théorique de généralisation n'est offerte par les modèles connexionnistes. Les approches statistiques et les méthodes à base de noyaux prennent en considération cet aspect puisqu'ils reposent sur de bonnes bases mathématiques.
- Nombreux sont les paramètres d'ajustement dans l'apprentissage neuronal. Les autres approches sont moins ennuyeuses sur cet aspect.
- Les RNs sont relativement plus adaptés dans les applications exigeant des réponses en temps réel. Néanmoins, cet aspect n'est pas assez contraignant en catégorisation des textes.

Nous optons dans notre étude d'utiliser la méthode SVM pour la classification des documents. Nous présentons dans la section (7.4) certaines mesures pour évaluer la performance d'un classifieur SVM.

6 Autres applications relatives à la RI

Avec le développement des nouvelles technologies de l'information, la *RI* moderne ne cesse de proposer des modèles de plus en plus efficaces couvrant l'indexation, la classification et la visualisation de l'information. Des méthodes issues du traitement automatique du langage naturel représentent des outils indispensables pour toute application RI. La composition de tout système de recherche d'information doit être vue dans un contexte plus général de l'analyse automatique du texte. Les fonctions telles que la meta-recherche, la visualisation des résultats, le suivi des thèmes, le text-mining ou la recherche sémantique partagent souvent des modèles communs pour l'organisation et l'indexation du contenu textuel (voir Figure 2-6).

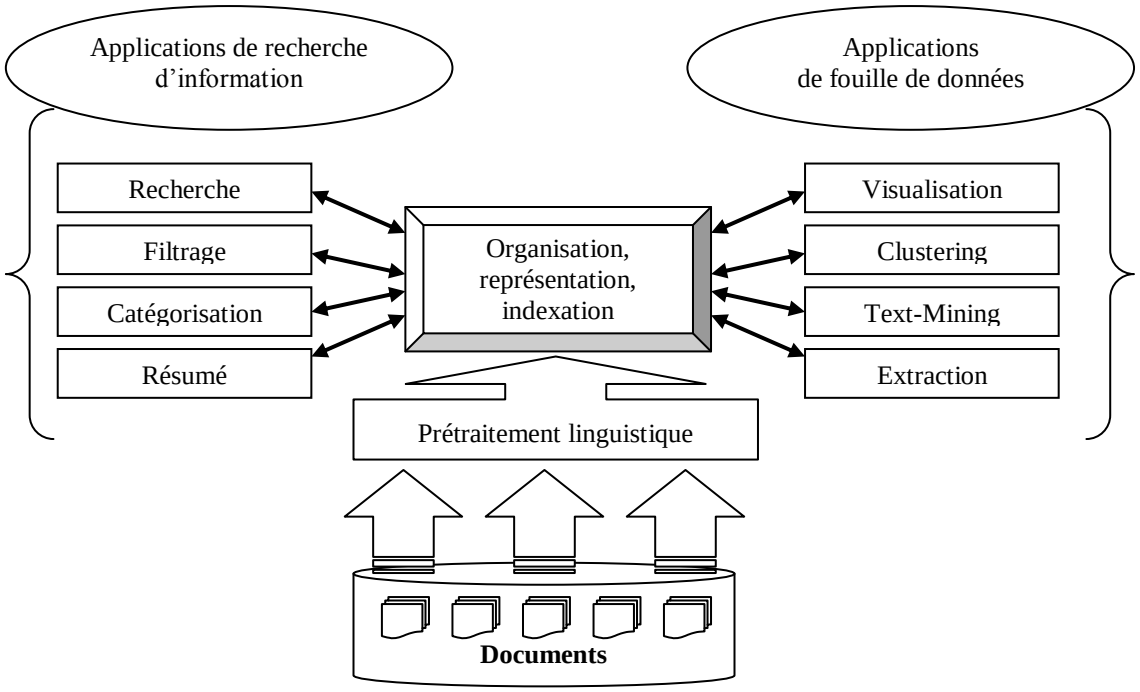


Figure 2-6. Fonctions d'analyse automatique du contenu textuel.

6.1 Le filtrage

Le filtrage d'information vise à extraire au sein d'un important volume d'informations générées dynamiquement, les documents susceptibles de correspondre aux intérêts de l'utilisateur. L'élimination du courrier indésirable, le blocage des expéditeurs malveillants ou le filtrage des sites sensibles représentent des domaines d'application très sollicitées. Le filtrage intègre aussi les opérations d'exploitation et de présentation des résultats. Les sources de filtrage sont dynamiques et évoluent dans le temps.

Contrarient à une recherche ad-hoc, l'outil de filtrage permet de repérer exclusivement les documents relatifs aux centres d'intérêt (requêtes préétablies) formulés par l'utilisateur sous forme de sélection thématique prédéfinie [Büttcher et al., 2010].

Deux types de filtrage peuvent être cités dans ce contexte : le filtrage par contenu et le filtrage collaboratif. Ce dernier est devenu un sujet de recherche d'actualité pour répondre aux besoins grandissants dans le commerce électronique et les réseaux sociaux. Il repose sur les opinions d'un groupe d'utilisateurs pour recommander les objets (produits, amis, sites …etc.) les appropriés. Un aperçu détaillé des techniques utilisées pour cette approche peut être consulté dans [Su et Khoshgoftaar, 2009].

6.2 Le résumé automatique

Un résumé est une transformation réductive du texte source par sélection et/ou généralisation de ce qui est le plus significatif du texte original. Trois étapes essentielles doivent être considérées dans cette approche : l'identification des thèmes, l'interprétation et la génération [Lloret et Palomar, 2010].

La détection des phrases pertinentes dans le texte constitue la fonction clé dans ce genre de système qui épuise des techniques du traitement automatique du langage naturel et offre une alternative efficace pour la consultation des quantités immenses d'information en-ligne. Le résumé automatique peut être réalisé à partir d'un seul document comme il peut être multi-document. Bien que l'appréciation des résumés générés soit assez sensible et compliquée, la méthode d'évaluation ROUGE est adoptée par la majorité de chercheurs.

6.3 L'extraction des entités nommées

C'est une sous-fonction de l'extraction d'information qui consiste à identifier et extraire, à partir d'un texte non-structuré, les unités lexicales correspondant aux noms des personnes, endroits, organisation et même des expressions numériques. Quatre approches peuvent être considérées pour réaliser un tel système : la représentation des segments textuels, les algorithmes d'inférence, l'utilisation des caractéristiques non-locales et l'exploitation des connaissances externes [Ratinov et Roth, 2009].

Groupant les techniques d'apprentissage automatique et le traitement automatique du langage naturel, la reconnaissance des entités nommées représente un outil indispensable pour le repérage et l'accès aux documents pertinents. Par ailleurs, elle constitue une phase incontournable dans toute analyse efficace du texte en vue de l'indexer dans un système de recherche.

6.4 Autres applications

De façon non exhaustive, nous citons autres tâches relatives qui couvrent certains aspects de la RI tels que l'organisation, la visualisation et la manipulation [Büttcher et al., 2010]:

- L'agrégation (clustering) des textes regroupe les documents selon leurs propriétés communes. A la différence de la catégorisation, le clustering ne dispose d'aucune information préalable relative aux caractéristiques des classes à construire. Les méthodes d'apprentissage non supervisé sont largement utilisées dans ce genre d'applications.
- La détection et le suivi des thèmes est la tâche qui permet de repérer les événements sur des flux d'articles de presse ou d'autre source d'information similaires.
- Le système de question-réponse intègre des informations collectées de sources différentes en vue de fournir des réponses concises. Afin de répondre aux questions spécifiques, un tel système déploie d'autre techniques de la RI comme la recherche, le résumé automatique et l'extraction.
- La recherche d'information multimédia étend le principe de classement de pertinence ainsi que d'autre techniques de la RI vers d'autre types d'information tels que l'image, la vidéo, la musique et la parole.

7 Evaluation

Depuis l'émergence des premiers modèles pour la recherche d'information, l'évaluation objective de leur efficacité représentait une pièce-maitresse dans le développement du domaine. Il était évident pour les chercheurs la nécessité de trouver des mesures standards pour estimer la qualité des résultats de recherche. Encore fallait-il élaborer et s'entendre sur des collections représentatives pour les tests d'évaluation. L'approche expérimentale qui avait été proposée et réalisée, au cours des années 1960, sur la collection *Cranfield* avait permis d'aboutir vers le jeu de caractéristiques souhaité pour un système de recherche [Cleverdon, 1967]. Bien qu'il y'a eu quelques débats durant des années, la communauté de la RI a approuvé deux mesures de base qui sont la précision et le rappel [Singhal, 2001].

7.1 Précision et Rappel

Pour une requête appliquée sur un ensemble de documents, nous identifions d'une part, l'ensemble des documents initialement pertinents et d'une autre part, l'ensemble des documents trouvés par le système. Ainsi les deux mesures seront définies comme suit :

La précision est définie par le taux des documents pertinents dans ceux qui sont trouvés.

$$\text{Precision } (P) = \frac{\#\{\text{documents pertinents trouvés}\}}{\#\{\text{documents trouvés}\}}$$

Le rappel peut être défini par le taux des documents trouvés parmi ceux qui sont pertinents.

$$\text{Rappel } (R) = \frac{\#\{\text{documents pertinents trouvés}\}}{\#\{\text{documents pertinents}\}}$$

Ces notions peuvent être clarifiées dans Figure 2-7 où nous définissons quatre quantités :

- TP (*vrai positif*) : les pertinents correctement trouvés par le système.
- TN (*vrai négatif*) : les non pertinents correctement exclus.
- FP (*faux positif*) : les non pertinents mais trouvés à tord.
- FN (*faux négatif*) : les non pertinents mais exclus à tord.

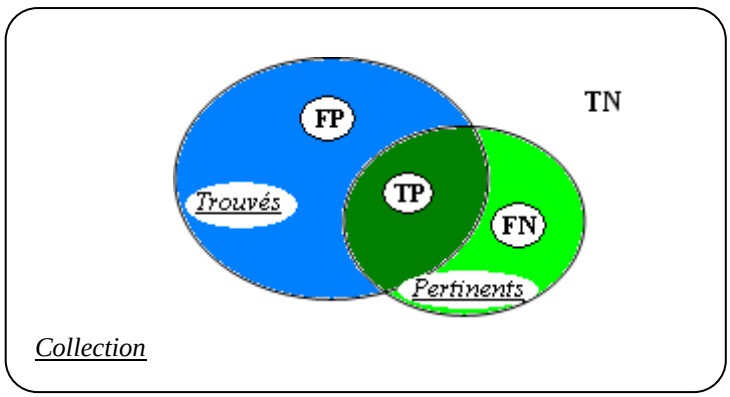


Figure 2-7. Représentation des éléments d'évaluation dans une recherche ad-hoc.

Ainsi, les deux mesures seront définies comme suit :

$$P = \frac{TP}{TP + FP} \quad \textbf{(2-1)}$$

$$R = \frac{TP}{TP + FN} \quad \textbf{(2-2)}$$

D'autres mesures peuvent être déduites des quatre quantités (*TP*, *TN*, *FP*, *FN*), parmi lesquelles, il existe une alternative naïve, celle de calculer le taux de reconnaissance pour évaluer la performance d'un système de recherche. Cette quantité est donnée par l'expression :

$$\text{Taux de reconnaissance}(TR) = \frac{TP + TN}{TP + TN + FP + FN} \quad \textbf{(2-3)}$$

Ceci semble plausible du moment où nous pouvons considérer, d'une part, deux classes (pertinent et non-pertinent) et, d'autre part, un jugement du système qui tente d'étiqueter les éléments trouvés comme pertinents. Bien qu'elle soit souvent utilisée dans l'évaluation des méthodes de classification par apprentissage supervisé, la mesure (*TR*) est loin d'être appropriée au problème de la recherche ad-hoc dans une grande masse de données. Dans la majorité des situations, les données sont extrêmement biaisées où plus de 99,9% des documents appartiennent à la catégorie des non-pertinents [Manning et al., 2008].

7.2 F-mesure

Un système de recherche conçu pour maximiser le taux de reconnaissance peut apparaître très performant en jugeant simplement tous les documents comme non-pertinents. Par ailleurs, un simple internaute espère avoir les documents qui l'intéressent sur la première page (haute précision) mais ils n'a pas le moindre intérêt pour connaître ou consulter ceux qui sont pertinents. En revanche, certains professionnels souhaitent, au détriment d'une faible précision, n'écarter aucune pièce pertinente à leur requête (rappel supérieur). Il est clair que la maximisation des deux mesures soit souvent contradictoire et donc, il est envisageable d'établir un compromis entre les deux quantités. *Van Rijsbergen* proposa de pondérer précision et rappel par une mesure relative de l'efficacité (E) [Rijsbergen, 1976]:

$$E_\alpha = 1 - \frac{1}{\alpha\left(\frac{1}{P}\right) + (1-\alpha)\frac{1}{R}}$$

Où α est un paramètre variant de 0 (aucune importance à la précision) à 1 (aucune importance pour le rappel). Lorsque l'utilisateur attache une importance équitable pour la précision et le rappel, on choisit α=½ et la mesure d'efficacité de *Rijsbergen* devient :

$$E_{0,5} = 1 - \frac{2PR}{P + R} \quad \textbf{(2-4)}$$

La performance d'une recherche est maximale lorsque cette mesure (E) soit proche de zéro. Elle est moins bonne avec des valeurs supérieures (voire proches de 1).

En pratique, les chercheurs utilisent F-Mesure qui représente la moyenne harmonique pondérée de précision et rappel et est définie comme suit [Manning et al., 2008] :

$$F_\beta = 1 - E_\alpha = \frac{(\beta^2 + 1)PR}{\beta^2 P + R}, \qquad \textbf{où :} \ \beta^2 = \frac{1-\alpha}{\alpha} \tag{2-5}$$

Cette mesure, dérivée de celle de *Rijsbergen*, est aussi connue par F-score ou même encore par F_1-score pour une pondération équitable entre précision et rappel ($\beta=1 \Leftrightarrow \alpha=0,5$). β peut prendre toute valeur réelle en donnant plus d'importance au rappel pour les valeurs $|\beta|<1$ alors que celles de $|\beta|>1$, focalisent beaucoup plus sur la précision $|\beta|<1$ [Büttcher et al., 2010].

7.3 Evaluation d'une recherche ordonnée

L'objectif fondamental de l'ordonnancement par degré de pertinence est souvent exprimé en termes de principe d'ordonnancement de probabilité ; Si pour chaque requête, la réponse d'un système de recherche est un classement dans une collection des documents par ordre décroissant de probabilité de pertinence, alors l'efficacité globale du système et maximale [Büttcher et al., 2010].

Les mesures citées ci-dessus sont calculées à base d'une vue ensembliste. Aucune hypothèse d'ordonnancement n'était introduite. Or la majorité des moteurs de recherche adopte la présentation des résultats selon un ordre décroissant de pertinence. De plus, l'utilisateur espère trouver ce qui l'intéresse dans les 10 premiers résultats.

L'efficacité d'une recherche sur des résultats ordonnés peut être estimée par l'extension des notions de précision et de rappel pour ne considérer que les *k* premiers documents trouvés. Une courbe de précision et de rappel peut être tracée en évaluant chaque ensemble de *k*-résultats. Il est clair que le rappel reste le même si le $(k+1)^{ème}$ document trouvé n'est pas pertinent mais il augmente avec la précision dans le cas contraire.

Afin d'éliminer cette perturbation en se limitant sur les *k* premiers résultats, une mesure, dite précision interpolée, a été introduite en retenant la précision supérieure pour chaque niveau de rappel $R' \geq R$. R étant un niveau de rappel fixé [Manning et al., 2008] :

$$P_{\text{interp}}(R) = \max_{R' \geq R} P(R)$$

Ceci peut être justifié par le fait qu'un utilisateur est prédisposé à consulter quelques documents supplémentaires en espérant satisfaire sa requête.

Bien que l'examen d'une telle courbe puisse être informatif sur l'efficacité d'un système de recherche, il est préférable parfois de lire une seule valeur. La précision moyenne interpolée à onze (11) points est une mesure pratique pour l'évaluation des résultats ordonnés d'une recherche. Le principe est de calculer, pour chaque requête, la moyenne arithmétique de la précision interpolée pour les 11 niveaux de rappel (0,0 ; 0,1 ; … 1,0).

7.4 Evaluation de la catégorisation

Le principe d'évaluation d'une recherche ad-hoc non ordonnée peut être facilement appliqué dans un problème de classification binaire. Il suffit, d'un côté, de considérer l'ensemble des documents pertinents comme étant initialement positifs, les négatifs interprètent les non-pertinents. D'un autre côté, les éléments trouvés sont assimilés à ceux classés positifs, les autres, non trouvés, étant négatifs. Ainsi les quantités (*TP*, *TN*, *FP*, *FN*) auront leur interprétation intuitive dans un contexte d'apprentissage supervisé pour une classification binaire.

7.4.1 Précision et rappel

Les mesures de précision, de rappel et de F-mesure garderont les mêmes définitions citées ci-dessus. Selon la nature des données (proportion des positifs dans l'ensemble de la collection), le taux de reconnaissance peut évaluer l'efficacité d'une catégorisation binaire (voir la discussion à la fin de la section 7.1).

Par ailleurs, dans une catégorisation multi-classe, le modèle d'évaluation peut être décomposé en *k*-évaluations binaires (*k* étant le nombre de classes). Pour chaque catégorie (i=1...k), nous calculons les quatre quantités (TP_i, TN_i, FP_i, FN_i). La mesure globale, de la précision et du rappel, peut être estimée soit par micro-moyenne soit par macro moyenne comme définies dans Tableau 2-6 [Sebastiani, 2005].

	Micro-moyenne	Macro-moyenne
Précision	$P_{micro} = \frac{\sum_{i=1}^{k} TP_i}{\sum_{i=1}^{k} TP_i + FP_i}$	$P_{macro} = \frac{1}{k}\sum_{i=1}^{k} \frac{TP_i}{TP_i + FP_i}$
Rappel	$R_{micro} = \frac{\sum_{i=1}^{k} TP_i}{\sum_{i=1}^{k} TP_i + FN_i}$	$R_{macro} = \frac{1}{k}\sum_{i=1}^{k} \frac{TP_i}{TP_i + FN_i}$

Tableau 2-6. Moyennes de précision et rappel pour une catégorisation à *k* classes.

Cette approche inclut les situations des documents multi-étiquetés du moment où les valeurs de base sont calculées séparément pour chaque classe. Le choix entre micro et macro moyenne dépend de la nature des données manipulées et de l'objectif de l'application elle même. Alors que la macro-moyenne pondère les catégories équitablement, la micro-moyenne donne plus d'attention à la classification individuelle de chaque document et par conséquent, les grandes classes dominent les plus petites [Manning et al., 2008].

7.4.2 Validation croisée

Par ailleurs, l'évaluation d'un classifieur nécessite la décomposition de la collection en deux sous-ensembles principaux : le premier pour l'apprentissage et le deuxième pour le test. Dans certains cas, il est préférable de réserver une troisième partie pour la validation. Ce découpage est subjectif et peut perturber l'estimation des paramètres d'apprentissage. Par conséquent, l'évaluation des performances d'un classifieur sur la même collection compromise par cette sélection.

Une technique est recommandée à cet effet qui consiste à opérer une validation croisée pour l'évaluation de la classification. Son principe est d'effectuer un découpage aléatoire (échantillonnage) pour retirer un sous-ensemble de données et les réserver à la phase du test. L'opération est répétée (*n*) fois afin de récupérer une moyenne de la mesure d'évaluation. On dénote ce processus par (*n*-fold) pour exprimer une validation croisée à (*n*) itérations.

Cette technique peut être étendue en retirant un seul exemple et laisser les (*M*-1) exemples pour l'apprentissage. L'évaluation est donc estimée par la moyenne de (*M*-1) tests indépendants. Néanmoins, la méthode *LOO* (*en angl. Leave-One-Out*) s'avère plus couteuse en temps de calcul et on préfère utiliser la validation croisée par (*n*-fold) pour *n*=3…10.

Le taux de reconnaissance, selon l'équation (2-3), obtenu par une validation croisée donne généralement une évaluation standard et satisfaisante de la catégorisation binaire ou multi-classe.

7.4.3 Ratio des vecteurs de support

La méthode SVM parte d'un fondement théorique solide (modèle statistique) pour modéliser le problème d'apprentissage binaire (puis multi-classe) dans un ensemble de données linéairement séparables. L'astuce du noyau et l'augmentation de la dimension permettent de traiter les cas non-séparables [Vapnik, 1995] [Brahmi, 2005].

Les vecteurs de support constituent les exemples constituant l'enveloppe convexe des classes. Vapnik donne une borne alternative du risque empirique lors de l'apprentissage de SVM sur un ensemble donnée [Vapnik, 1995]:

$$E[P(erreur)] \leq \frac{E[\text{nombre de vecteurs de support}]}{\text{nombre des exemples d'apprentissage}} = r\text{-}SV \quad \textbf{(2-6)}$$

Où $E[P(erreur)]$ est l'estimation du risque empirique sur les (M-1) exemples d'apprentissage. E[nombre de vecteurs de support] est l'estimation du nombre des vecteurs de supports dans une validation croisée par *LOO*. Le ratio des vecteurs de support *r-SV* dans l'équation (2-6), par rapport au nombre des exemples d'apprentissage, donne une indication sur la capacité de généralisation dans un apprentissage par SVM. Moins est ce ratio, plus est la performance de capacité de prédiction. Néanmoins, certaines études ont contesté cette intuition et préfère considérer le principe de minimisation du risque structurel [Burges, 1998].

De notre part, nous optons pour l'utilisation de cette mesure *r-SV* à titre d'indication empirique de la performance d'une classification SVM. En la combinant avec les autres mesures des sections précédentes, nous impliquons *r-SV* afin d'arbitrer certains modèles qui fournissent des évaluations similaires.

8 Corpus de test

L'évaluation des méthodes en RI nécessite un jeu de données (collection de documents) représentatif ayant des jugements de référence. Parmi les défis majeurs, pour développer l'approche expérimentale dans le domaine, était de mettre à la disposition des chercheurs des corpus standard appropriés aux tâches diversifiées de recherche d'information.

Le terme "corpus" peut faire référence à tout ensemble de textes sous forme numérique. Mais pour un linguiste ou toute autre personne impliquée dans la description générale de la langue, un corpus est un vaste échantillon représentatif illustrant le répertoire complet des types de textes dans une langue, comme les romans, le journalisme, l'écriture académique et de nombreuses autres variétés de texte [Evans, 2007]. On peut distinguer plusieurs types de corpus linguistiques :

- ***Le corpus général (de référence)*** : c'est une collection très large et assez diversifiée pour capturer les différents cas d'utilisation de la langue. Un exemple type est celui du corpus national britannique (BNC) qui continent prés de 100 millions de mots et dont l'objectif est de représenter l'univers de l'anglais britannique contemporain.
- ***Le corpus spécialisé*** : sa taille est généralement moins importante et contient des textes d'un type particulier ou relatifs à une période ou un contexte spécifique.
- ***Les corpus comparables*** : réfèrent à deux ou plusieurs collections construites selon des paramètres similaires mais dans des langues différentes.
- ***Les corpus parallèles*** : similaires aux corpus comparables sauf qu'ils sont alignés de telle sorte que la recherche d'un terme dans une langue puisse récupérer les documents relatifs avec toutes les traductions équivalentes.

- ***Le corpus historique (diachronique)*** : c'est une collection de textes de différentes périodes. Il permet d'étudier les changements de langue au fil du temps. Par exemple, on peut citer le corpus ARCHER qui contient 1,7 millions de mots de l'anglais couvrant les périodes entre 1650 et 1990.
- ***Le corpus moniteur*** : c'est une collection qui est complétée régulièrement par de nouveaux textes. Ceci est fait de telle manière que la proportion des types de texte reste constante ce qui signifie que chaque nouvelle version du corpus est comparable à toutes les versions précédentes.

8.1 Caractéristiques d'un corpus

La crédibilité des expérimentations relatives au traitement automatique de la langue naturelle dépend en partie de la nature des collections de test utilisées. L'évaluation des méthodes d'analyse automatique du texte, dont les tâches de RI en font partie, nécessite l'élaboration préalable de corpus appropriés respectant les critères suivants :

8.1.1 La taille

Le corpus doit évidemment atteindre une taille assez suffisante pour permettre des traitements statistiques fiables. Il est devenu primordial d'apprécier toute méthode d'analyse de texte dans deux configurations possibles : la première traite une collection de taille modérée (des centaines ou quelques milliers de documents). La deuxième concerne l'utilisation d'une collection à grande échelle (des dizaines de milliers voire des centaines de milliers et plus). La taille du document lui-même (nombre de mots ou de phrase) peut aussi être déterminante de la qualité du test dans certaines situations.

8.1.2 La représentativité

La représentativité se réfère à un échantillon de texte incluant la variabilité complète dans une population. Ainsi, l'élaboration d'un corpus linguistique peut être évaluée par la gamme de types de texte dans une langue et de la variation des distributions linguistiques [Biber, 1993]. Il est nécessaire, pour les linguistes, de sélectionner des textes à partir de plusieurs domaines différents (rapports de presse, fiction romanesque, statuts juridiques, rédaction scientifique, poésie, …etc.) pour bien représenter la langue.

8.1.3 La forme lisible par la machine

De nos jours, le terme "corpus" implique souvent le sens "lisible par la machine". Ce ne fut pas toujours le cas, dans le passé le mot "corpus" n'a été utilisé en référence qu'au texte imprimé. Les données des corpus sont parfois disponibles dans d'autres formes de médias comme les microfiches (concordance complète de mot-clé du corpus LOB, corpus parlé d'anglais par Lancaster / IBM). La lisibilité numérique des corpus permet un traitement accéléré et facilite son enrichissement. La mise à disposition des routines normalisées d'accès aux données du corpus est une option facultative mais très sollicitée par les chercheurs.

8.1.4 La standardisation

Il est souvent souhaitable qu'un corpus constitue une référence standard pour la variété de langue(s) qu'il représente. Cela suppose que le corpus sera mis à la disposition d'autres chercheurs. Un avantage d'un corpus largement diffusé fournit une mesure à laquelle les études successives peuvent être comparées, tant que la méthodologie n'est pas changée, de nouveaux résultats sur des sujets connexes peuvent être directement comparés avec les résultats déjà publiés sans avoir besoin de les recalculer. En outre, un corpus standard signifie aussi qu'une base de données continue soit utilisée, ceci implique que toute variation entre les résultats des études différentes serait à la cause de la pertinence des hypothèses et de la méthodologie, moins que la différence dans les données.

Parmi les autres caractéristiques on peut ajouter le ***temps*** couvert par les textes du corpus, car le temps joue un rôle important dans l'évolution du langage, par exemple le français parlé aujourd'hui ne ressemble pas au français parlé il y a 200 ans, ni au français parlé il y a 10 ans, à cause notamment des néologismes. C'est un phénomène à prendre en compte pour toutes les langues vivantes. Donc un corpus ne doit pas contenir de textes rédigés à des intervalles de temps trop larges.

Aussi, il ne faut pas non plus mélanger des catalogues différents de plusieurs domaines. Par exemple un corpus construit à partir de textes scientifiques ne peut être utilisé pour extraire des informations sur les textes littéraires, et un corpus mélangeant des textes scientifiques et littéraires ne permettra de tirer aucune conclusion crédible, il faut que le corpus soit ***homogène*** [Spousta, 2006].

D'autres linguistes introduisent le critère de ***balance***. La taille du texte, dans les différentes catalogues, doit être presque la même ; la mesure généralement recommandée est estimée entre 2,000 et 5,000 mots [Evans, 2007].

8.2 Construction des corpus

Les corpus traditionnels pour la recherche en linguistique sont créés principalement à partir de textes imprimés, comme des articles de journaux et de livres. Aujourd'hui, la plupart des corpus sont construit à partir du texte qui est déjà numérisé, le coût de mise en forme de texte électronique qui n'existe que sur papier est beaucoup plus cher que le coût d'une simple copie par téléchargement et la collection de données qui sont déjà numérisés.

Il y a plusieurs avantages à la création d'un corpus à partir des données sur le Web, plutôt que du texte imprimé. Toutes les données sur le Web sont déjà sous forme électronique et donc lisible par les machines, alors que les données imprimées ne sont pas toujours disponibles sous forme électronique. La grande quantité de texte disponible sur le Web (plus de 5 millions Téra octet) est un avantage majeur qui permet de collecter des quantités suffisante dans n'importe qu'elle domaine.

8.2.1 Techniques de Crawling

Afin de construire un corpus à partir du Web, plusieurs questions et problèmes techniques doivent être résolus. Tout d'abord, le Web Crawler doit être choisi par rapport au volume de données de texte que nous voulons obtenir. Pour certaines tâches, il suffirait de télécharger quelque milliers de documents, pour d'autres tâches, il faudrait une quantité énorme de données. Les données brutes doivent passer par plusieurs filtres comme le recodage de caractères, détection de la langue, la conversion de format et l'extraction de texte. Le format cible doit être correctement sélectionné selon le style prévu pour l'accès aux données.

Il y a de nombreux défis dans la création d'un corpus Web. Comme le contenu du Web n'est pas structuré et ne possède pas de répertoire définitif, aucune méthode simple et directe n'existe pour rassembler un vaste échantillon représentatif depuis le Web. Une approche radicale consiste à explorer le Web à travers l'échantillonnage des adresses IP du réseau mondial (Internet). Cette méthode requiert des ressources considérables car de nombreuses tentatives sont faites pour chaque site Web trouvé. En répertoriant près de 16 millions de serveurs, une étude a signalé qu'un seul serveur Web a été trouvé sur 269 essais [Lawrence et Giles, 2000].

Nous résumons, dans ce qui suit, l'essentiel des approches d'exploration du Web pour en concevoir un corpus ou indexer son contenu :

1. ***Exploration en largeur d'abord*** : en visitant, à partir d'un ensemble d'adresses initiales, toutes les pages Web du premier niveau avant d'explorer le niveau suivant [Najork et Wiener, 2001].
2. ***Exploration du meilleur d'abord*** : en sélectionnant, dans une frontière de liens, les meilleures pages selon certains critères plus ou moins complexes [Cho et al., 1998].
3. ***Exploration du plus important*** (ou PageRank): en choisissant les pages ayant les plus grand scores de référencement récursive [Brin et Page, 1998].
4. ***Marche aléatoire*** : en échantillonnant un ensemble uniforme de nœuds dans un graphe du Web (ou dans une version simulée) [Henzinger et al., 2000].
5. ***Exploration thématique*** (ou ciblée) : en appliquant des heuristiques pour sélectionner un certain type de pages sur un sujet précis ou dans une langue particulière [Mencser et al., 2003].

8.2.2 Fonctionnement d'un moteur d'exploration du Web

Rassembler les documents Web nécessite un robot, généralement appelé *Web-Crawler*. Ce programme est conçu pour télécharger les documents à partir du serveur Web donné, les sauvegarder et les traiter, afin de découvrir des liens vers d'autres documents ou sites Web. Cette tâche est plus difficile que l'on peut s'attendre, le Crawler doit résoudre de nombreux défis pour une exploration efficace: téléchargement parallèle massive, traitement rapide de page, les enregistrements de domaines (DNS) de pré-recherche, la limitation de débit, éviter les boucles infinies, etc.

La première tâche dans le Crawl du Web consiste en l'***acquisition*** des documents texte dans la langue désirée. Les pages Web appropriées sont présentes sur des serveurs Web dans divers domaines, nous devons sélectionner les domaines à inclure dans la recherche. Nous pouvons limiter notre attention aux serveurs Web en fonction du domaine de premier niveau (p. ex *.dz* pour l'Algérie, *.fr* et *.es* pour les domaines français et espagnol). Cette approche convient pour les principales langues parlées dans un pays. Une autre option est de traiter tous les noms d'un domaine et filtrer les pages Web appropriées en utilisant une fonction de détection automatique de la langue. Dans les deux cas nous avons besoin d'un bon point de départ (*seed*) qui peut être obtenu depuis un catalogue (*dmoz.org* ou *Yahoo.com*) de pages Web ou des résultats des moteurs de recherche pour un terme générique dans la langue désirée.

Une deuxième étape assez cruciale dans la construction de corpus est le "***nettoyage*** des données". Il impératif dans cette phase de uniformiser le codage des caractères de pages (UTF-8 par exemple) et de traduire les entités HTML dans leurs caractères équivalents. Les limites de la phrase doivent être identifiées. Les phrases grammaticales doivent être séparées des phrases non grammaticales.

En troisième phase, les documents collectés doivent être correctement répertorié dans un processus d'***annotation***. Il s'agit d'ajouter toute information utile pour le repérage et la classification du document. L'en-tête d'une page Web contient souvent des méta-données comme le nom de l'auteur, le titre de l'œuvre, l'année de publication et même parfois des étiquettes de référencement. L'inclusion de ces informations rend le corpus, une plate-forme de recherche beaucoup plus puissante, car les éléments de la langue dans un corpus peuvent être étudiés non seulement de façon linguiste pure, mais aussi en tant que phénomènes sociaux. La méthode et le style des annotations appliquées dépendent des tâches utilisatrices notamment dans l'indexation d'une base d'un moteur de recherche ou d'un catalogue à classification thématique.

8.2.3 Exemples de moteurs de Crawling

Plusieurs Web Crawler appropriées existent, certains d'entre eux font partie des moteurs de recherche Internet, d'autres sont des robots indépendants pour collecter les pages Web. A titre d'exemple, nous citons :

- ***Heritrix***[6] est le moteur de Crawl du site Internet Archive, écrit en Java, utilise une interface Web de contrôle sous la licence GPL.
- ***Egothor***[7] est un moteur de recherche Web à haute performance, développé dans l'université de Charles à Prague. Ce Crawler est l'un des plus rapides Crawler, mais la version actuelle n'est pas accessible au public.
- ***Sherlock-Holmes***[8] est une implémentation complète d'un moteur de recherche Web, écrit en langage C, sous deux licences GNU GPL et commerciale. Holmes a une architecture modulaire et comprend le jeu de caractères et les filtres de détection de la langue. Une partie du programme contient des outils d'indexation et de recherche [Spousta, 2006].

8.3 Quelques corpus de référence

Il est toujours souhaitable d'évaluer une méthode de recherche sur une collection standard. Pour chaque tâche en RI, on trouve des corpus de référence appropriés qu'il est intéressant de citer.

8.3.1 Cranfield en premier

Le corpus Cranfield, relative aux publications des revues scientifiques d'aéronautique, contient 1398 résumés et 225 requêtes avec des jugements de pertinence pour chaque pair (document, requête). Il est considéré comme le pionnier des corpus de test dans le domaine de RI [Cleverdon, 1967]. Avec les nouveaux besoins de recherche dans les grandes collections et notamment avec l'apparition du Web, il était impératif d'élaborer des corpus de taille plus importante. La diversité des sujets, le texte non-anglais et le multilingue représentait d'autres aspects spécifiques de la RI, suscitant l'élaboration des collections de test assez conséquentes. Nous listons ci-dessous, mais de façon non exhaustive, le standard des corpus relatifs aux applications de recherche et de catégorisation des textes.

8.3.2 TREC et GOV2 pour la recherche

Initiée en 1992 par l'institut national des standards et de technologie des USA (NIST), la conférence de la recherche du texte (TREC) a conduit une série de compétitions sur diverses tâches en RI. Différentes collections spécifiques ont été élaborées et mise à la disposition des chercheurs en vue d'uniformiser l'évaluation des méthodes. En l'occurrence, l'ensemble des collections présentées dans les 8 premières conférences (entre 1992 et 1999) pour la tâche TREC-Ad-Hoc, compte près de 1,89 million de documents avec des jugements de pertinence pour 450 besoins d'information (appelés thèmes). Les documents étaient principalement collectés des articles de presse.

Afin d'offrir une plate forme de test proche de l'échelle des moteurs de recherche sur le Web, NIST a conçu ensuite le corpus GOV2 comprenant près de 25 million de page Web. La

[6] https://webarchive.jira.com/wiki/display/Heritrix/Heritrix

[7] http://www.egothor.org/

[8] http://www.ucw.cz/holmes/

qualité des résultats des méthodes de recherche n'est plus le seul critère visé mais c'est la performance globale du système, incluant l'efficacité d'indexation et le temps de réponse, qui doit être évaluée.

8.3.3 Reuters-21578 et 20Newsgroups pour la catégorisation

Pour la catégorisation des textes, Reuters-21578 est la collection la plus utilisée comptant 21.578 dépêches de presse réparties sur 118 classes. Les documents, collectés de l'agence de presse Reuters, sont multi-étiquetés mais certains n'appartiennent à aucune classe. Reuters a élaboré dernièrement un corpus plus consistant en taille et en annotation. RCV1 qui compte plus de 800.000 articles Reuters avec même des objets multimédia, représente la meilleure base de test pour les travaux futurs de recherche [Manning et al., 2008].

Une autre collection aussi importante et largement utilisée est celle de 20-Newsgroups. Elle composée de près de 20.000 articles équitablement répartis sur 20 groupes de discussion (considérés comme catégories).

8.3.4 NTCIR et CLEF pour le multilingue

Les collections, citées ci-dessus, sont toutes en anglais. Cependant, il était vital pour la communauté de la RI de créer des collections dans d'autres langues afin de répondre aux besoins d'évaluation des méthodes de recherche sur le texte non-Anglais. Deux catégories de corpus devraient être considérées, les monolingues et les multilingues (ou corpus parallèle).

Le projet de NTCIR (collections de test NII pour les systèmes RI) a conçu une variété de corpus multilingues ou parallèles. L'objectif étant d'avoir des bancs de test (benchmark) similaire à celui des conférences TREC focalisé surtout sur les langues asiatiques surtout. Les requêtes, généralement dans une seule langue, sont appliquées sur des documents d'une ou plusieurs langues. Par ailleurs, le forum CLEF (*en angl. Cross-Language Evaluation Forum*) présentait des collections de tests centrées sur les langues européennes.

8.4 Corpus arabe

Les recherches en RI, comme d'ailleurs pour les autres tâches incluant le traitement automatique du langage naturel (TALN), sont devenues plus actives dans la langue arabe. Un des défis majeurs pour la RI arabe réside dans la disponibilité des corpus de test. De nombreuses tentatives ont été menées pour construire des corpus en arabe, mais certaines ont été infructueuses et d'autres étaient pour des fins commerciales.

Un corpus de 42.591 articles arabes collectés de l'archive du journal Al-Hayat pendant l'année 1998 fut conçu pour vérifier une série d'évaluation statistiques [Goweder et De Roeck, 2001]. Les auteurs déduisirent que le texte arabe est moins dense que l'anglais. Dans une autre étude, un Crawl spécifique a été utilisé pour extraire du site Web de l'ONU[9] les documents en version bilingue (anglais, arabe). Un corpus parallèle de 38.000 paires de documents, avec plus de 50 millions de mots en anglais, fut obtenu [Xu et al., 2001]. Les auteurs l'ont utilisé pour résoudre les problèmes de pluriels en arabe. Durant la dernière décennie, une séries d'études ont été réalisées dans le domaine de la RI arabe dont chacune proposait un nouveau corpus spécifique généralement de taille modérée et rarement standard [Abdelali et al., 2005] [Al-Shammari et Lin, 2008] [Said et al., 2009].

Il est intéressant de noter que le corpus *AFP_ARB* (ou TREC-2001 arabe) est l'un des rares corpus arabes standards incluant des jugements de référence. Il contient 25 requêtes pour

[9] http://www.un.org

383.872 articles issus de l'Agence France-Presse (AFP) et couvrant la période entre 1994 et 2000 [Xu et al., 2001] [Larkey et al., 2002]. *AFP_ARB* est actuellement publié par le consortium des données linguistique (LDC) de l'université de Pennsylvanie mais avec une licence payante. Néanmoins, le corpus *AFP_ARB* n'est pas représentatif bien que sa taille soit significative [Abdelali et al., 2005]. En 2011, la cinquième version du corpus *Gigaword*[10], toujours payante, a été publiée par le même consortium incluant plus 3,3 millions d'articles extraits de 9 journaux arabes et agences de presse.

9 Conclusion

Tout au long de ce chapitre, nous avons tracé un fil d'histoire tant pratique que théorique du comportement humain pour satisfaire son besoin en organisation et accès à l'information. La pratique de classification documentaire, telle que nous la comprenons de nos jours, naquit avec les premiers travaux d'Ibn An-Nadim (987) bien avant sa standardisation par Dewey (1876) dans les bibliothèques publiques. Cette approche d'hiérarchisation de la connaissance humaine, afin de rendre plus accessible les ressources documentaires, constituait la motivation principale pour les tâches de catégorisation automatique des documents électroniques.

Les ambitions de l'être humain, de voir son besoin d'information rapidement satisfait, sont extraordinairement amplifiées avec l'apparition du premier ordinateur. L'idée imaginaire du Memex (1945) avait connu sa concrétisation révolutionnaire dans les moteurs de recherche sur le Web et dans les bibliothèques électroniques collaboratives.

Le décalage conceptuel, entre concordance de terme et cohérence du sens, duquel souffraient les SRI, semble se rétrécir avec la naissance de nouvelles habitudes des internautes. Une étude récente d'un groupe de chercheurs de trois universités (*Columbia*, *Wisconsin* et *Harvard*) a signalé une nouvelle tendance d'oublier ce qui peut être facilement retrouvé par un moteur de recherche sur Internet. Dénommé par l'*effet-Google*, le phénomène décrit dans cette étude proclame que : *Les gens peuvent se rappeler d'une information s'ils ne savent pas où la trouver ; ils peuvent se rappeler comment trouver leur besoin s'ils ne se rappellent pas de l'information* [Sparrow et al., 2011]. Ainsi, notre société[11] commence à développer une mémoire *Google*[12] basée sur le rappel des formulations des requêtes pour satisfaire un besoin spécifique.

Par ailleurs, si la recherche ad-hoc constitue la tâche principale en RI, la catégorisation des textes, le filtrage, l'agrégation sont d'autres formes utiles pour certaines situations. La recherche d'information moderne partage souvent avec les applications de fouille de données des techniques d'intelligence artificielle et de traitement automatique du langage naturel. Il est devenu évident de voir la RI moderne dans un contexte plus général de l'analyse automatique du texte. Nous décrirons dans le chapitre suivant les principaux modèles d'indexation des documents avant d'étudier de près une alternative utile pour l'indexation sémantique du texte avec un traitement linguistique adéquat.

[10] http://www.ldc.upenn.edu/Catalog/catalogEntry.jsp?catalogId=LDC2011T11

[11] Près d'un tiers des habitants du globe utilise Internet, http://www.internetworldstats.com/stats.htm

[12] Dont l'usage représente plus de 80% de la recherche mondiale sur le Web, http://marketshare.hitslink.com/

Chapitre 3 :
MODELES D'INDEXATION TEXTUELLE

1 Introduction

La tâche de représentation des documents constitue la plate-forme sur laquelle tout SRI puisse capturer le contenu documentaire avant de pouvoir l'indexer et de mesurer sa pertinence par rapport à une requête donnée. Un modèle de RI fournit le cadre théorique pour cette représentation qui peut conduire à des performances de recherche plus ou moins satisfaisantes. Ainsi, chaque modèle de représentation des documents définit une stratégie de recherche impliquant une méthode d'indexation et une fonction de calcul de pertinence.

L'idée de base, initiée par [Luhn, 1957] pour indexer le texte en calculant la fréquence de ses termes, est commune pour la plupart des modèles de représentation des documents. Néanmoins nous pouvons classer ces modèles d'une part selon leurs bases mathématiques : approche de la théorie des ensembles, approche algébrique ou approche probabiliste. D'une autre part, la distinction entre les différents modèles peut être établie sur la base de la prise en compte de l'interdépendance des termes d'indexation. Nous esquissons dans Figure 3-1 les principaux modèles de documents.

Nous décrivons dans ce chapitre les principaux modèles de RI. En particulier, nous nous intéressons au mode de représentation du texte pour capturer au mieux son contenu lexical et sémantique. Le formalisme de mesure de la pertinence ou la similarité des documents sera décrit afin de comprendre et apprécier la capacité du modèle à s'intégrer dans un système réel basé sur des techniques pratiques de mesure et d'apprentissage automatique. Les aspects relatifs au prétraitement linguistique seront différés au (Chapitre 5 :1).

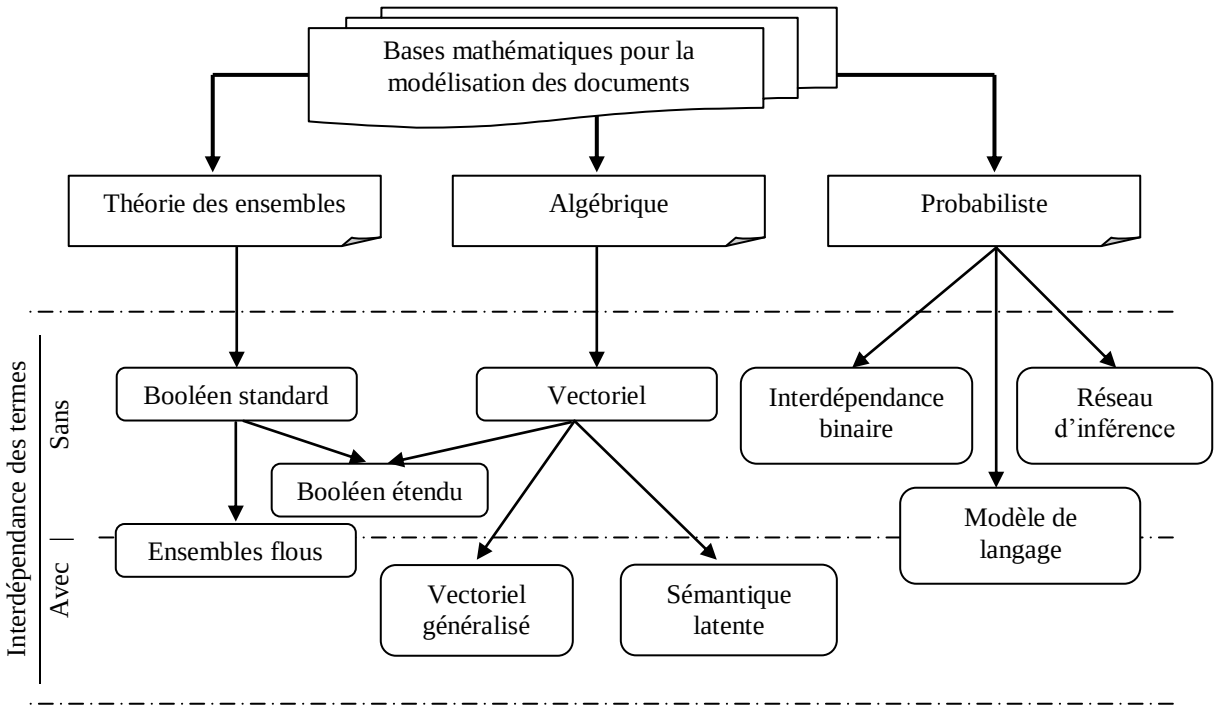


Figure 3-1. Classification des principaux modèles de document pour la RI.

2 Modèles booléens

L'approche Booléenne est la plus ancienne des stratégies de recherche et de modélisation des documents en RI. Partant d'un modèle ensembliste, l'approche Booléenne caractérise chaque document par l'appartenance (ou non) des termes au même document. L'ensemble de tous les termes utilisés (T) est appelé vocabulaire d'indexation. Ainsi, nous pourrons construire, pour une collection de documents (D), une matrice d'incidence binaire (D,T) où chaque élément (d_i,t_j) prend la valeur 1 si le document d_i contient au moins une fois le terme t_j. Autrement, la valeur 0 interprète l'absence dans le document du terme d_i du terme t_j.

Prenons, à titre d'exemple, les phrases suivantes :

D_1 : L'Algérie compte huit millions d'élèves scolarisés.

D_2 : Le vaccin antigrippal est disponible par millions aux Algériens.

D_3 : Les algériens ont payé leur indépendance par des millions de martyrs.

D_4 : L'équipe nationale est qualifiée à la coupe du monde.

D_5 : L'économie nationale dépend fortement des hydrocarbures.

En ne sélectionnant que les noms de ces phrases, nous construisons le vocabulaire d'indexation avec lequel la matrice d'incidence peut être établie comme dans Tableau 3-1.

	Algérie	**algériens**	**millions**	**antigrippal**	**coupe**	**économie**	**équipe**	**élèves**	**hydrocarbures**	**indépendance**	**martyrs**	**monde**	**nationale**	**scolarisés**	**vaccin**	**…**
D_1	1	0	1	0	0	0	0	1	0	0	0	0	0	1	0	
D_2	0	1	1	1	0	0	0	0	0	0	0	0	0	0	1	
D_3	0	1	1	0	0	0	0	0	0	1	1	0	0	0	0	
D_4	0	0	0	0	1	0	1	0	0	0	0	1	1	0	0	
D_5	0	0	0	0	0	1	0	0	1	0	0	0	1	0	0	

Tableau 3-1. Matrice d'incidence (document-terme) selon le modèle booléen.

2.1 Recherche booléenne

Dans un modèle de recherche booléenne, la fonction de pertinence $RSV(D_i,Q)$ pour chaque document (D_i) de la collection, par rapport à une requête ($Q=\{q_k\}$) composée des termes $\{q_k\}$, est donnée par[13] :

$RSV(D_i, q_k) = 1$ si $q_k \in D_i$; 0 sinon

$RSV(D_i, \neg q_k) = \neg RSV(D_i, q_k)$

$RSV(D_i, q_1 \vee q_2) = RSV(D_i, q_1) \vee RSV(D_i, q_2)$

$RSV(D_i, q_1 \wedge q_2) = RSV(D_i, q_1) \wedge RSV(D_i, q_2)$

A titre d'exemple, pour chercher les documents contenant le terme *millions* mais ne cite pas les *martyrs*, il suffit de formuler la requête :

Q_1 = "*millions* ET NON *martyrs*"

En calculant pour chaque document la fonction de pertinence relative à la requête (Q_1), nous obtenons, dans un système de recherche idéal, les valeurs de pertinence suivantes par rapport à la requête :

$RSV(D_1, Q_1) = 1$; $RSV(D_2, Q_1) = 1$; $RSV(D_3, Q_1) = 0$; $RSV(D_4, Q_1) = 0$; $RSV(D_5, Q_1) = 0$;

Ce qui permet au système de recherche de répondre à la requête (Q_1) par les deux documents D_1 et D_2.

En se basant sur un calcul binaire, le modèle booléen paraît assez simple et puissant ; il permet une recherche restrictive qui est très utile pour un utilisateur expérimenté désirant obtenir une information exacte et spécifique. Cependant, le besoin en information peut ne pas être correctement formulé dans la requête. L'exactitude de recherche pour une telle situation pourrait être pénalisante. Ainsi, nous pouvons résumer les insuffisances suivantes dans ce modèle :

- absence d'une recherche approximative,
- pondération équitable des termes,
- impossibilité d'établir un ordre de pertinence sur les résultats.

[13] Dans le sens d'un calcul en logique classique

Nous verrons par la suite d'autres alternatives et extensions du modèle booléen afin de mieux représenter et rechercher les documents d'une collection.

2.2 Index inversé

Il est intéressant à ce stade de voir une autre manière pour sélectionner les documents pertinents qui consiste à exploiter la matrice d'incidence (document-terme) en transposée. Chaque ligne représente un terme décrit par sa présence, ou non, dans les documents de la collection. Ainsi, l'évaluation d'une requête se limite au calcul des seuls vecteurs-termes composant la requête. Cette technique, connue par la représentation en index inversé, évite de passer sur des milliers de documents pour évaluer la fonction *RSV* correspondante à la requête utilisateur.

L'index inversé, ou fichier inversé, doit être préalablement conçu afin d'accélérer le calcul de la fonction de pertinence et, par conséquent, d'améliorer la performance du système de recherche. Dans le modèle de recherche booléenne, chaque requête sera donc formulée par une combinaison de termes et de connecteurs booléens {NON($\neg$), OU($\vee$), ET($\wedge$)}.

En revenant sur l'exemple précédent, cette technique propose de prendre les vecteurs relatifs aux termes de la requête {*millions, martyrs*}, et appliquer un calcul binaire comme suit :

$11100 \wedge \neg(00100) = 11100 \wedge 11011 = 11000$

La recherche donnera donc la même réponse constituée des deux premiers documents {D1, D2}.

2.3 Modèle à base des ensembles flous

Introduite par Lotfi Zadeh, la théorie des ensembles flous est une extension de la théorie classique des ensembles où chaque élément possède un degré (de certitude) d'appartenance à cet ensemble [Zadeh, 1965]. Cette théorie, qui a servi comme fondement théorique pour la logique floue, a inspiré les chercheurs en RI pour prendre en charge les notions de vague et d'imprécision dans le processus d'indexation [Bordogna et Pasi, 2001].

Par extension du modèle booléen de base, un document est modélisé par un ensemble flou des termes $D_i = \{(t_j, p_j)\}$; où p_j est le degré d'appartenance, déterminé par l'importance, du terme t_j au document D_i. Selon Bordobna et Pasi, l'importance du terme dans un document peut être déterminée de façon graduelle selon son appartenance à une section spécifique dans un texte structuré (titre, mots clés, titre de section, corps du texte) [Bordogna et Pasi, 2001].

Du moment où la majorité des utilisateurs ne cherchent pas à pondérer les termes lors d'une recherche, la représentation des requêtes peut garder le même principe du modèle booléen de base. Une des formulations communes de la fonction de pertinence entre un document ($D_i=\{(t_j , p_j)\}$) et une requête ($Q=\{q_k\}$) est définie par :

$RSV(D_i, t_j) = p_j$

$RSV(D_i, \neg q_k) = 1\text{-}RSV(D_i, q_k)$

$RSV(D_i, q_1 \vee q_2) = \max(RSV(D_i, q_1), RSV(D_i, q_2))$

$RSV(D_i, q_1 \wedge q_2) = \min(RSV(D_i, q_1), RSV(D_i, q_2))$

Il faut signaler que cette évaluation proposée par L. Zadeh peut paraître inconvenable au contexte de la RI. L'exemple trivial dans la logique classique $RSV(D_i, t_j \wedge \neg t_j) = 0$ n'est vrai que lorsque la certitude est complète pour l'appartenance ou non du terme , q_1 au document D_i. Par ailleurs, remarquons que dans cette formulation, la conjonction et la disjonction sont

interprétées par les fonctions numériques *min* et *max*. La pertinence d'un document par rapport à une requête, composée deux termes, ne dépend que du degré de pertinence d'un seul. Le deuxième étant ignoré par la fonction *min* ou *max*. Contrairement à celle de L. Zadeh [Zadeh, 1965], la formulation de Lukasiewicz implique les deux parties en même temps dans l'évaluation de la mesure de pertinence :

$$RSV(D_i, q_1 \vee q_2) = RSV(D_i, q_1) * RSV(D_i, q_2)$$

$$RSV(D_i, q_1 \wedge q_2) = RSV(D_i, q_1) + RSV(D_i, q_2) - RSV(D_i, q_1) * RSV(D_i, q_2)$$

Certains travaux ont proposé de généraliser la représentation floue pour les requêtes en attribuant un poids d'importance pour le terme recherché. Trois approches peuvent être considérées [Bordogna et al., 1991] [Kraft et al., 1995] [Crestani et Pasi, 1999] :

- Pondération numérique des termes interprétant le degré de signification idéal du terme.
- Pondération linguistique des termes selon une graduation d'importance du terme dans la requête
- Pondération des opérateurs logiques pour exprimer l'importance désirée d'un terme par rapport à un autre.

Comparée au modèle booléen, cette extension basée sur les ensembles flous présente les avantages suivants :

- Prendre en charge l'imprécision qui caractérise le processus de recherche.
- Contrôler l'imprécision dans la requête de l'utilisateur.
- Possibilité d'attribuer un degré de correspondance entre un document et une requête.

L'inconvénient majeur de ces modèles réside dans l'inadaptation au classement des documents sélectionnés. La formulation initiale de L. Zadeh pour la fonction de calcul de pertinence *RSV* ne considère pas l'importance graduelle de tous les termes de la requête. Cependant, il existe d'autres formulations avec des extensions plus récentes afin d'améliorer l'ordonnancement des résultats de recherche [Boughanem et al., 2005].

2.4 Modèle booléen étendu

En plus des inconvénients relevés précédemment dans le modèle booléen de base, il est intéressant de noter que :

- Pour une requête sous forme d'une longue conjonction, un document qui satisfait la majorité des termes est aussi mauvais qu'un document qui ne satisfait aucun terme.
- Pour une requête sous forme d'une disjonction, un document qui satisfait un terme est aussi bon qu'un document qui satisfait tous les termes.

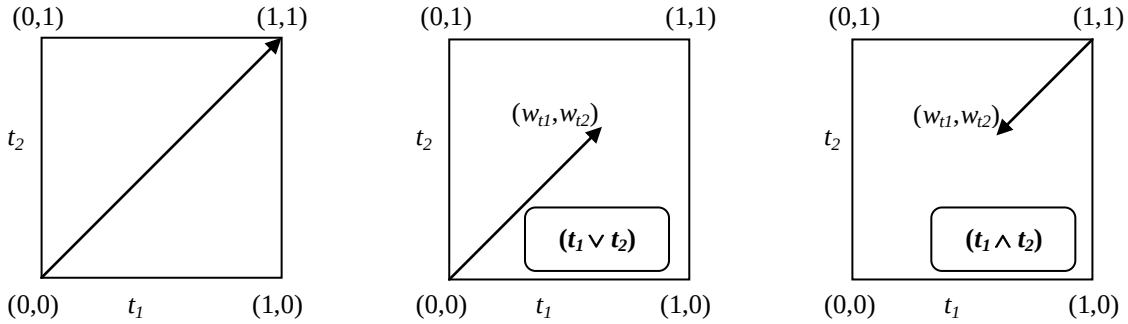
Afin de surmonter ces limitations, le modèle booléen étendu a été proposé [Salton et al., 1983]. Il se base sur les deux modèles booléen et vectoriel avec des requêtes P-normées.

Les caractéristiques principales du modèle booléen étendu (ou modèle P-Norme) peuvent être résumées comme suit [Salton et al., 1983]:

- Considérer tous les termes dans une requête.
- Ajuster la restriction des opérateurs ($\vee$,$\wedge$) par des P-valeurs.

- Proposer un modèle général dont celui du booléen de base n'est qu'un cas particulier, avec une interprétation des ensembles floues ($p\to\infty$) et un modèle d'espace vecteur muni du produit scalaire comme similarité (p=1)

Le modèle propose de pondérer les termes par leur fréquence et de les normaliser dans l'intervalle [0,1]. La pondération la mieux adaptée et celle du *tf.idf* qui sera décrite dans la section suivante. Considérons, par exemple, une requête composée de deux termes k_1 et k_2. un document de la collection possède les pondérations (…, w_{k1},w_{k2}, …) relatives à ces deux termes (voir Figure 3-2). La représentation dans l'espace bidimensionnel des termes présente deux cas extrêmes : le premier (1,1) lorsque les documents sont totalement pertinents ; et le deuxième (0,0) lorsque les documents ne contiennent aucun des deux termes. La distance euclidienne maximale entre ces deux points est donc $2^{½}$.



(a) distance euclidienne max. **(b)** distance euclidienne du OU. **(c)** distance euclidienne du ET.

Figure 3-2. Schématisation des opérateurs logiques dans un espace bidimensionnel.

D'après la représentation géométrique dans Figure 3-2, il en résulte la distance euclidienne suivante du (*OU*) entre la requête ($k_1 \vee k_2$) et le document (…, w_{k1},w_{k2}, …) :

$$dist_{OU} = \sqrt{(w_{k1}-0)^2 + (w_{k2}-0)^2} = \sqrt{w_{k1}^{\ 2} + w_{k2}^{\ 2}}$$

De même pour la distance euclidienne du (ET) :

$$dist_{ET} = \sqrt{2} - \sqrt{(1-w_{k1})^2 + (1-w_{k2})^2}$$

Ces deux équations doivent être normalisées pour exprimer un score de similarité en divisant par $2^{½}$. Ensuite, on peut généraliser pour n'importe quel nombre de termes (m) et pour un paramètre p pour définir une similarité en P-Norme :

$$Sim(D,Q_{OU}) = \left(\frac{\sum_{j=1}^{m} w_{kj}^{p}}{m} \right)^{1/P}$$

$$Sim(D,Q_{ET}) = 1 - \left(\frac{\sum_{j=1}^{m} (1-w_{kj})^{P}}{m} \right)^{1/P}$$

Des propriétés intéressantes peuvent être constatées sur le modèle P-Norme. Pour p=1, c'est le modèle vectoriel qui est exprimé avec l'unification des opérateurs logiques (*OU*, *ET*). Cependant lorsque $p\to\infty$, on obtient le modèle booléen standard ou basé sur les ensembles flous [Salton et al., 1983].

3 Modèles vectoriels

Le modèle vectoriel (ou d'espace vecteur) est l'un des plus anciens et des mieux connus parmi les modèles classiques de la RI [Salton et al., 1975]. Durant trois décennies depuis son apparition vers la fin des années 1960, plusieurs études et variantes avaient été investi par les chercheurs dans le domaine [Salton, 1968] [Salton et McGill, 1983] [Baeza-Yates et Ribeiro-Neto, 1999].

Les documents est les requêtes sont représentés par des vecteurs de poids des termes descripteurs. Chaque poids interprète l'importance du terme correspondant dans le texte. Les vecteurs sont décrits dans un espace vectoriel définis par l'ensemble des termes préalablement établi lors de l'indexation. Le nombre (N) des termes d'indexation (ou du dictionnaire) détermine la dimension de l'espace vectoriel R^N.

Le modèle vectoriel, qui représente le noyau du système SMART pour l'analyse et la recherche automatique des textes [Salton, 1971], repose sur les hypothèses suivantes :

- Document et requête possèdent tous les deux une représentation uniformisée dans le même espace des termes.
- Aucune dépendance entre les termes n'est considérée.
- Les documents pertinents sont ceux qui ont des représentations vectorielles les plus proches de celle de la requête.
- La fonction de pertinence *RSV* (similarité entre un document et une requête) est obtenue par la définition d'une distance dans l'espace vectoriel.

Il est intéressant de noter dans ce contexte que les termes utilisé pour définir l'espace de représentation d'un corpus ne sont pas tous présents dans le même document. Par conséquent, les vecteurs-documents résultants sont creux. Ceci permettra d'utiliser des techniques adaptées pour accélérer le calcul des similarités.

3.1 Fonctions de similarité

Il est donc possible, dans le modèle vectoriel, d'ordonner les résultats sélectionnés par ordre de décroissant de correspondance entre la requête et chaque document. Pour utiliser une mesure de similarité dans l'espace R^N, notons ce qui suit :

- N : la dimension du dictionnaire considéré pour la représentation des textes de la collection (nombre de termes uniques)
- $D_i = (d_{i1}, d_{i2}, \ldots d_{iN})$: vecteur représentant du i-ème document de la collection.
- $Q = (q_1, q_2, \ldots q_N)$: vecteur représentant d'une requête.
- d_{ij}, q_j : poids du j-ème terme du document (resp. de la requête)

Ainsi, nous pouvons définir les principales mesures de similarité :

Le produit scalaire : $$RSV(D_i, Q) = \sum_{j=1}^{N} d_{ij}.q_j$$

La mesure cosinus : $$RSV(D_i, Q) = \frac{\sum_{j=1}^{N} d_{ij}.q_j}{\sqrt{\sum_{j=1}^{N} d_{ij}^2}.\sqrt{\sum_{j=1}^{N} q_j^2}}$$

La mesure de Jaccard : $$RSV(D_i, Q) = \frac{\sum_{j=1}^{N} d_{ij}.q_j}{\sum_{j=1}^{N} d_{ij}^2 + \sum_{j=1}^{N} q_j^2 - \sum_{j=1}^{N} d_{ij}.q_j}$$

La fonction *RSV* est assimilée à toute mesure de similarité pouvant être définie dans un espace vectoriel. Dans le cas d'une distance, son calcul est inversement proportionnel à la fonction de pertinence. Une des plus élémentaires est la distance euclidienne qui peut être exploitée de la façon suivante :

$$RSV(D_i, Q) = \frac{1}{\sqrt{\sum_{j=1}^{N} (d_{ij} - q_j)^2}}$$

Ces mesures de similarité permettent de classer les documents d'une collection par ordre décroissant de pertinence. Un système de recherche sélectionne les documents les plus similaires à la requête soit selon un nombre prédéfini, soit par satisfaction d'un seuil préalablement fixé pour la valeur *RSV*.

3.2 Pondération des termes

L'approche la plus simple pour attribuer un poids à un terme *t* descripteur d'un document *d* consiste dans le nombre de son appariation dans le texte lui-même. Cette pondération est plus connue par la fréquence du terme ou le schéma *tf* (*en angl. Term frequency*). Chaque document *d* est caractérisé par un ensemble de poids relatifs aux termes du dictionnaire. Dans ce schéma, aussi connu par le modèle de sac de mots (*en angl. Bag of words*), l'ordre des termes dans le document est ignoré sans aucune considération syntaxique.

Néanmoins, la pondération *tf* est très sensible à la longueur des documents et peut perturber l'évaluation de similarité entre les documents et les requêtes. La mesure *idf* (*en angl. Inverted document frequency*) détermine si un terme *t* est discriminant sur toute la collection de documents. Le schéma de pondération *tf.idf* constitue une alternative plus stable en faisant intervenir à la fois :

- L'importance du terme pour le document (*tf*)
- Le pouvoir de discrimination de ce terme dans la collection (*idf*)

Ces deux mesures sont formulées comme suit :

- $tf_{t,d}$: fréquence du terme *t* dans le document *d*.
- $idf_t = log(M/df_t)$; où *M* est le nombre de documents de la collection (taille du corpus) et df_t est le nombre de documents contenant le terme *t*.
- $tf.idf = tf_{t,d} * log(M/df_t)$
- $tf.idf = (tf_{t,d} / max(tf_{t,d})) * log(M/df_t)$: mesure normalisée en divisant par le maximum des fréquences dans un document.

3.3 Loi de Zipf

La loi de Zipf est le résultat d'une étude statistique qui met en évidence un rapport entre la fréquence d'un terme et son rang dans le vocabulaire utilisé [Zipf, 1949]. Si l'on compte le nombre de fois qu'un terme apparaît dans une grande collection, et que l'on met les termes par ordre de fréquence, il est possible d'examiner le lien entre la fréquence globale (tf_t) d'un terme et sa position dans la liste (son rang r_t).

Zipf énonce que la fréquence d'un terme est proportionnelle à l'inverse de son rang ; autrement dit, il existe une constante *K* telle que :

$$tf_t.r_t = K \quad \textbf{(3-1)}$$

Si on trace une courbe du rang en fonction de la fréquence (et qu'on utilise des axes logarithmiques), on obtient essentiellement une droite de pente -1. Nous présentons dans Tableau 3-2 la distribution des premiers dix mots les plus fréquents dans la collection Brown qui a été collectée en 1961 et regroupant plus d'un million de mots [Manning et Schütze, 2001].

Mot	Rang	Fréquence	%
the	1	69975	6,90%
be	2	39175	3,86%
of	3	36432	3,59%
and	4	28872	2,85%
to	5	26190	2,58%
a	6	23073	2,28%
in	7	20870	2,06%
he	8	19427	1,92%
have	9	12458	1,23%
it	10	10942	1,08%

Tableau 3-2. Distribution des 10 mots les plus fréquents dans la collection Brown.

Bien qu'elle reste empirique et parfois contestée, la loi de Zipf nous renseigne qu'il existe quelques mots communs et représentent la plus grande partie des textes. Généralement, ces mots ne sont pas discriminants et peuvent être considérés comme mots vides (*en angl. stopwords*). Leur suppression peut réduire l'index de 30 à 50% de sa taille.

3.4 Formule de Rocchio

La rétroaction de pertinence est une technique très répandue pour la reformulation de requêtes. Dans le modèle vectoriel, la formule de Rocchio permet une transformation automatique d'une requête initiale (*Q*) [Rocchio, 1971]. L'idée consiste à trouver un nouveau vecteur de requête (*Q'*) qui permet de maximiser la pertinence. Sachant que l'ensemble D_P représente les documents jugés pertinents et D_{np} les documents non pertinents. La formule calcule la nouvelle requête (*Q'*) en rajoutant à (*Q*) les vecteurs des documents pertinents et en retranchant ceux qui ne le sont pas.

$$\vec{Q}' = \alpha\vec{Q} + \beta\frac{1}{|D_p|}\sum_{\vec{d}\in D_p}\vec{d} - \gamma\frac{1}{|D_{np}|}\sum_{\vec{d}\in D_{np}}\vec{d}$$

Les paramètres (α, β, γ) sont choisi en fonction de l'importance que l'on souhaite donner à chaque terme. Cette reformulation est automatique du moment où seul le jugement initial du système est considéré. Elle peut être semi-automatique ou interactive en prenant en compte l'appréciation de l'utilisateur.

La formule et l'algorithme de Rocchio, sont introduits dans la version initiale du système SMART qui a été amélioré et muni d'autres techniques de rétroaction de pertinence [Salton, 1971] [Salton, 1983] [Salton et Buckley, 1990]. Buckley et Salton réclamaient que les requêtes optimisées, par amélioration dynamique des poids à base de la formule de Rocchio,

augmente de 10 à 15% les performances d'un système de recherche [Buckley et Salton, 1995].

3.5 Modèle vectoriel généralisé

Le modèle d'espace vecteur de base suppose que les vecteurs associés aux termes descripteurs (dans R^N) sont linéairement indépendants. Cette hypothèse d'orthogonalité ne permet pas d'exprimer les relations sémantiques qui peuvent exister entre les termes. Pour surmonter cette insuffisance, le modèle vectoriel généralisé propose une extension en considérant un nouvel espace vectoriel où chaque vecteur-terme t_i est exprimé en fonction d'une combinaison linéaire de $N'=2^N$ vecteurs [Wong et al., 1985] [Wong et al., 1987].

La similarité entre un document D et une requête Q peut être exprimée par la mesure du cosinus comme suit :

$$RSV_{\cos}(D,Q) = \frac{\sum_{j=1}^{N'}\sum_{i=1}^{N'} w_{D,i}.w_{Q,j}.\vec{t}_i.\vec{t}_j}{\sqrt{\sum_{i=1}^{N'} w_{D,i}^2}.\sqrt{\sum_{j=1}^{N'} w_{Q,i}^2}}$$

Où t_i, t_j sont des vecteurs-termes dans l'espace à $N'=2^N$ dimensions. $w_{D,,i}$, $w_{Q,,i}$ étant, respectivement, les nouveaux poids des vecteurs, document (D) et requête (Q), dans le nouvel espace. Le calcul des vecteurs (t_i, t_j) dans la formule n'est pas nécessaire tant qu'il n'y a pas de corrélation entre les termes associé. Dans le cas particulier où il n'existe aucune corrélation sémantique entre les termes (orthogonalité de tous les termes deux à deux), la formule précédente se réduit à la mesure de similarité du cosinus dans le modèle vectoriel de base.

Cette approche constitue une bonne alternative aux modèles d'espace vecteur pour prendre en charge la sémantique du texte. Deux directions de recherche peuvent être notées dans ce contexte [Tsatsaronis et Panagiotopoulou, 2009] :

- Calculer les corrélations sémantiques entre les termes,
- Calculer des statistiques de cooccurrences des termes dans les grandes collections.

Le cadre global, qu'offre le modèle vectoriel généralisé en utilisant l'espace des cooccurrences des termes (au lieu des termes), a suscité plusieurs travaux de recherche pour incorporer les la sémantique dans l'indexation des textes. La construction de la matrice des cooccurrences terme×terme peut exploiter les relations taxonomiques (synonymie, hyperonymie et hyponymie) extraites à partir de ressources linguistiques externes (telles que WordNet et Wikipedia) [Wang et Domeniconi, 2008] [Tsatsaronis et Panagiotopoulou, 2009].

3.6 Indexation sémantique latente

Une des plus importantes extensions du modèle d'espace vecteur est celle basée sur la méthode d'analyse sémantique latente (LSA) [Deerwester et al., 1990]. Cette méthode utilise un procédé, de décomposition en valeur singulière (SVD), très connu de l'algèbre linéaire pour la factorisation des matrices rectangulaires. Le modèle d'indexation en sémantique latente (LSI) s'inspire de la LSA pour réduire la représentation d'une collection de textes étant initialement sous forme d'une matrice (A) terme×document [Dumais, 1993].

Notre matrice $A_{N\times M}$, pour M documents et N termes, peut être décomposée sous la forme :

$$A = U\,S\,V^T$$

Avec $U_{N\times M}$ une matrice orthogonale ($UU^T=I_N$) pour la description des termes ; et $V_{M\times M}$ une matrice orthogonale ($VV^T=I_M$) pour la description des documents. $S_{N\times N}$ est la matrice

diagonale qui contient les valeurs singulières de *A*. Une convention courante consiste à ranger par ordre décroissant les valeurs de *S* afin de la décrire de façon unique.

On peut constater qu'à partir d'un certain nombre $k<N$, les valeurs singulières deviennent insignifiantes. Il est possible donc de trouver une approximation A_k de A en sélectionnant seulement les k premières valeurs et en réduisant la matrice A aux k termes les plus significatifs :

$$A_k = U_k S_k V_k^T$$

A_k représente l'approximation d'ordre k de la matrice A. La réduction de dimensionnalité obtenue par cette technique de SVD permet, dans un sens, de supprimer le bruit de dimension qui cachait la structure latente.

Pour calculer la similarité du cosinus entre chaque document du corpus (ligne de V_k) et le vecteur de requête Q_k obtenu par projection du vecteur initial Q dans le nouvel espace selon la formule suivante :

$$Q_k = Q^T U_k S_k^{-1}$$

Le paramètre k peut être choisi empiriquement selon la collection utilisé et la performance désiré. L'objectif d'une indexation en sémantique latente est d'obtenir la meilleure réduction de dimension avec le minimum de perte d'information utile pour la description du problème.

3.6.1 Exemple

En revenant sur l'exemple dans la section 2 de la page 46, et en utilisant le codage *tf* pour la pondération des termes des (M=5) documents dans la matrice A (terme×document). Nous proposons d'appliquer une décomposition en valeurs singulières et réduire la représentation des textes de (N=15) à la dimension k=3 (voir Figure 3-3). Les résultats sont obtenus par les fonctions prédéfinies dans Matlab.

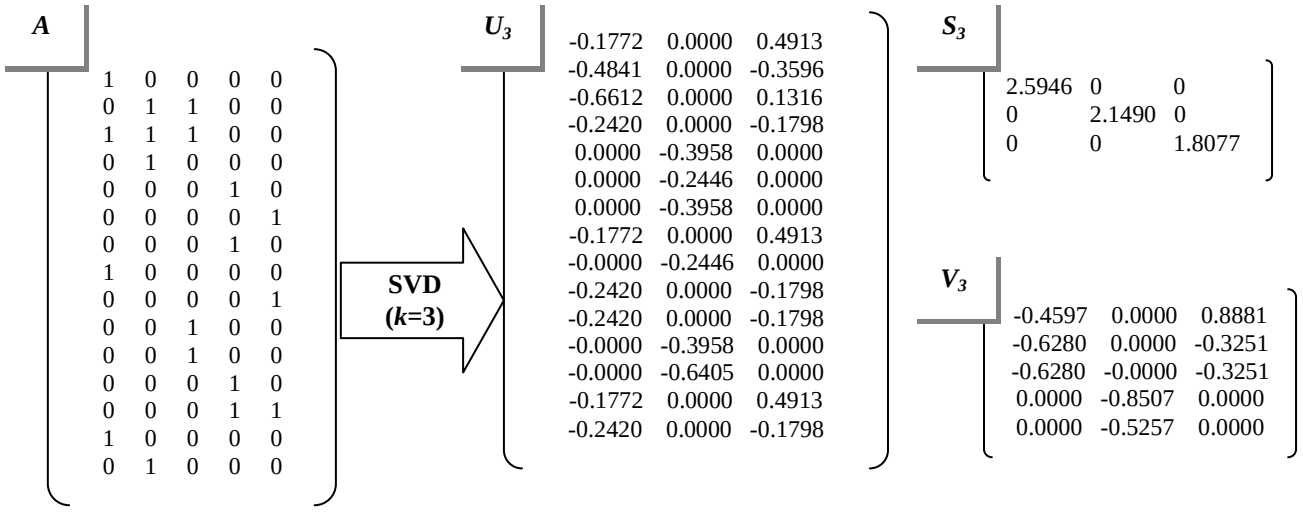


Figure 3-3. Décomposition en valeurs singulières avec réduction de l'espace (k=3) .

Nous posons la requête "*millions martyrs*" qui sera codée dans l'espace vectoriel initial par Q = (0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0)

$$Q_3 = Q^T U_3 S_3^{-1} = (-0.3481, 0, -0.0267)$$

Et en calculant la similarité de cosinus de cette requête avec les cinq documents tel qu'ils sont représentés dans l'espace réduit par V_3 ; nous obtenons les valeurs de pertinence suivantes :

$RSV_{LSI3}(D_1, Q) = 0{,}39042513$; $RSV_{LSI3}(D_2, Q) = 0{,}92061828$;

$RSV_{LSI3}(D_3, Q) = 0{,}92061828$; $RSV_{LSI3}(D_4, Q) = 0$; $RSV_{LSI3}(D_5, Q) = 0$;

Ce qui implique l'ordre de pertinence suivant : D_2, D_3, D_1, D_4, D_5.

3.6.2 Remarques

L'application du modèle LSI, sur de grandes collections, peut donner un aperçu sur la cognition humaine. Les problèmes de synonymie (même sens pour deux termes) et de polysémie (même terme pour deux sens) peuvent être pris en charge par la technique de réduction contrôlée des valeurs singulières. Trois idées principales caractérisent l'approche d'analyse en sémantique latente [Dumais, 1993] [Steyvers et Griffiths, 2007] :

- Que la sémantique peut être dérivée de la matrice de cooccurrences terme×document,
- Que la réduction de dimension est une partie essentielle de cette dérivation,
- Que les mots et les documents peuvent être représentés dans un espace Euclidien.

Néanmoins, le modèle LSI manque de fondement théorique justifiant son usage pour la modélisation des textes en RI. Nous verrons dans le chapitre suivant une approche (pLSI) qui partage avec LSI les deux premières caractéristique ci-dessus, mais qui diffèrent dans la troisième.

4 Modèles probabilistes

Cette famille de modèles est basée sur le principe général indiquant que les documents d'une collection doivent être ordonnés par probabilité descendante de leur pertinence à une requête. Le principe (*RPR*) d'ordonnancement probabiliste (*en angl. The probabilistic ranking principle*), fût annoncé depuis les années 1960 avant d'être revitalisé à la fin des années 1970.

Van Rijsbergen l'énonça comme suit : Si la réponse d'un système de recherche de référence pour chaque demande est un classement des documents de la collection dans un ordre décroissant de probabilité de pertinence pour l'utilisateur qui a soumis la demande, où les probabilités sont estimées aussi précisément que possible sur la base de toutes les données disponibles au système à cet effet, l'efficacité globale du système pour son utilisateur sera la meilleure qu'on puisse obtenir sur la base de ces données [Maron, 1964] [Robertson, 1977] [Rijsbergen, 1979].

Dans les modèles probabilistes, il n'est plus question de calculer pour chaque document un score (absolu) de pertinence mais d'ordonner (relativement) les documents d'une collection par probabilité de pertinence. L'idée commune est d'estimer la probabilité de pertinence d'un document (D) de la collection par rapport à une requête donnée (Q) :

$P(P|D,Q)$: dénote la probabilité de pertinence du document D à la requête Q.

$P(NP|D,Q)$: dénote la probabilité de non-pertinence du document D à la requête Q.

La fonction de correspondance *RSV* est assimilée à l'ordonnancement de probabilité de pertinence en mesurant la chance logarithmique (log-odds) de pertinence d'un document :

$$RSV(D,Q) \equiv \log\left(O(P|D,Q)\right) = \log\left(\frac{P(P|D,Q)}{P(NP|D,Q)}\right) \quad \textbf{(3-2)}$$

Où on peut appliquer la règle de Bayes pour inverser la probabilité conditionnelle :

$$O(P|D,Q) = \frac{P(P|D,Q)}{P(NP|D,Q)} = \frac{\frac{P(P|Q)P(D|P,Q)}{P(D|Q)}}{\frac{P(NP|Q)P(D|NP,Q)}{P(D|Q)}} = \frac{P(P|Q)}{P(NP|Q)}.\frac{P(D|P,Q)}{P(D|NP,Q)} = O(P|Q).\frac{P(D|P,Q)}{P(D|NP,Q)}$$

Et sachant que la chance de pertinence d'une requête donnée est constante on peut formuler la mesure *RSV* en fonction de la pertinence du document *D* seulement :

$$RSV(D,Q) \equiv \log\left(\frac{P(D|P,Q)}{P(D|NP,Q)}\right) \quad \textbf{3-3)}$$

A partir des hypothèses fixées pour estimer cette quantité dans (3-3, plusieurs modèles de recherche probabiliste peuvent être considérés. Cependant, l'hypothèse commune dans la plupart des modèles probabilistes énonce que la pertinence d'un document est indépendante des autres documents de la collection [Robertson, 1977].

4.1 Modèle d'indépendance binaire

C'est l'une des plus simples formes des modèles probabilistes où on suppose, d'une part, que les termes sont mutuellement indépendants. Et d'une autre part, qu'ils sont binairement représentés, comme dans le modèle booléen, pour interpréter leur présence (ou non) dans un document [Robertson et Sparck, 1976]. En dénotant le document considéré par $D=(t_1,\ldots,t_N)$ où $t_i \in \{0,1\}$ et en séparant les termes présents des absents dans *D*, nous exprimons la formule précédente par [Robertson et Sparck, 1976] [Rijsbergen, 1979] [Manning et al., 2008] :

$$\log\left(\frac{P(D|P,Q)}{P(D|NP,Q)}\right) = \log\left(\prod_{t_i \in Q,D} \frac{P(t_i=1|P,Q)}{P(t_i=1|NP,Q)}.\prod_{t_i \in Q, t_i \notin D} \frac{P(t_i=0|P,Q)}{P(t_i=0|NP,Q)}\right)$$

Le problème se réduit donc à l'estimation des probabilités $p_i=P(t_i=1|P,Q)$ et $q_i=P(t_i=1|NP,Q)$. Et en multipliant par $\prod_{t_j \in Q}\frac{(1-q_i)}{(1-p_i)}$, on peut réécrire l'équation précédente comme suit :

$$\log\left(\frac{P(D|P,Q)}{P(D|NP,Q)}\right) = \log\left(\prod_{t_j \in Q,D} \frac{p_i}{q_i}.\prod_{t_j \in Q, t_j \notin D} \frac{1-p_i}{1-q_i}\right) = \log\left(\prod_{t_j \in Q,D} \frac{p_i(1-q_i)}{q_i(1-p_i)}.\prod_{t_j \in Q} \frac{(1-p_i)}{(1-q_i)}\right)$$

Qui peut être réduite par la suppression du membre constant à droite pour avoir une fonction d'ordonnancement de pertinence simplifiée :

$$RSV(D,Q) \equiv \log\left(\prod_{t_j \in Q,D} \frac{p_i(1-q_i)}{q_i(1-p_i)}\right) = \sum_{t_j \in Q,D} \log \frac{p_i(1-q_i)}{q_i(1-p_i)} \quad \textbf{(3-4)}$$

Il est possible de dresser la table (Tableau 3-3) de distribution des documents dans la collection selon leur pertinence à une requête donné et la contenance d'un terme (t_i) :

	Document pertinent	**Document non-pertinent**	**Total**
Terme présent (t_i =1)	r_i	df_i - r	df_i
Terme absent (t_i =0)	R - r_i	M - R - df_i + r_i	M - df_i
Total	R	M - R	M

Tableau 3-3. Table de contingence pour la distribution des documents dans la collection.

Où M est la taille de la collection et df_i est le nombre de documents contenant le terme t_i.

Ceci nous permettra d'estimer les deux quantités (p_i,q_i) à partir de l'échantillon (collection) de taille M :

$$p_i = r_i / R \; ; \; q_i = (df_i - r_i) / (M - R)$$

Il en résulte l'estimation du poids d'un terme t_i en substituant le compte de la table de contingence (Tableau 3-3) dans le terme utilisé dans l'équation de pertinence précédente :

$$w_i = K(M, df_i, R, r_i) = \frac{r_i /(R - r_i)}{(df_i - r_i)/(M - R - df_i + r_i)} \quad \textbf{(3-5)}$$

Ainsi le poids d'un document peut être déterminé par :

$$w(D) = \sum_{t_i} t_i . w_i = \sum_{t_i} t_i . \frac{r_i /(R - r_i)}{(df_i - r_i)/(M - R - df_i + r_i)} \quad \textbf{(3-6)}$$

4.2 Généralisation de la pondération des termes

Remarquons que si nous supposons que le taux des documents pertinents est très faible dans la collection considérée, nous pourrions approximer la probabilité du terme présent dans les documents non pertinents q_i par (df_i/M), et par conséquent le membre de l'équation (3-4), relatif aux documents non pertinents, peut être simplifié comme suit :

$$\log \frac{1 - q_i}{q_i} = \log \frac{M - df_i}{df_i} \approx \log \frac{M}{df_i} \quad \textbf{(3-7)}$$

Ce qui nous rappelle la fréquence inversé des documents *idf* dans le modèle vectoriel. Par ailleurs, les deux formulations précédentes (3-5) et (3-6) permettent de considérer des poids pour les termes d'indexation du document mais aussi, d'une certaine manière, de pondérer les termes des requêtes dans une éventuelle phase de reformulation (rétroaction de pertinence) [Robertson et Sparck, 1976] [Rijsbergen, 1979]. Ainsi, la formulation (3-4) devient similaire à la fonction de pertinence *RSV* dans le modèle d'espace vecteur avec une pondération unitaire des termes.

Une première généralisation du modèle probabiliste binaire proposa d'attribuer une pondération non binaire aux termes [Croft, 1981]. Une espérance mathématique peut être donc exprimée par :

$$E(D) = \sum_{i=1..N} P(\delta_i) . \log \frac{p_i (1 - q_i)}{q_i (1 - p_i)} \quad \textbf{(3-8)}$$

Où $P(\delta_i)$ représente la probabilité d'indexation du document D par le terme t_i.

4.3 Okapi BM25

Le modèle probabiliste d'indépendance binaire était initialement conçu pour des collections modérées avec de courts textes et surtout une dimension assez moyenne du dictionnaire d'indexation. Avec la généralisation des méthodes de recherche aux grandes collections et textes réels accessibles sur le Web, il était nécessaire de réviser ce modèle, solide mathématiquement, pour considérer la fréquence des termes et la longueur des documents.

Le modèle Okapi, ou le schéma de pondération BM25 communément appelé Okapi-BM25, est un modèle probabiliste basé sur le modèle d'indépendance binaire [Robertson et Sparck, 1976] avec des extensions tirées du modèle 2-Poisson proposé dans [Robertson et Walker, 1994] pour capturer la fréquence du terme. En gardant les mêmes notations précédentes avec L_D et L_{moy} représentent la longueur (en nombre de mots) du document D et la longueur moyenne des documents de la collection ; la fonction de similarité Okapi-BM25 entre un document D et un requête Q est exprimée par :

$$RSV_{BM25}(D,Q) = \sum_{t \in Q} \left(\log \frac{M - df_t}{df_t} \right) . \frac{(k_1 + 1).tf_{t,D}}{k_1 \left((1-b) + b.L_D / L_{moy} \right) + tf_{t,D}} . \frac{(k_3 + 1).tf_{t,Q}}{k_3 + tf_{t,Q}} \qquad \textbf{(3-9)}$$

Sachant que k_1, b et k_3 sont des paramètres d'ajustement de la fonction Okapi-BM25 selon la nature de la collection considérée. Robertson et Walker ont recommandé de choisir les valeurs (k_1=1.2, b=0.75, k_3=1000) qui ont été utilisées avec succès dans des études ultérieures [Robertson et Walker, 1999] [Bennett et al. 2008].

L'expression de la fonction Okapi-BM25 peut être lue comme une composition de trois termes : le premier donne la fréquence inversée des documents *idf*, tandis que le deuxième (et le troisième resp.) reflète le poids du terme dans le document (et dans la requête resp.).

4.4 Dépendance hiérarchique des termes

Le modèle probabiliste de base suppose que les termes d'indexation sont mutuellement indépendants. Cette hypothèse est loin d'être réelle comme dans les expressions à double mots (telles que Hong et Kong, operating et system) qui sont fortement dépendants. L'élimination de l'hypothèse d'indépendance, qui a permis de transformer la quantité $P(P|D,Q)$ en un simple produit de pertinence des termes, nous oblige à réviser toute la construction mathématique du modèle probabiliste expliqué dans la section précédente. Van Rijsbergen considère cette dépendance comme stochastique et non pas logique ; il propose un modèle simple de dépendance hiérarchique où chaque terme peut être directement dépendant à un seul terme [Rijsbergen, 1979]. Cette idée a été reformulée et appliquée avec succès sur une variété de collections d'apprentissage automatique [Friedman et Goldszmidt, 1996].

4.5 Réseaux Bayesiens

Les réseaux Bayésiens constituent une combinaison de l'approche probabiliste avec la théorie des graphes. L'idée centrale d'un réseau Bayésien réside dans l'utilisation d'un graphe acyclique orienté pour la représentation des dépendances probabilistes entre les variables (document, requête, terme). Des formulations efficaces pour l'apprentissage (représentation de structures) des réseaux Bayésiens et l'inférence (propagation dans la structure) ont été proposées depuis les années 1980 pour prendre en charge des problèmes complexes de grande taille [Howard et Matheson, 1981] [Pearl, 1988] [Jensen, 2001]. Deux types de modèles de cette approche ont été proposés pour la modélisation des textes dans le contexte de la RI : les

réseaux d'inférence [Turtle et Croft, 1990] ; et les réseaux de croyance [Ribeiro-Neto et Muntz, 1996] [Baeza-Yates et Ribeiro-Neto, 1999].

Dans le modèle des ***réseaux d'inférence***, les nœuds représentent des concepts, des groupes de mots ou des documents. Un requête est elle aussi représentée par un nœud particulier. Les arcs correspondent aux relations sémantiques entre les nœuds avec des probabilités de croyance. La recherche peut être donc considérée comme un processus de raisonnement incertain pour estimer la probabilité qu'un document satisfasse une requête [Turtle et Croft, 1990]. Des systèmes efficaces, basés sur ce modèle, ont été développés et même commercialisés ; on peut citer dans ce contexte le système GRANT [Cohen et Kjeldsen, 1987] et INQUERY [Turtle et Croft, 1990].

Dans les ***réseaux de croyance***, la sélection d'un document repose sur la similarité entre un document et une requête donnée. La définition préalable de l'espace d'échantillonnage permet de séparer les portions de documents des portions de requêtes et, par conséquent, de calculer d'une manière efficace les degrés de croyance [Ribeiro-Neto et Muntz, 1996]. Le processus de recherche (propagation dans le réseau) est déclenché par la réception d'un besoin utilisateur et l'instanciation d'une requête. Parmi les apports majeurs du modèle des réseaux de croyance réside dans son cadre théorique solide qui permet de généraliser les modèles classiques de la RI tels que les modèles booléens, vectoriels, probabilistes [Ribeiro-Neto et Muntz, 1996] [Baeza-Yates et Ribeiro-Neto, 1999].

Sans pour autant rentrer dans les détails théoriques du modèle on peut dire que l'avantage principal, apporté par l'approche des réseaux Bayésien en RI, est la possibilité de combiner des informations provenant de différentes sources pour dresser un schéma de pertinence d'une requête dans une collection de documents. De plus, des techniques efficaces ont été proposées pour adapter le modèle des réseaux Bayésiens aux données hétérogènes mais aussi pour optimiser l'espace de représentation et accélérer les calculs [Indrawan et al., 1998] [Acid et al., 2003] [Crestani et al., 2003].

4.6 Modèles de langages

Une astuce commune aux utilisateurs, qui veulent formuler leur besoin d'information dans une bonne requête, consiste à penser aux mots les plus probables d'apparaître dans un document pertinent. L'approche des modèles de langage part de cette idée en supposant qu'un document correspond au mieux à une requête s'il est le plus susceptible de la générer.

Initiés vers la fin des années 1990, les modèles de langage reposent sur une approche statistique pour représenter les documents et les requêtes en intègrant les processus d'indexation et de recherche dans un seul modèle [Ponte et Croft, 1998]. Les modèles de langage permettent de générer directement, à partir des documents de la collection, les termes de la requête. Plusieurs variantes ont été proposées [Ponte et Croft, 1998] [Song et Croft, 1999] [Lavrenko et Croft, 2001].

De façon formelle, chaque texte est vu comme une distribution de probabilité sur ces termes constituants. Pour calculer la probabilité d'une séquence d'un texte (t_1, … , t_k), on le décompose en probabilité d'événements conditionnés successives relatifs à la rencontre des termes t_1 :

$$P(t_1,\ldots,t_k) = P(t_1).P(t_2|t_1).P(t_3|t_1t_2)..P(t_k|t_1..t_{k-1}) \quad \textbf{(3-10)}$$

La forme la plus simple de l'équation (3-10) annule toute dépendance au contexte et donne une estimation indépendante des termes. Ainsi, le modèle de langage uni-gramme peut être formulé comme suit :

$$P_{uni}(t_1,..., t_k) = P(t_1).P(t_2)...P(t_k) \quad \textbf{(3-11)}$$

Le modèle uni-gramme ne considère aucune interdépendance ni aucun ordre sur les termes d'un texte ; c'est la même hypothèse du "sac de mots" déjà vue dans le modèle vectoriel. Néanmoins, cette variante doit être considérée plutôt comme une distribution multinomiale sur les termes du moment où la probabilité d'une séquence reste la même quel que soit l'ordre des termes [Manning et al., 2008].

L'altération de cette hypothèse, en considérant une dépendance binaire entre chaque paire de termes successifs, engendre une nouvelle formulation (3-12) d'un autre modèle appelé le modèle de langage bi-gram :

$$P_{bi}(t_1,\ldots,t_k) = P(t_1).P(t_2|t_1).P(t_3|t_2)..P(t_k|t_{k-1}) \quad \textbf{(3-12)}$$

En incluant l'aspect syntaxique, ce modèle (bi-gram) plus complexe, paraît plus adapté aux tâches linguistiques nécessitant une analyse contextuelle telles que la traduction automatique ou la correction d'orthographe. Contrairement au modèle uni-gramme, la complexité et le coût élevé du calcul dans modèle bi-gram n'encouragent pas de l'utiliser en RI où l'apport des considérations syntaxiques n'est pas assez justifié [Manning et al., 2008] [Bennet et al., 2008].

Le processus de recherche avec un modèle de langage s'articule sur trois étapes principales :

- Un modèle de langage est estimé pour chaque document de la collection,
- Une probabilité de la séquence des termes de la requête est calculée,
- Un classement des documents est établi selon leurs valeurs de probabilité.

Trois approches peuvent être suivies pour développer un modèle de langage dans un processus de recherche :

- ***Modèle de probabilité de requête***, où on cherche à estimer, pour chaque document *D*, la probabilité qu'une requête *Q* soit générée par ce modèle M_D : P(Q |M_D).
- ***Modèle de probabilité de document***, où on cherche à estimer, pour une requête *Q*, la probabilité que le document *D* soit générée par ce modèle M_Q : P(D |M_Q).
- ***Comparaison de modèle***, où document et requête sont spécifiés par le modèle de langage uni-gramme (multinomial) avant de classer les documents de la collection selon la mesure de divergence de Kullback-Leibler (*KL*)par rapport au modèle de langage de la requête M_Q :

$$RSV_{KL}(D,Q) = KL\left(M_D \| M_Q\right) = \sum_{t \in Q} P(t|M_Q) \log \frac{P(t|M_Q)}{P(t|M_D)} \quad \textbf{(3-13)}$$

La divergence *KL* est une mesure asymétrique qui indique à quel point la distribution de probabilité M_Q est mauvaise en considérant le modèle M_D. Certains travaux ont montré une meilleure performance du modèle de comparaison par rapport aux deux modèles de probabilité de requête ou de probabilité de document [Lafferty et Zhai, 2001].

4.7 Autres modèles

Bien que les modèles présentés dans ce chapitre constituent l'essentiel des approches développées pour l'indexation du texte, nous trouvons en littérature d'autres variantes faisant intervenir des méthodes spécifiques.

Par exemple, les relations entre les termes, les documents et la requête peuvent être modélisées par un réseau de neurones multicouche. Les valeurs de la couche de sortie interprètent généralement des critères de pertinence d'un document ou d'expansion d'une requête [Kwok, 1995] [Boughanem et al., 1999].

Par ailleurs, une approche de la logique non-classique a été introduite en RI en se basant sur la sémantique des mondes possibles. Dans un cadre de la logique modale, Rijsbergen proposa d'estimer le degré d'incertitude de $P(D \rightarrow Q)$ par un processus appelé "imaging" [Rijsbergen, 1986] [Zuccon et al., 2008].

Une autre approche, pour modéliser par thème les documents, avait émergé avec le début du millénaire et présente une bonne alternative pour prendre en charge certains aspects sémantiques du texte. Nous décrivons les modèles de thème dans le chapitre suivant en présentant les démarches d'implémentation et d'évaluation.

5 Classification et comparaison des modèles

Les modèles d'indexation textuelle en RI peuvent être classés en trois catégories selon leurs fondements mathématiques (théorie des ensembles, algébrique ou probabiliste) [Kuropka, 2004]. L'interdépendance des termes dans un modèle de RI peut être un autre critère de classification. L'hypothèse de dépendance entre les termes peut être intrinsèque au modèle comme elle peut être transcendante. En ajoutant l'hypothèse d'indépendance entre les termes, trois classes d'indépendance peuvent être définies pour les modèles de RI (indépendance, dépendance intrinsèque, dépendance transcendante) [Kuropka, 2004].

D'autres taxonomies ont été proposées pour classer les modèles de RI. Par exemple, le type d'appariement entre requête et document peut scinder ces modèles en "comparaison flexible" ou "comparaison stricte" [Nie, 1990]. Une autre étude avait proposé une classification selon deux points de vue : Le premier consiste en une taxonomie verticale selon la manière de représentation des documents et des requêtes. Le deuxième s'articule autour des types des tâches en RI dans une taxonomie horizontale [Canfora et Cerulo, 1990].

Comme récapitulatif des modèles de document en RI, nous recensons trois approches principales selon leur fondement mathématique : ensembliste, algébrique ou probabiliste. Nous pouvons affiner cette classification en ajoutant une quatrième approche statistique relative aux modèles des réseaux Bayésiens et les modèles de langage. Néanmoins, on peut garder ces modèles dans l'approche globale probabiliste. Les modèles de thème, que nous développons dans le chapitre suivant, s'apparentent à l'approche statistique, et donc probabiliste, mais on peut les considérer dans une nouvelle classe, celle des modèles à variable latente. Par ailleurs, nous devons citer l'approche logique comme cinquième classe regroupant les modèles proposés dans [Rijsbergen, 1986] [Zuccon et al., 2008]. Nous avons préféré de ne pas les détailler davantage dan la section précédente puisqu'ils ne sont pas assez étudiés et développés.

Afin d'apprécier la qualité de chaque modèle, nous proposons de dresser un tableau comparatif des modèles en RI sur la base des critères suivants :

- L'approche mathématique,
- La solidité du fondement théorique,

- La pondération des termes (possibilité d'ordonnancement),
- L'interdépendance des termes,
- La flexibilité de la mesure de similarité ou de la pertinence,
- La performance globale des SRI relatifs,
- La possibilité de réduction de dimension,
- La prise en compte des aspects sémantiques.

Notons que la capacité d'un modèle d'ordonnancement les résultats dans une recherche ad-hoc est une caractéristique fortement liée à sa faculté d'attribuer un poids aux termes d'indexation. Dans Tableau 3-4, nous décrivons les principaux modèles selon les critères cités précédemment.

Caractéristiques / **Modèles**	**Approche**	**Fondement théorique**	**Pondération des termes**	**Interdépendance des termes**	**Flexibilité de similarité**	**Performance**	**Réduction de dimension**	**Aspect sémantique**
Booléen	Ensembliste	+	-	-	-	-/+	-	-
Booléen étendu	Ensembliste / Algébrique	+	-/+	-	-/+	-/+	-	-
Ensembles flous	Ensembliste	+	-/+	-/+	-/+	-/+	-	-
Vectoriel tf-idf	Algébrique	+	+	-/+	+	+	-	-
Vectoriel généralisé	Algébrique	+	+	+	+	+	-	-/+
LSI	Algébrique	+	+	+	+	+	+	+
Indépendance binaire	Probabiliste	++	-/+	-	+	-/+	-	
Okapi-BM25	Probabiliste	++	+	-	+	+	-	-
Réseaux Bayésiens	Probabiliste	++	+	+	+	+	-/+	-/+
Modèle de langage	Probabiliste	++	+	+	-/+	+	-	-/+

Tableau 3-4. Comparaison des modèles de document en RI.

Le modèle pionnier en RI est le modèle booléen. Bien qu'il soit très rapide et simple à implémenter, il ne propose de pondération des termes (et par conséquent l'ordonnancement de la pertinence) que dans certaines variantes. L'appariement est effectué de façon exacte et l'interdépendance des termes n'est prise en charge que dans le modèle des ensembles flous.

Le modèle vectoriel utilise un formalisme algébrique pour définir des mesures de similarité flexibles et par conséquent, pouvoir ordonner les résultats de recherche. La requête peut être formulée en langage naturel (et non pas par composition booléenne de mots clés). Elle est représentée par un vecteur dans le même espace vectoriel des documents. L'interdépendance des termes ne peut être prise en compte que dans le modèle vectoriel généralisé. Dans certaines variantes de ce modèle, il est possible d'introduire une connaissance linguistique externe traçant des liens taxonomiques utiles à l'indexation sémantique. Mais c'est le modèle LSI qui propose une formulation algébrique efficace offrant

deux avantages en même temps : la prise en charge des aspects sémantiques (synonymie et polysémie) et la réduction de la dimension de l'index.

L'approche probabiliste offre les mêmes qualités du modèle vectoriel mais avec une base théorique saine [Croft et al., 92]. De plus, le tri des documents respecte une probabilité de pertinence, des documents par rapport au besoin d'information des utilisateurs, au lieu d'une mesure de similarité avec la requête comme c'est le cas dans le modèle vectoriel. Néanmoins, le modèle vectoriel reste l'un des modèles de RI classique les plus influents, les plus étudiés et les mieux acceptés.

6 Conclusion

Ce chapitre était consacré aux principaux modèles de représentation des textes pour les tâche de RI. L'objectif de cette exploration d'identifier une approche efficace pour prendre en charge la sémantique imbriquée dans le texte non-structuré tout en réduisant la dimensionnalité. Les modèles de langage multi-grammes présentent un cadre unifié pour l'indexation et la recherche textuelles. Partant d'une approche statistique solide et élégante, ils capturent tous les avantages de la modélisation documentaire pour les tâches de RI. La complexité grandissante des implémentations, en bi-grammes ou tri-grammes, ralentisse leur intégration dans des applications réelles à grande échelle. Tirer profit de la théorie des modèles de langage et des techniques de réduction par indexation sémantique latente, peut aboutir à des modèles plus efficaces que nous décrivons dans le chapitre suivant.

Chapitre 4 :
MODELISATION PAR THEME DES TEXTES NON STRUCTURES

1 Introduction

Le chapitre précédent nous a tracé une vue panoramique des principales approches d'indexation des documents pour les tâches de RI. Les méthodes décrites opèrent toutes au niveau lexical du document en vue de représenter son contenu textuel propre. La pertinence d'un document devrait être mesurée par rapport à une représentation équivalente de la requête (ou à une description caractérisant la catégorie). Cependant, la satisfaction d'un utilisateur dépend de son appréciation par rapport à un besoin contextuel d'information. La formulation lexicale de ce besoin, qui n'est pas toujours exacte, interprète plutôt un thème spécifique ou une combinaison de sujets reliés au besoin contextuel de l'utilisateur. Avec le début du troisième millénaire, une nouvelle vision d'indexation de texte émergea en proposant la modélisation par thème pour prendre en charge l'aspect sémantique du contenu. Cette approche tente d'indexer les documents, non par les composantes lexicales du vocabulaire, mais plutôt par des variables latentes interprétant les thèmes cachés par les termes utilisés dans le document lui-même.

Partant d'une approche probabiliste, solide théoriquement, nous commençons par décrire un modèle de base simple (uni-gramme) avant de présenter le modèle d'indexation en sémantique latente probabiliste (pLSI) puis l'allocation latente de Dirichlet (LDA). Nous analysons les méthodes d'évaluation et certains aspects pratiques de la modélisation par thème.

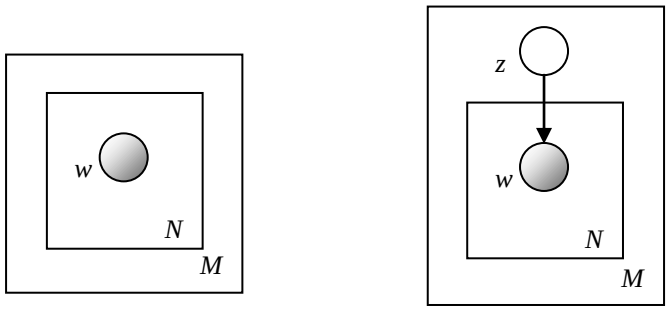
2 Le modèle de mélange d'uni-grammes

Nous avons déjà vu dans le chapitre précédent que le modèle de langage uni-gramme suppose l'interdépendance des termes "hypothèse du sac de mots". Il est interprété tout simplement en une unique distribution multinomiale. Ce modèle peut être augmenté par une variable aléatoire discrète *z* dans l'équation (3-11) pour obtenir un modèle génératif pour modéliser les documents d'une collection selon l'équation (4-1) :

$$P(d) = \sum_{z} P(z) \prod_{i=1}^{N} P(w_i | z) \qquad \textbf{(4-1)}$$

Ce modèle de mélange d'uni-grammes génère chaque document par le choix initial d'un thème *z* avant de générer indépendamment *N* mots à partir de la distribution multinomiale conditionnelle $P(w|z)$ [Nigam et al, 2000] [Blei et al., 2003]. En d'autres termes, chaque document contient un seul thème où l'ensemble des thèmes possibles doit être prédéterminés. Ce modèle convient au problème de classification supervisée où chaque valeur de *z* correspond à l'étiquette d'une classe. Dans ce cas, on peut estimer les paramètres de ces distributions à partir d'une collections de documents préalablement mono-étiquetés.

Nous pouvons utiliser une représentation graphique, selon la notation de rectangle (plate) [Buntine, 1994], pour illustrer le modèle de mélange d'uni-grammes comme dans la Figure 4-1. Chaque nœud du graphe représente une variable aléatoire, une transition interprète une dépendance statistique et le rectangle reflète la répétition au nombre inscrit sur le coin bas droit.



(a) uni-gramme **(b) mélange d'uni-gramme**

Figure 4-1. Illustration graphique des modèles uni-gramme et mélange d'uni-grammes.

Le modèle de mélange d'uni-gramme présente une alternative utile au problème de catégorisation des textes. Dans le cas d'absence d'étiquetage préalable, le problème est reformulé dans un cadre d'agrégation (*en angl. clustering*) dont la solution peut être traitée par l'algorithme de maximisation d'espérance (*EM*) introduit par [Dempster et al., 1977]. Cependant, le modèle d'uni-gramme présente certaines insuffisances pour modéliser les textes dans des collections volumineuses. D'une part, un document n'est supposé contenir qu'un seul thème (catégorie) et d'autre part, les distributions manquent de probabilités à priori et sont supposées être apprises complètement de la collection [Blei et al., 2003].

3 Le modèle pLSI

Le modèle vectoriel d'indexation sémantique (LSI) que nous avons décrit dans le chapitre précédent utilise efficacement une technique de réduction de dimension (SVD) dans la matrice terme×document afin de créer un espace de sémantique latente. Néanmoins, le modèle LSI manque, en premier lieu, de justification théorique solide pour son usage en RI.

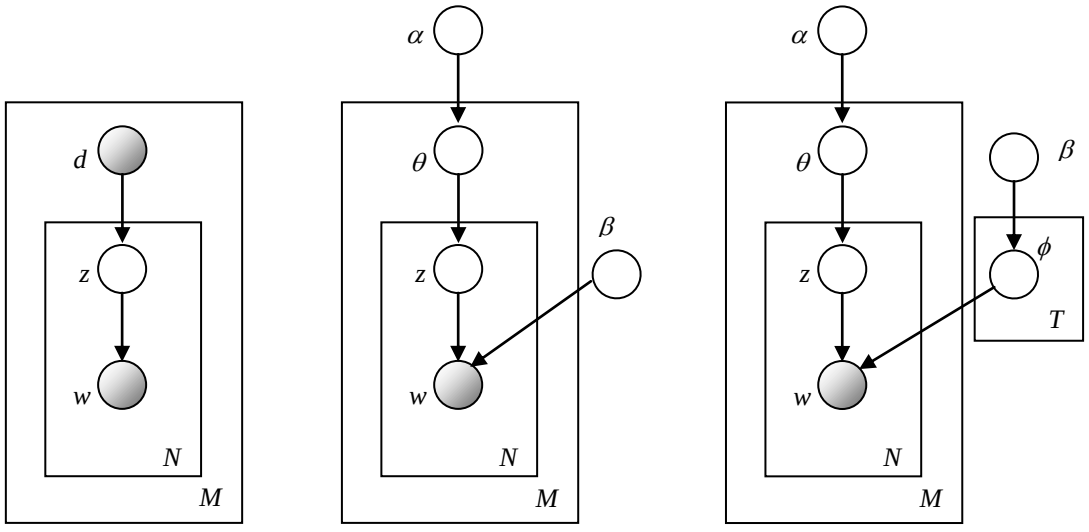
Le modèle d'indexation sémantique latente probabiliste (pLSI) part d'une base théorique solide et propose une formulation probabiliste (4-2) pour atteindre le même objectif que la traditionnelle LSI. Aussi connu par le modèle d'aspect, pLSI propose de relaxer l'hypothèse simplificatrice du mélange d'uni-grammes où chaque document ne puisse être généré que d'un seul thème [Hofmann et al., 1999]. Ainsi, pLSI suppose qu'un document *d* peut contenir plusieurs thèmes et qu'il est conditionnellement indépendant d'un mot w_i sachant le thème latent *z* [Hofmann et al., 1999] [Blei et al., 2003] :

$$P(d, w_i) = P(d)\sum_z P(w_i|z)P(z|d) \qquad \textbf{(4-2)}$$

Bien qu'il présente un cadre formel pour l'analyse multi-thème, le modèle pLSI est critiqué pour deux insuffisances assez importantes : d'une part, le modèle crée un index fictif sur les seuls documents d'apprentissage et ne fournit pas de moyen clair pour estimer une probabilité d'un nouveau document hors collection. D'une autre part, le nombre de paramètres à estimer augmente linéairement avec le nombre *M* des documents ce qui complique les calculs dans les collections volumineuses même avec des heuristiques de lissage [Popescul et al., 2001] [Blei et al., 2003].

4 Le modèle LDA

L'allocation latente de Dirichlet (LDA) est un modèle génératif de thèmes pour les documents textuels [Blei et al., 2003]. Le modèle est proposé dans l'objectif d'améliorer la flexibilité du pLSI dans les problèmes de RI. LDA peut être illustrée graphiquement dans (Figure 4-2) où deux hyperparmatres (α,β) sont rajoutés pour avoir une distribution de Dirichlet à priori dans l'estimation du modèle. La version lissée de LDA est proposée dans l'objectif de prendre en charge les mots non observés (hors vocabulaire) dans les nouveaux documents. Ceci évite d'attribuer une probabilité nulle aux mots non observés dans le corpus d'apprentissage.



(a) pLSI **(b) LDA** **(c) LDA lissée**

Partant de l'hypothèse de "sac de mots", l'ordre (d'apparition) des mots dans un document, comme celui des documents dans la collection, est ignoré. En statistique, cette hypothèse est assimilée à la propriété d'interchangeabilité dans une séquence de variables aléatroires. LDA considère chaque document (*d*) comme une combinaison de thèmes (θ_d).

Chaque thème est défini par une distribution de probabilité sur les mots (w) dont chacun peut être estimé comme suit sachant un document d :

$$P(w_i|d) = \sum_{j=1}^{T} P(w_i|z_i = j)P(z_i = j|d)\,. \qquad \textbf{(4-3)}$$

Où $P(w|z=j)=\phi_j$ désigne la distribution multinomiale sur les mots w sachant le thème j. $P(z|d)$ représente la distribution sur les thèmes z dans une collection de mots (document d) [Blei et al., 2003]. L'hyperparamètre α interprète la distribution à priori sur les thèmes avant toute observation d'un document du corpus. L'hyperparamètre β peut être vu comme le nombre de fois à priori où les mots sont échantillonnés d'un thème (avant toute observation d'un mot du corpus). Ces hyperparamètres doivent être déterrminés arbitrairement par ajustement préliminaire dans chaque situation. Néanmoins, certaints travaux de recherche ont trouvé que l'estimation du modèle LDA est satisfaisante dans plusieurs collections lorsqu'on choisit (α=50/T, β=0,01) [Griffiths et Steyvers, 2004] [Steyvers et Griffiths, 2007].

4.1 Processus génératif

Pour un nombre donné de thèmes T, l'apprentissage (ou l'inférence) du modèle LDA est est appmiqué dans une collection de documents définie comme suit :

> N : nombre de mots du vocabulaire.
>
> N_d : nombre de mots dans le document d.
>
> M : nombre de documents dans la collection.
>
> T : nombre de thèmes, donné en entrée.
>
> $P(z)$: distribution sur les thèmes z dans un document.
>
> $P(w|z)$: distribution de probabilité sur les mots w sachant le thème z.

Ensuite nous définissons le processus génératif comme suit :

> Pour chaque document $d = 1 \ldots M$ (dans la collection) faire :
>
> 1- Echantillonner une distribution de thème $\theta_d \sim Dir(\alpha)$
>
> 2- Pour chaque mot $w_{di} = 1 .. N_d$ faire :
>
> 2a- Choisir un thème $z_{di} \in \{1, \ldots, T\} \sim Multinomial(\theta_d)$
>
> 2b- Choisir un mot $w_d \in \{1, \ldots, N\} \sim Multinomial(\beta_{zdi})$

Où α est un vecteur de dimension T paramétrant la distribution de Dirichlet des thèmes (θ) dans chaque document. Les $\{\beta_i\}$ représentent les paramètres multinomiaux de thèmes dans toute la collection. Chaque β_i attribue une probabilité élevé pour un ensemble spécifique de mots sémantiquement consistants. C'est cette distribution sur le vocabulaire qui désignera le thème. Dans la formulation originale du modèle LDA, les auteurs indiquent qu'il est possible de traiter seulement une partie du document en choisissant, tout d'abord, une longueur inférieure à N_d pour chaque document selon une distribution de Poisson [Blei et al., 2003]. Néanmoins, il est préférable, autant que possible, de considérer la totalité du document. Ceci n'affecte en rien les paramètres de génération (θ,z) mais cette alternative reste une variante de choix avec certaines contraintes de calcul ou de nature des textes.

Par ailleurs, les hyper-paramètres $\{\beta_i\}$ peuvent être lissés afin d'attribuer une probabilité non-nulle à tous les mots possibles du vocabulaire, en particulier pour ceux des nouveaux documents non rencontrés dans la phase d'apprentissage [Blei et al., 2003] [Wallach, 2006].

4.2 Interprétation géométrique

Pour comparer le modèle LDA aux autres modèle de thème, une illustration géométrique peut être tracée dans l'espace des thèmes [Blei et al., 2003]. Dans la Figure 4-3, un simplexe est défini pour trois mots où chaque point intérieur représente une distribution multinomiale sur ces trois mots. Les sommets du simplexe correspondent aux distributions qui affecte une probabilité maximale à l'un des mots et annule celle des deux autres. Autrement, toute autre distribution alternative peut être considérée. En particulier, on peut considérer un autre simplexe de trois thèmes dont chacun représente une distribution sur les trois mots.

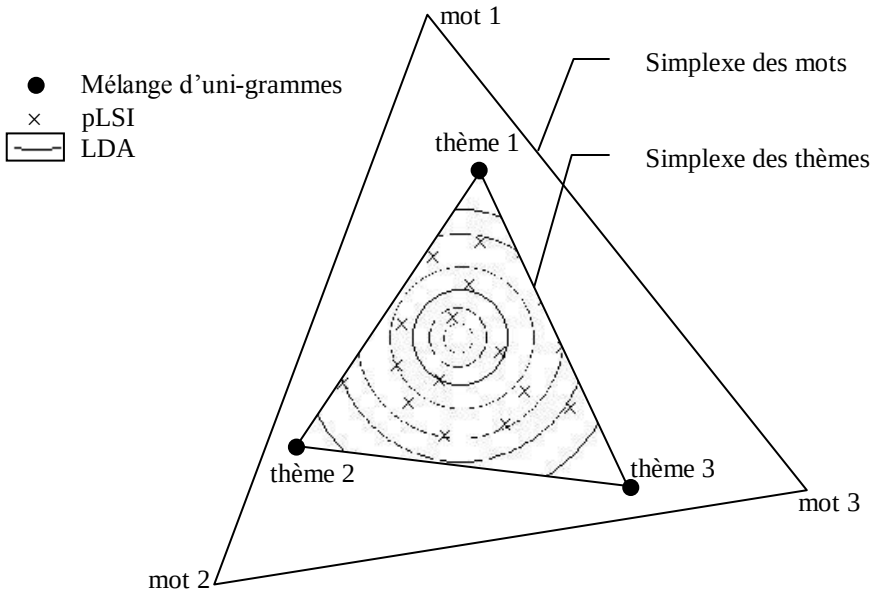


Figure 4-3. Illustration géométrique des modèles de thème.

Le modèle de mélange d'uni-grammes place chaque document, qui ne contient qu'un seul thème, sur l'un des sommets. Le modèle pLSI attribue une distribution sur les thèmes à chaque document de la collection qui sera donc représenté par un point (×) à l'intérieur du simplexe des thèmes. Cependant, le modèle LDA élargie la représentation en traçant des contours à l'intérieur du simplexe des thèmes pour capturer même les nouveaux documents.

4.3 Apprentissage du modèle LDA et échantillonnage de Gibbs

L'inférence du modèle LDA revient à estimer les distributions (θ) et (ϕ) sachant les distributions à priori (α) et (β) et les observations des mots (w) dans chaque document de la collection. Blei et al. ont proposé un algorithme d'inférence variationnel pour l'estimation des paramètres Bayésien empiriques [Blei et al., 2003]. Bien que cette technique achève l'apprentissage dans un temps raisonnable, les distributions postérieures inférées manquent de précision [Porteous et al., 2008].

Griffiths et Steyvers ont proposé un algorithme d'apprentissage basé sur l'échantillonnage de Gibbs [Griffiths et Steyvers, 2004]. Cet algorithme est devenu plus recommandé pour LDA puisqu'il fournit une estimation plus précise des paramètres du modèle [Porteous et al., 2008]. Certaines implémentations sont actuellement disponibles en ligne avec un code source

ouvert. Parmi lesquels, le package *LingPipe*[14] et *Dragon*[15] qui proposent des implémentations évolutives pour l'apprentissage LDA dans un cadre global d'analyse linguistique et statistique des documents textuels.

L'idée proposée dans [Griffiths et Steyvers, 2004] part du fait qu'il suffit d'estimer la distribution à postériori sur z (l'attribution de chaque occurrence de mot observé à l'un des thèmes). Ceci évite de chercher à estimer directement les distributions ϕ (thème-mot) et celle des thèmes θ pour chaque document. Nous pouvons exprimer cette approche dans l'equation (4-4) :

$$P(z|w) = \frac{P(w|z)}{\sum_z P(w|z)} \qquad \textbf{(4-4)}$$

Vu la taille élevée qu'une collection de texte puisse avoir (des millions de mots), cette distribution est difficile à résoudre analytiquement (factorisation du dénominateur sur T^N termes). On peut reformuler le problème, par la distribution conditionnelle $P(z_i \mid z_{-i}, w)$, en estimant l'attribution d'un thème z_i sachant les autres distributions z_{-i} [Griffiths et Steyvers, 2004]. Ce qui donne l'équation (4-5) :

$$P(z_i = j|z_{-i}, w) \propto \frac{C^{NT}_{w_i j} + \beta}{\sum_{w=1..N} C^{NT}_{w_i j} + N\beta} \cdot \frac{C^{MT}_{d_i j} + \alpha}{\sum_{t=1..T} C^{MT}_{d_i t} + T\alpha} \qquad \textbf{(4-5)}$$

En gardant les mêmes notations dans ce chapitre, on dénote par C^{NT} la matrice de dimension $N \times T$ dont chaque élément c_{wj} représente le nombre d'attributions du mot w au thème j sans inclure l'occurrence actuelle i. De façon analogue la matrice C^{NT} est définie pour les attributions document×thème.

L'estimation de telles distributions jointes à haute dimensionnalité peut être réalisée par des techniques itératives d'approximation telles que la méthode de Monte-Carlo par chaines de Markov [Gilks et al. 1996]. Une forme spécifique et efficace de cette méthode est l'échantillonnage de Gibbs appliqué séquentiellement jusqu'à l'approximation de la distribution cible [Griffiths et Steyvers, 2004] [Steyvers et Griffiths, 2007].

Les distributions ϕ et θ peuvent être déduites par estimation du maximum à posteriori (MAP) selon les deux équation suivantes :

$$P(w|t) = \phi'^{(t)}_w = \frac{C^{NT}_{wt} + \beta}{\sum_{w=1..N} C^{NT}_{wt} + N\beta}, \qquad \textbf{(4-6)}$$

$$P(t|d) = \theta'^{(d)}_t = \frac{C^{MT}_{dt} + \alpha}{\sum_{t=1..T} C^{MT}_{dt} + T\alpha} \qquad \textbf{(4-7)}$$

[14] Disponible sur http://alias-i.com/lingpipe/index.html

[15] Disponible sur http://dragon.ischool.drexel.edu/default.asp

Vu l'efficacité des approximations résultantes, l'échantillonnage de Gibbs a été largement appliqué dans les implémentations du modèle LDA [Griffiths et Steyvers, 2004] [Zhou et al., 2007].

4.4 Exemple d'analyse sémantique dans le corpus CF

L'apprentissage du modèle LDA est appliqué dans le présent travail par échantillonnage de Gibbs après adaptation de l'implémentation dans le package *LingPipe*. Afin de présenter un exemple de la modélisation par thème dans un cas réel, nous choisissons le corpus "Cystic-Fibrosis" (*CF*), sous-ensemble de la base de données *MEDLINE*, contenant 1.239 articles publiés entre 1974 et 1979 discutant des aspects de la mucoviscidose (ou la fibrose kystique) avec 100 requêtes jugées par quatre catégories de personnes [Shaw et al., 1991]. Nous utilisons, dans cet exemple illustratif, 1.229 résumés pour l'apprentissage du modèle LDA avec 8 thèmes latents. Les résumés ont été prétraités pour ne garder que les noms et les verbes significatifs[16].

Nous affichons, dans Tableau 4-1 et Tableau 4-2 et selon l'ordre décroissant des probabilités, les 20 premiers mots pour chaque distribution de thème. Les titres ont été attribués manuellement, par un spécialiste en hématologie et cytogénétique, après lecture préliminaire des mots les plus probables dans chaque thème.

Thème 1		**Thème 2**		**Thème 3**		**Thème 4**	
Infection à pseudomonas en mucoviscidose pédiatrique		**Aspect familial de la mucoviscidose**		**Caractéristiques biochimiques en mucoviscidose**		**caractéristiques enzymatiques en mucoviscidose**	
Mot	**prob.**	**mot**	**prob.**	**mot**	**prob.**	**mot**	**prob.**
patient	,091	fibrosi	,043	patient	,076	serum	,042
infection	,033	disease	,035	cf	,059	activity	,035
sputum	,028	patient	,023	control	,044	protein	,032
cf	,023	child	,018	serum	,041	fibrosi	,029
fibrosi	,021	diagnosi	,012	fibrosi	,037	cf	,027
pseudomona	,020	treatment	,009	normal	,029	alpha	,024
strain	,020	cf	,009	level	,027	control	,020
antibody	,019	disorder	,008	concentration	,024	plasma	,019
response	,014	family	,008	subject	,023	sample	,016
mucoid	,014	study	,008	age	,023	fraction	,016
precipitin	,014	include	,008	activity	,020	normal	,016
test	,013	discu	,008	value	,018	plu	,015
child	,013	review	,007	found	,017	enzyme	,014
cent	,013	affect	,007	compare	,017	minu	,013
found	,013	parent	,007	difference	,016	patient	,013
antigen	,012	population	,006	study	,015	ph	,012
asthma	,012	report	,006	mean	,014	detect	,011
serum	,011	age	,006	cla	,013	factor	,010
type	,011	development	,006	increase	,011	gel	,010
tract	,011	adult	,006	child	,010	trypsin	,010

Tableau 4-1. Distribution des thèmes (1-4) dans un modèle LDA_8 du corpus *CF*.

[16] Nous décrivons dans le chapitre suivant les méthodes de prétraitement linguistique.

A titre d'exemple, deux documents de la collection sont présentés dans Figure 4-4. Leur distribution sur les 8 thèmes latents est présentée dans Figure 4-5. Le modèle LDA nous informe que l'article 126 discute du *syndrome de malabsorption chez les enfants* (thème 5) mais concerne aussi l'*étude cellulaire dans la sécrétion de mucoviscidose* (thème 8) et ces *caractéristiques biochimiques* (thème 3). Le document 127 discute principalement de l'*aspect familial de la mucoviscidose* (thème 2).

Thème 5		**Thème 6**		**Thème 7**		**Thème 8**	
syndrome de malabsorption chez les enfants		**Exploration fonctionnelle respiratoire chez les enfants atteints de mucoviscidose**		**Analyse du méconium chez les enfants atteints de mucoviscidose**		**Etude cellulaire dans la sécrétion de mucoviscidose**	
mot	**prob.**	**mot**	**prob.**	**mot**	**prob.**	**mot**	**prob.**
acid	,055	patient	,055	test	,051	cf	,049
child	,030	fibrosi	,027	fibrosi	,048	cell	,045
patient	,027	lung	,021	sweat	,042	fibrosi	,025
fibrosi	,024	volume	,020	meconium	,027	normal	,022
bile	,021	flow	,015	infant	,024	fibroblast	,019
vitamin	,021	therapy	,014	chloride	,018	secretion	,018
fatty	,020	treatment	,014	screen	,017	culture	,015
absorption	,016	function	,014	method	,015	study	,014
deficiency	,016	study	,014	sodium	,014	increase	,012
enzyme	,015	rate	,013	child	,014	serum	,011
disease	,013	child	,013	pancrea	,013	glycoprotein	,011
level	,013	increase	,013	normal	,013	effect	,011
diet	,012	day	,011	patient	,012	mucu	,011
fat	,012	airway	,011	positive	,012	gland	,011
excretion	,012	effect	,010	result	,012	transport	,010
malabsorption	,011	change	,010	found	,011	factor	,010
growth	,010	measure	,009	diagnosi	,010	observe	,009
plasma	,010	treat	,009	report	,010	difference	,009
increase	,009	disease	,008	newborn	,010	medium	,008
therapy	,009	dose	,008	albumin	,009	membrane	,008

Tableau 4-2. Distribution des thèmes (5-8) dans un modèle LDA_8 du corpus *CF*.

The nature of the metabolic defect in cystic fibrosis (CF) has, to date, not been elucidated. Elliott reported that intravenous administration of soybean oil and lecithin to children suffering from CF caused their sweat sodium concentration to decrease towards normal values. This prompted us to investigate the fatty acid spectra of the various serum lipid classes of children with CF. We have found that affected children are deficient in essential fatty acids. The data suggest another approach to the study of the metabolic defect in cystic fibrosis. There is an apparent deficiency in essential fatty acids, and a marked reduction in serum vitamin E levels has also been reported in CF. These conditions might be partially corrected by intravenous injection, or possibly feeding of, lipids containing essential fatty acids with concomitant administration of vitamin E. The observed deficiency in essential fatty acids may be produced by a reduced ability to absorb these acids from the diet and this, in turn, may result in defects in membrane structure or stability. Reduced levels of vitamin E in CF may reflect a reduced requirement for this vitamin because of the lower levels of essential fatty acids. It is also possible that the observed deficiency in serum levels of essential fatty acids may lead to less than normal production of prostaglandins.

Article 126

A tool to assess psychosocial adjustment of school-age children with cystic fibrosis was developed using three instruments: a standardized open-ended parent interview, a self-administered teacher questionnaire, and a self-administered parent demographic data form. The three instruments enabled the pediatric nurse practitioner to make a clinical assessment of the child's psychosocial adjustment. This assessment was then checked for validity by means of a social worker interview. Implications for further research are discussed.

Article 127

Figure 4-4. Exemples d'articles de la collection *CF*.

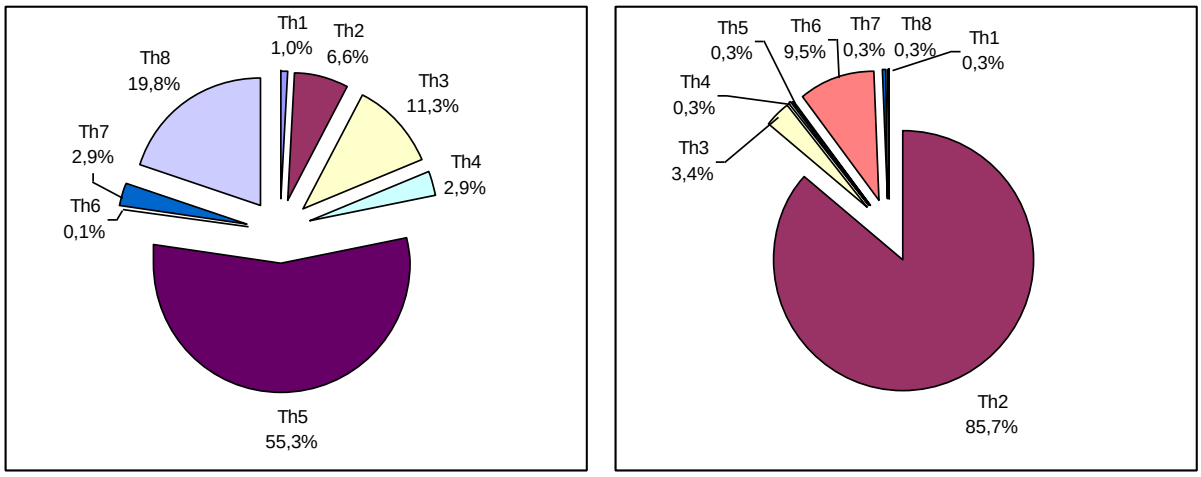


Figure 4-5. Distribution des articles 126 et 127 sur 8 thèmes latents.

4.5 Travaux et développements

L'introduction de LDA avait permis de l'utiliser comme un modèle de langage pour représenter les documents dans les tâches de classification et de filtrage collaboratif [Blei et al., 2003]. Les auteurs ont utilisé principalement deux collections (articles bibliographiques CGC et Reuters-21578) pour comparer la capacité de généralisation (par mesure de probabilité du modèle sur l'ensemble de test). En comparant aux autres modèles de thèmes (mélange d'uni-grammes et pLSI, les expérimentations ont montré une nette amélioration des performances avec LDA.

Depuis son introduction originale, différentes applications et variantes du modèle LDA se sont succédées. Nous pouvons résumer ces développements dans quatre directions :

1. Amélioration de l'algorithme d'apprentissage [Griffiths et Steyvers, 2004] [Nallapati et Cohen, 2008] [Porteous et al., 2008],
2. Traitement des collections volumineuses [Newman et al., 2007],
3. Développer d'autres variantes du modèle de thème [Chemudugunta et al., 2008] [Nallapati et Cohen, 2008],
4. Visualisation et interprétation des résultats [Iwata et al., 2008].

Par ailleurs, l'hypothèse de "sac de mots" sur laquelle repose le modèle LDA néglige toute caractéristique syntaxique pouvant être d'une utilité capitale dans la description du document. Une extension du modèle LDA a été proposée dans ce sens afin prendre en compte l'ordre des mots et, par conséquent, apprendre l'aspect syntaxique du texte [Griffiths et al., 2005].

Néanmoins, peu de travaux sur le modèle LDA se sont intéressés aux aspects multi-langage ou, laissons-nous dire clairement, aux applications sur le texte non-anglais. Ce n'est que récemment, que certaines études ont commencé à établir (timidement) un cadre compréhensif pour la modélisation par thèmes multi-langages ou non-anglais. A titre d'exemple, une méthode a été proposée et appliquée pour extraire des thèmes anglais-espagnols et anglais-allemands à partir de collections parallèles [Jagaralamudi et Daumé, 2010]. Une autre étude a proposé d'étendre le modèle pLSI pour une analyse en sémantique latente probabiliste en langages croisés [Zhang et al., 2010].

5 LDA pour la classification multi-thèmes

5.1 Similarité dans l'espace des thèmes

Comme avec pLSI, le modèle LDA calcule, pour chaque document du corpus d'apprentissage, une distribution multinomiale relative à T variables latents (thèmes). Malheureusement, le modèle pLSI reste muet envers la modélisation d'un document hors corpus. LDA, par contre, propose un cadre bien fondé pour formuler tout nouveau document (non observé) en calculant la distribution a posteriori de chacun de ses mots sur les thèmes appris du modèle. L'équation (4-6) peut être utilisée pour estimer une nouvelle collection de mots (document) à partir du modèle déjà appris d'un corpus.

Ceci permet de caractériser tout document dans un espace réduit de thèmes. Ainsi, toute mesure de similarité pourra être considérée dans l'espace vectoriel latent (voir Chapitre 3 :3.1). On peut donc facilement ramener le problème de classification supervisée (catégorisation) vers ce nouvel espace engendré par l'apprentissage du modèle LDA dans un corpus donné. La catégorisation multi-classe peut être réalisée en mono-étiquète comme en multi-étiquète. Les séparateurs à vaste marge (SVM) ont été appliqués dans l'espace réduit des thèmes [Blei et al., 2003].

Les auteurs du modèle LDA original ont obtenu des performances similaires sur deux classifications binaires dans l'espace des thèmes ou celui des mots [Blei et al., 2003] [Blei et al., 2006]. Néanmoins, ces tests ont été appliqués sur un texte anglais sans aucune description du prétraitement linguistique adopté. Par ailleurs, la classification binaire n'est pas assez satisfaisante pour apprécier l'impact de l'approche de classification dans l'espace des thèmes. Nous proposons dans ce qui suit d'évaluer la catégorisation multi-classes dans l'espace des thèmes anglais de collections réelles. En plus, nous analyserons, dans Chapitre 6 :5, la faisabilité de considérer les aspects linguistiques pour la catégorisation dans l'espace des thèmes du texte arabe.

5.2 Expérimentations

Nous présentons dans cette section quelques résultats des évaluations préliminaires de la catégorisation du texte anglais dans l'espace des thèmes. Nous retenons la mesure du taux de reconnaissance par validation croisée à *5-fold* (voir Chapitre 2 :7.4). Nous utilisons la méthode SVM pour un problème de classification multi-classes avec le simple noyau linéaire (produit scalaire) en fixant le paramètre de coût à 10. Deux corpus de texte anglais sont utilisés pour les tests :

1. ***WebKb*** est une collection de pages Web des départements d'informatique de plusieurs universités américaines. Ce corpus est réalisé dans le cadre du projet ***World Wide Knowledge Base (Web->Kb)***[17] pour développer des bases de connaissances reflétant le contenu Web. 4518 pages sont manuellement classées dans 6 catégories (*student, faculty, staff, department, course, project*). Nous ne considérons pas le reste des documents non étiquetés (3764 pages classées dans *others*). La taille du corpus est d'environ 8,5Mo avec plus de 1,3 millions de mots, soit une longueur moyenne de 295 mots par document.

[17] http://www.cs.cmu.edu/afs/cs.cmu.edu/project/theo-11/www/wwkb/index.html

2. ***AFP-22k*** est une collection d'articles anglais de l'agence de *France-Presse*. Le corpus a été automatiquement extrait[18] par une application de Web-Crawling (voir Chapitre 6 :2). Ce corpus contient 22.135 articles relatifs à la période 2008-2010 et répartis sur 11 catégories (*Afrique, Amérique, Asie-Pacific, Economie, Culture, Europe, France, Santé, Moyen-Orient, Sciences et Sport*). La taille du corpus dépasse les 50Mo avec plus de 9 millions de mots, soit une longueur moyenne de 407 mots par document.

En utilisant la même liste des mots vides à ignorer[19], le contenu (texte anglais) de chaque corpus a été préalablement analysé selon deux méthodes de stemming (nous décrivons ultérieurement les méthodes de prétraitement linguistique, Chapitre 5 :4) :

- Stemming léger ou pseudo-racinisation par l'algorithme de *Porter* [Porter, 1980],
- Lemmatisation basée sur les ressources lexicales de *WordNet* [Miller et al., 1990].

L'apprentissage du modèle LDA a été réalisé pour une large gamme de nombre de thèmes (entre 2 et 500). Nous reportons dans ce qui suit les configurations qui ont abouti sur les meilleures performances (T = 100, 200, 300). Pour la comparaison, nous avons appliqué la catégorisation dans l'espace des termes (stem ou lemme) en utilisant la pondération ***TF*** et ***TF-IDF***.

Corpus	**Nature du mot**	**TF**	**TF-IDF**	**LDA-100**	**LDA-200**	**LDA-300**
WebKb	**Stem**	**85,88**	83,64	80,35	**81,41**	80,85
	Lemme	**82,82**	74,55	79,55	81,16	**81,23**
AFP-22k	**Stem**	**78,00**	74,65	80,03	**81,86**	81,70
	Lemme	**78,01**	74,04	79,85	81,85	**82,00**

Tableau 4-3. Taux de reconnaissance de la classification des corpus *WebKb* et *AFP-22k*.

Les performances de classification enregistrées dans **Tableau 4-3** nous indiquent que la classification dans l'espace des thèmes est légèrement inférieure à celle appliquée dans l'espace des mots. Cependant elle s'améliore de façon significative lorsque l'apprentissage des thèmes (LDA) s'applique dans de larges corpus où la longueur des documents est plus importante. La nature des mots descripteurs (stems ou lemmes) influence, parfois même considérablement, l'efficacité de la classification dans l'espace des termes. La modélisation par thème permet donc de réaliser une classification globalement équivalente à celles opérée dans l'espace des termes. Elle peut être plus performante lorsque les textes soient assez nombreux et plus développés. De plus, elle s'avère très efficace pour la réduction de l'espace de représentation des documents (de quelques dizaines de milliers de mots vers quelques centaines de thèmes).

Par ailleurs, l'impact des méthodes de prétraitement linguistique reste presque insignifiant dans l'espace des thèmes. La faisabilité d'opérer une lemmatisation très couteuse (par rapport à la pseudo-racinisation) n'est pas assez justifié pour la tâche de classification. Ce constat concerne principalement le texte anglais dans lequel sont appliqués la majorité des modèles de

[18] Du site http://www.france24.com/en/

[19] 423 mots vides proposés par le projet ONIX http://www.lextek.com/manuals/onix/stopwords1.html

recherche d'information. La généralisation de ce résultat dans d'autres langues n'est pas aussi évidente et nécessite une large investigation (voir Chapitre 5 :1).

6 LDA pour la recherche ad-hoc

La faisabilité d'introduire la modélisation par thème dans la recherche ad-hoc n'est pas assez claire. Peu de travaux récents ont proposé certaines approches pour améliorer le calcul de la pertinence mais les résultats restent mitigés et suscitent de plus amples recherches [Wei et Croft, 2006] [Yi et Allan, 2009]. Néanmoins, nous pouvons introduire le modèle de thème dans une recherche ad-hoc en suivant trois approches principales :

6.1 Modèles de recherche combinés

Parmi les approches du modèle de langage pour l'estimation de la pertinence d'une requête par rapport à un modèle de document, il est possible d'utiliser le modèle de probabilité de requête (voir Chapitre 3 :4.6). Sous l'hypothèse de sac de mots (indépendance des termes), et pour chaque document D[20] et une requête Q composée des termes $\{q\}$, un score de pertinence est estimé par :

$$P(Q|D) = \prod_{q \in Q} P(q|D) \qquad \textbf{(4-8)}$$

Partant du principe de lissage de Dirichlet et intégrant le modèle LDA, une étude proposa de combiner linéairement trois modèles pour estimer la probabilité conditionnelle de générer un terme q sachant le document D [Wei et Croft, 2006] :

1. Le maximum de vraisemblance du terme q dans le document D : $P_{MV}(q|D)$,
2. Le maximum de vraisemblance du terme q dans le corpus C : $P_{MV}(q|C)$,
3. La probabilité du terme q selon les distributions estimées par LDA : $P_{LDA}(q|D)$.

Sachant que N_D est la longueur (en mots) du document D, que μ est le paramètre à priori de lissage de Dirichlet et que λ (entre 0 et 1) est un paramètre d'ajustement de la contribution du modèle LDA, la pertinence d'un terme par rapport à un document est mesurée par :

$$P(q|D) = \lambda\left(\frac{N_D}{N_D+\mu} P_{MV}(q|D) + \frac{\mu}{N_D+\mu} P_{MV}(q|C)\right) + (1-\lambda)\left(P_{LDA}(q|D)\right) \qquad \textbf{(4-9)}$$

Les auteurs dans [Wei et Croft, 2006] confirmèrent avoir obtenu des résultats légèrement meilleurs que celles réalisés par le modèle de requête en choisissant (μ=1000, λ=0,7). Néanmoins, aucune comparaison, incluant d'autres modèles probabilistes ou vectoriels, n'a été appliquée. En testant sur le corpus *AP*[21], les auteurs reportèrent une amélioration de la précision moyenne à 0,2651 par rapport au modèle de requête (0,2179) ou celui à base de cluster (0,2326).

Ce modèle de recherche souffre de certaines limites telles que la complexité de la formule de pertinence et la faible contribution du modèle LDA (1-λ=0,3). Par ailleurs, nos expérimentations sur deux corpus moyens (*CF* et *CACM*) étaient infructueuses et les

[20] Rappelons qu'il s'agit du modèle de document M_D.

[21] Dépêches de presse de la période 1988-90 de l'*Associated Press*

performances obtenues étaient très faibles par rapport à d'autres modèles de références. Par contre, nous avons exploré d'autres alternatives que nous décrivons dans ce qui suit.

6.2 Similarité dans l'espace des thèmes

Du moment où le modèle LDA permet de calculer, dans l'espace des thèmes, la distribution de tout document (même nouveau), nous pouvons donc estimer la distribution à posteriori d'une requête *Q* et mesurer la valeur de pertinence *RSV* de chaque document *D*.

Rappelons que la divergence de *Kullback-Leibler* peut être utilisée pour mesurer la dissemblance entre deux distributions *p* et *q* comme suit [Heinrich, 2008] :

$$D_{KL}(p,q) = \sum_{j=1}^{T} p_j \log_2 \frac{p_j}{q_j} \quad \textbf{(4-10)}$$

Selon l'équation (4-10), la divergence D_{KL} est une fonction non-négative mais asymétrique. Il est possible de dériver une forme plus pratique d'une mesure symétrique :

$$SymKL(p,q) = \frac{1}{2}\left[D_{KL}(p,q) + D_{KL}(q,p)\right] \quad \textbf{(4-11)}$$

C'est la forme symétrique de la divergence de *Kullback-Leibler* qu'on utilise comme mesure de similarité entre les distributions des documents.

Bien que la divergence de *Kullback-Leibler* parait comme la plus appropriée pour mesurer la similarité entre ce genre de distributions, les fonctions de similarité vectorielle, telles que le cosinus, ont produit les meilleures performances [Wei et Croft, 2006]. Nos expérimentations ont confirmé le même constat.

6.3 Extension thématique de la requête

L'approche d'extension de la requête part du fait que le besoin d'information ne peut être satisfait complètement dans un système de recherche à base de mots clés (termes d'une requête). L'objectif consiste à enrichir la requête par de nouveaux termes exprimant le même sujet. Du moment où il permet de regrouper les mots les plus consistants dans le même thème, le modèle LDA paraît comme une alternative astucieuse pour découvrir les termes contextuellement rattachés à une requête donnée.

En considérant les relations associatives entre les termes, on peut mesurer une similarité basée sur les distributions conditionnelles. Pour chaque terme w_i, on exprime le degré de dépendance par rapport au terme *w* comme suit [Steyvers et Griffiths, 2007] :

$$P(w_i|w) = \sum_{j=1}^{T} P(w_i|z=j)P(z=j|w) \quad \textbf{(4-12)}$$

Les termes w'_i maximisant l'expression dans (4-12) interprètent une forte dépendance au terme *w* ($w \rightarrow w'_i$). En fixant un seuil δ de tolérance (pour accepter les termes les plus rattachées), il est possible de définir un cadre pour reconstruire la requête étendue *QE* à partir de la requête de base $Q = \{q\}$.

$$QE_{\delta}^{-} = Q \bigcup_{q \in Q} \left\{ q', P(q'|q) \geq \delta \wedge P(q'|q) \geq P(q|q') \right\} \quad \textbf{(4-13)}$$

$$QE_{\delta}^{+} = Q \bigcup_{q \in Q} \{q', P(q|q') \geq \delta \wedge P(q|q') < P(q'|q)\} \quad \textbf{(4-14)}$$

$$QE_{\delta} = QE_{\delta}^{-} \bigcup QE_{\delta}^{+} \quad \textbf{(4-15)}$$

L'ensemble QE^{-}, dans l'équation (4-13), représente l'extension de la requête Q par les hypo-termes dépendants ($q \rightarrow q'$). Alors que QE^{+}, dans l'équation (4-14), contient la requête étendue par les hyper-termes dont ceux de la requête Q en dépendent ($q' \rightarrow q$).

Afin d'estimer l'effet de l'approche d'extension de la requête, nous avons calculé le nombre total des termes obtenus dans les requêtes de chaque corpus. Sur 704 termes des requêtes *CF* (635 termes dans *CACM*) nous reportons dans **Tableau 4-4** le taux d'extension global ($QE^{-} \cup QE^{+}$) en fonction du nombre de thèmes LDA et du seuil de tolérance δ.

Corpus	**CF**			**CACM**		
δ	**10%**	**20%**	**30%**	**10%**	**20%**	**30%**
LDA-100	1 566,3%	392,9%	130,3%	1 048,5%	155,1%	100,9%
LDA-200	1 596,6%	181,4%	120,3%	925,0%	254,5%	123,0%
LDA-300	1 164,1%	209,9%	110,8%	906,9	195,0%	104,7%

Tableau 4-4. Taux d'extension des requêtes dans les corpus *CF* et *CACM*.

Corpus	**Terme**	**Extension**
CF	Activity	cytoplasmic, esteras, extend, enzyme, chromatographic, proteolytic, arginine
	Bronchial	Asthmatic, aerosol, tracheostomy, asthma, endotrach, bronchoscopic
	Effect	serial, beneficial, dialyze, extraction, days, replic, adverse, intralipid, student
	Mucus	hypersecretion, variance, nonspecific, shed, responsive, production, viscid, technician
	Respiratory	pneumonia, colonization, escherichia, harbour, enterobacteriaceae, predominate, tract
	Role	issue, prospect, play
	Test	reliable, equipment, false, negative, bronchoconstrict, maximum, perform, northern, procedure, positive, strong, reasonably, reliably, collect
CACM	Approximation	cubic, spline, piecewis, quintic, ln, mesh, horowitz, minimax, periodic, ki, rational, reme, extrapolation
	Bit	register, shift
	Compiler	meta, translator, gri, assembler, compilation, compile, neliac, incremental
	Language	ability, familiar, assemble, advancement, macro, translation, deficiency, applicative, nonprogramm, calculus, meta, imperative, orient, grasp, assembly, crespi, lambda, semantics
	Mathematical	partly, apt, trigonometric, photocomposit, mathematics, pi, rodriguez, font, typeset
	Matrix	qr, column, band, row, eigen, multiplication, tridiagon, perturbation, inverse, latent, ergodic, subchain, reversible, diagonal, pentadiagon, pei, vector, eigenvalue, pencil, submatrix, f, eigenvector, symmetric, product, inversion
	Pattern	induce, chemical, shrink, waveform, electrocardiogram, match, carotid, arterial, peak, higgin, stark, pulse, recognition, structural, vault

Tableau 4-5. Exemples d'extension des termes dans les corpus *CF* et *CACM* (T=300, δ=10%).

La variation des taux d'extension et inversement proportionnelle au seuil de tolérance dans le même modèle de thème. Néanmoins, le seuil δ doit être soigneusement choisi afin d'obtenir une extension raisonnable pour chaque modèle LDA. Notre investigation a exploré une large gamme de modèles (T = 8, 16, 32, 64, 100, 200, 300, 400, 500, 600, 700) avec un seuil δ = 5%, 10%, 20%, 30%, 50%).

En plus, nous dressons dans **Tableau 4-5** l'extension de certains termes des requêtes dans les deux corpus selon modèle LDA à 300 thèmes latents et avec un seuil de tolérance δ=10%:

Nous proposons d'utiliser cette approche d'extension de requête avant d'appliquer un modèle de recherche dans l'espace de mots tel que *TF*, *IDF* ou *Okapi*.

6.4 Expérimentations

Nous appliquons dans cette section plusieurs approches d'intégration du modèle LDA dans la recherche ad-hoc. Pour cet effet, nous utilisons deux corpus de référence : le premier est le corpus *CF* déjà décrit dans la section 4.4. Le deuxième est le corpus *CACM* contenant 3.204 résumés des articles publiés dans le journal *ACM*[22] entre 1958 et 1979.

6.4.1 Configuration et prétraitement

Le corpus *CF* contient 100 requêtes avec des jugements provenant de quatre sources différentes. Chaque jugement attribue un score de pertinence pour chaque document dont l'interprétation est comme suit (2 : pertinent, 1 : marginalement pertinent, 0 : non pertinent). Pour notre évaluation, nous avons considéré comme pertinent à une requête tout document totalisant un score minimal de 4/8. En plus, nous avons ignoré toute requête ayant un seul document pertinent. Cette configuration nous a ramené à évaluer 97 requêtes.

Le corpus *CACM* inclut 63 requêtes dont 51 possèdent des jugements de référence (pertinent/non-pertinent). En ignorant celles qui donnent un seul document pertinent, nous nous sommes limités à 49 requêtes pour les évaluations.

Comme pour les tests de catégorisation (voir 5.2) nous avons appliqué deux types de prétraitement linguistique sur les deux corpus (documents et requêtes) : le stemming léger et la lemmatisation. En plus, nous avons mené les mêmes évaluations dans le texte brut sans prétraitement préalable.

Pour la comparaison des performances de recherche, nous considérons deux groupes de modèles selon l'approche de pondération des termes, absolue ou relative. Pour la pondération absolue, nous appliquons deux modèles (***Bool*** et ***TF***) avec lesquels nous comparons notre proposition d'extension de la requête ***LDA-QE-TF*** avec une pondération ***TF*** :

1. ***Bool*** : pour le modèle booléen de base où la combinaison des termes de la requête et réalisée par l'opérateur "OU".
2. ***TF*** : pour le modèle vectoriel où les termes sont pondérés par leur fréquence dans le document.
3. ***LDA-QE-TF*** : pour le même modèle ***TF*** avec une augmentation de la requête par les termes les plus dépendants dans le modèle de thèmes LDA (voir 6.3).

Dans le deuxième groupe (pondération relative), nous considérons deux modèles de référence (***TF-IDF*** et ***Okapi***) avec lesquels nous comparons le modèle d'extension de la

[22] Association of Computing Machinery

requête avec une pondération ***IDF*** des termes de recherche (***LDA-QE-IDF***). En plus, nous évaluons la recherche dans l'espace de thèmes ***LDA-TS*** :

4. ***Okapi*** : pour le modèle probabiliste Okapi-BM25 selon la formule (3-9) en fixant (k_1=1.2, b=0.75, k_3=1000)
5. ***LDA-TS*** : pour le modèle de recherche dans l'espace des thèmes. Les résultats présentés sont celles obtenues par la mesure du cosinus puisque les performances de recherche obtenues par la divergence de *Kullback-Leibler* étaient moins bonnes (voir 6.3).
6. ***TF-IDF*** : pour le modèle vectoriel où les termes sont pondérés par leur fréquence combinée avec la fréquence inversée dans la collection (voir Chapitre 3 :3.2).
7. ***LDA-QE-IDF*** : pour le modèle de pondération ***IDF*** des termes de requêtes augmentés par les termes les plus dépendants dans le modèle de thèmes LDA (voir 6.3).

Il faut noter qu'avec un nombre de thèmes entre 100 et 500, nous avons obtenu les meilleures performances de recherche dans les modèles intégrant LDA (***LDA-QE-TF***, ***LDA-TS*** et ***LDA-QE-IDF***). Pour le schéma d'extension de requête, les meilleurs résultats ont été obtenus en choisissant un seuil de tolérance δ entre 10% et 30% (voir l'équation (4-15)).

6.4.2 Résultats

Pour l'évaluation des performances de la recherche ad-hoc ordonnée, nous calculons les précisions obtenues pour les dix et les 100 premiers documents retrouvés (**Pre-10** et **Pre-100**). En plus, nous reportons la précision moyenne (**Pre-Moy**) et nous traçons la courbe de la précision interpolée à 11 points. Les résultats listés dans cette section concernent la moyenne globale relative à l'évaluation de toutes les requêtes du corpus (97 pour *CF* et 49 pour *CACM*).

Nature du mot	Modèle	Pre-10	Rap-10	Pre-100	Rap-100	Pre-Moy
Brut	**Bool**	0,017	0,011	0,024	0,141	0,023
	TF	0,152	0,115	0,053	0,363	0,116
	LDA-QE-TF	0,157	0,119	0,054	0,369	0,120
	Okapi	0,158	0,159	0,054	0,385	0,143
	LDA-TS	0,269	0,238	0,074	0,512	0,228
	TF-IDF	0,277	0,263	0,078	0,520	0,261
	LDA-QE-IDF	0,280	0,257	0,078	0,523	0,256
Stem	**Bool**	0,014	0,010	0,023	0,131	0,023
	TF	0,168	0,129	0,060	0,403	0,129
	LDA-QE-TF	0,184	0,141	0,062	0,415	0,141
	Okapi	0,186	0,182	0,058	0,402	0,161
	LDA-TS	0,274	0,240	0,077	0,507	0,250
	TF-IDF	0,313	0,288	0,087	0,564	0,283
	LDA-QE-IDF	0,301	0,267	0,087	0,563	0,280
Lemme	**Bool**	0,016	0,010	0,023	0,133	0,023
	TF	0,170	0,137	0,059	0,396	0,130
	LDA-QE-TF	0,190	0,158	0,063	0,443	0,161
	Okapi	0,175	0,176	0,057	0,405	0,158
	LDA-TS	0,279	0,246	0,076	0,510	0,247
	TF-IDF	0,301	0,290	0,085	0,555	0,282
	LDA-QE-IDF	0,303	0,275	0,086	0,559	0,284

Tableau 4-6. Evaluation de la recherche dans le corpus CF.

Pour le corpus *CF*, nous résumons dans **Tableau 4-6** l'évaluation des 7 modèles de recherche appliqués sur trois types d'unités lexicales (mot brut, stem et lemme). Tout d'abord, les résultats montrent une performance légèrement supérieure des modèles de recherche avec un prétraitement de lemmatisation. Nous nous contenons, dans ce qui suit, de reporter les évaluations dans le texte lemmatisé seulement. Nous revenons sur l'analyse de l'impact et l'utilité des prétraitements linguistiques dans le chapitre suivant.

En plus, l'analyse de la précision interpolée dans **Figure 4-6**, montre bien que l'approche d'extension thématique des requêtes permet d'améliorer la performance de recherche que ce soit avec une pondération absolue (*TF*) ou avec une pondération relative (*IDF*). Par ailleurs, le modèle de recherche dans l'espace des thèmes apporte une nette amélioration par rapport aux modèles (booléen, vectoriel TF et Okapi).

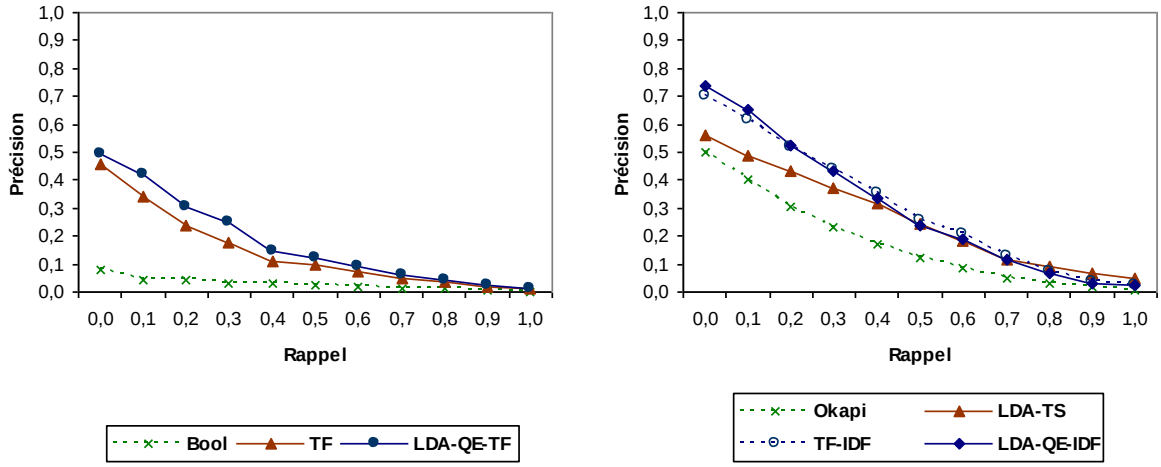


Figure 4-6. Précision interpolée de la recherche dans le corpus CF (lemme).

Les mêmes constatations ont été confirmées lors de l'évaluation des modèles de recherche dans le corpus *CACM* (voir **Tableau 4-7** et **Figure 4-7**).

Modèle	Pre-10	Rap-10	Pre-100	Rap-100	Pre-Moy
Bool	0,002	0,001	0,006	0,041	0,009
TF	0,280	0,215	0,075	0,541	0,204
LDA-QE-TF	0,259	0,208	0,074	0,529	0,206
Okapi	0,282	0,267	0,090	0,631	0,267
LDA-TS	0,300	0,220	0,101	0,624	0,255
TF-IDF	0,351	0,320	0,100	0,685	0,331
LDA-QE-IDF	0,400	0,352	0,101	0,701	0,339

Tableau 4-7. Evaluation de la recherche dans le corpus CACM (Lemme).

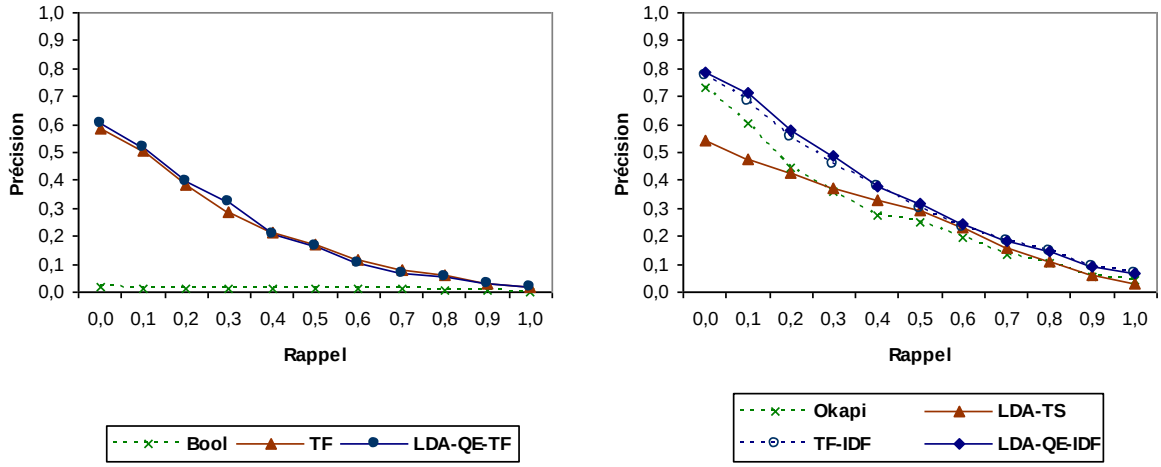


Figure 4-7. Précision interpolée de la recherche dans le corpus CACM (lemme).

7 Evaluation du modèle LDA

Une approche évidente pour évaluer le modèle de thème LDA est de comparer la performance de son utilisation dans les tâches suivantes en RI (recherche ad-hoc, classification, agrégation, …etc.). Cependant, la modélisation par thème peut être appréciée sur plusieurs autres niveaux tels que la perplexité, la stabilité des thèmes et la similarité des documents et des mots [Blei et al., 2003] [Steyvers et Griffiths, 2007] [Heinrich, 2008].

7.1 Détermination du nombre de thèmes

Le nombre de thèmes (T), qui est fixé arbitrairement par l'utilisateur, peut affecter directement l'interprétabilité des résultats. L'apprentissage du modèle LDA avec un nombre (T) trop faible produit génralement des thèmes vagues. Cependant, un mmodèle avec un très grand nombre de thèmes conduit à un modèle non-interprétable [Steyvers et Griffiths, 2007]. Plusieurs thechniques ont été proposées pour choisir un nombre convenable de thèmes pour l'apprentissage du modèle LDA. Une manière expérimentale consiste à selectionner les nombre (T) dont le modèle résultant donne les meilleures performances dans les application finale de recherche, classification, …etc.

7.2 Perplexité

Un modèle de thème inférré pour un sous-ensemble de documents doit être capable de prédire (générer) le choix des mots dans le reste de la collection. La mesure de perplexité a été proposée pour évaluer la capacité de généralisation d'un modèle de langage pour des textes non observés [Azzopardi et al., 2003]. Elle est définie comme la moyenne géométrique inversée de la probabilité d'une collection de test. La perplexité d'un modèle de langage (distribution de probabilités) relatif à l'ensemble de test D_{test} comprenant K documents est formulée comme suit [Blei et al., 2003] :

$$Perplexity(D_{test}) = \exp\left(-\frac{\sum_{d=1}^{K}\log p(d)}{\sum_{d=1}^{K} N_d}\right) \quad \textbf{(4-16)}$$

Où N_d est le nombre de mots de chaque document d dans la collection de test.

Un score faible de la perplexité indique une meilleure performance de généralisation. Cette mesure a été appliquée sur 10% de deux collections (résumés scientifiques de *C.Elegans*

et les dépêches de presse de *TREC-AP*) pour comparer quatre modèles (uni-grammes, mélange d'uni-gramme, pLSI et LDA). En variant les thèmes entre 2 et 200, les résultats ont montré une meilleure performance du modèle LDA [Blei et al., 2003].

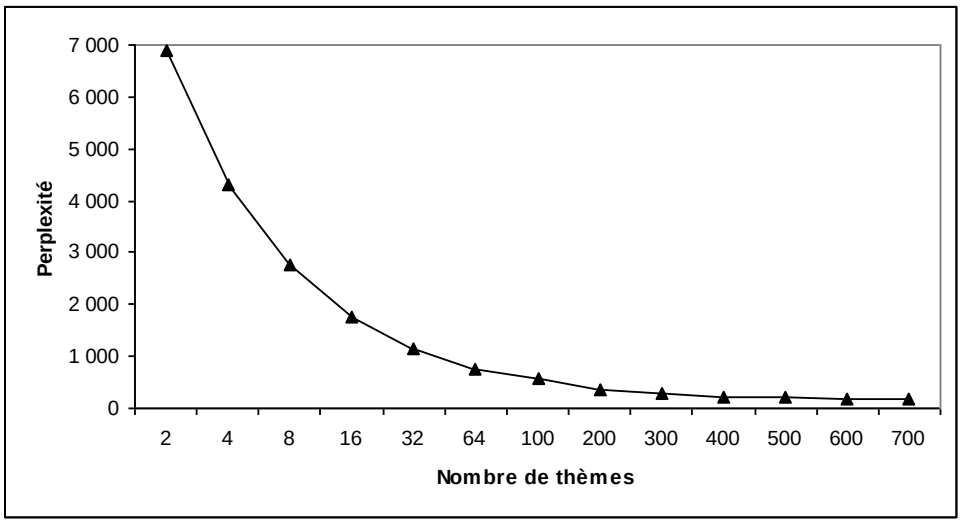


Figure 4-8. Perplexité du modèle LDA en fonction du nombre de thèmes *T*.

A titre d'évaluation préliminaire, nous calculons la perplexité du modèle LDA inféré dans une collection réelle d'articles de presse[23] où ¼ des documents, soit K=1000, a été réservé pour l'ensemble test (D_{test}). En variant T de 2 à 700, la Figure 4-8 montre que la capacité de généralisation du modèle LDA s'améliore proportionnellement avec le nombre de thèmes. Cette constatation est reste la même avec d'autres collections testées dans ce travail. Néanmoins, nous verrons dans les sections suivantes d'autres approches d'évaluation du modèle LDA qui vont relativiser ce constat.

7.3 Stabilité des thèmes

Puisque les thèmes correspondent à des distributions de probabilités sur les mots, la stabilité des thèmes dans un modèle peut être évaluée par le calcul de divergence des distributions inférées dans plusieurs exécutions de l'apprentissage du modèle LDA [Steyvers et Griffiths, 2007].

Nous utilisons dans ce qui suit la forme symétrique de la divergence de *Kullback-Leibler* comme définie dans (4-11). Si on considère deux modèles M_{T1} et M_{T2} résultants de deux inférences sur la même collection et pour un nombre T de thèmes, on peut calculer donc la dissemblance pour chaque pair de thèmes (θ_{1i}, θ_{2j}), le *i-ème* (resp. *j-ème*) thème dans la première (resp. deuxième) inférence. Du moment où les thèmes sont interchangeables, deux thèmes des deux modèles M_{T1} et M_{T2} sont supposés équivalents s'ils sont les moins divergents au sens de la fonction (4-11). Nous réordonnons les thèmes dans le deuxième modèle selon cette équivalence afin de pouvoir utiliser l'indice i de façon uniforme dans les deux modèles M_{T1} et M_{T2}.

Ainsi, nous pouvons définir un facteur d'instabilité du modèle en calculant la moyenne de dissemblance des thèmes dans les deux modèles M_{T1} et M_{T2} selon :

[23] La collection *Ech-4000* sera décrite ultérieurement dans Chapitre 6 :2.1.

$$S_T = S_T(M_1, M_2) = \frac{1}{T}\sum_{i=1}^{T} SymKL(\theta_{1i}, \theta_{2i}) \quad \textbf{(4-17)}$$

Dans une expérimentation préliminaire, nous avons calculé ce facteur S_T sur les thèmes latents inférés de la collection *Ech-4000*. Pour une série d'apprentissage (T=2…700), nous présentant dans Figure 4-9 la courbe du facteur d'instabilité du modèle LDA.

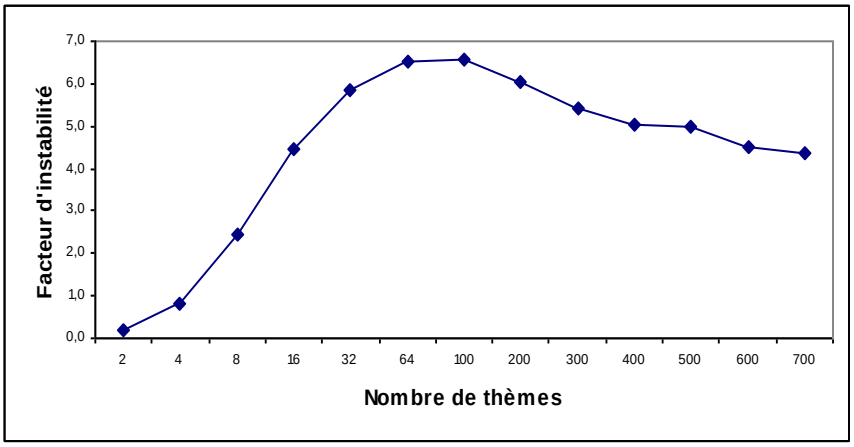


Figure 4-9. Facteur d'instabilité (S_T) du modèle LDA en fonction du nombre de thèmes T.

Le graphe dans Figure 4-9 montre que la divergence des inférences du modèle est minime pour un nombre faible de thèmes. Ce qui peut signifier que l'apprentissage LDA est moins stable lorsque le nombre de thèmes augmente. Néanmoins, ce constat, qui est recensé même pour d'autres collections, paraît contradictoire avec celui retenu lors de l'évaluation de la perplexité. Nous avons déjà mentionné précédemment que la capacité de généralisation d'un modèle LDA est meilleure en augmentant le nombre de thèmes à apprendre. Il est donc nécessaire de penser à un autre critère afin de sélectionner le nombre de thèmes adéquat pour l'apprentissage LDA dans une collection donnée. Nous proposons, dans ce qui suit, une approche empirique pour trouver un compromis judicieux face à ce problème.

7.4 Mesure combinée à base de Kullback-Leibler

Afin de trouver le nombre de thèmes (T) approprié, deux critères peuvent être considérés sur les distributions de thèmes lors de l'apprentissage LDA dans une collection donnée :

- Etre stables sur plusieurs inférences,
- Etre distinctes dans le même modèle.

Notre idée part du fait que les thèmes inférés doivent être, autant que possible, stables (S_T faible) et distincts à la fois. Pour ceci nous proposons de calculer, selon l'équation (4-11) et dans le même modèle, la divergence symétrique de *Kullback-Leibler* pour chaque paire de thèmes (θ_i, θ_j), où (θ_i) dénote la distribution du *i-ème* thème. Dans la même inférence, nous calculons la divergence minimale D_T^{Min} et la moyenne D_T^{Avg} de divergence des thèmes comme suit :

$$D_T^{Min} = \min_{i=1..T, j=1..T}^{i \neq j} \{SymKL(\theta_i, \theta_j)\} \quad \textbf{(4-18)}$$

$$D_T^{Avg} = \frac{2}{T(T-1)} \sum_{i=1}^{T-1} \sum_{j=i+1}^{T} SymKL(\theta_i, \theta_j) \quad \textbf{(4-19)}$$

Du moment où nous supposons que les modèles inférés peuvent être instables, nous estimons les quantités D_T^{Min} et D_T^{Avg} en calculant une moyenne de plusieurs exécutions (inférences).

Ainsi, nous pouvons définir une mesure empirique combinant le facteur d'instabilité (4-17) avec les divergences, minimale (4-18) et moyenne (4-19), dans un apprentissage du modèle LDA pour un nombre de thèmes T donné :

$$BKL_T = D_T^{Avg} \left(\frac{D_T^{Min}}{S_T} \right) \quad \textbf{(4-20)}$$

Cette mesure, que nous dénotons par *BKL* (*Brahmi-KL*), peut être justifiée par l'ambition d'avoir un modèle ayant la meilleure stabilité par rapport au niveau minimal de discrimination entre les thèmes (D_T^{Min}). Cette mesure est amplifiée par la moyenne de divergence des thèmes (D_T^{Avg}) afin de favoriser les modèles dont les thèmes sont les moins confus.

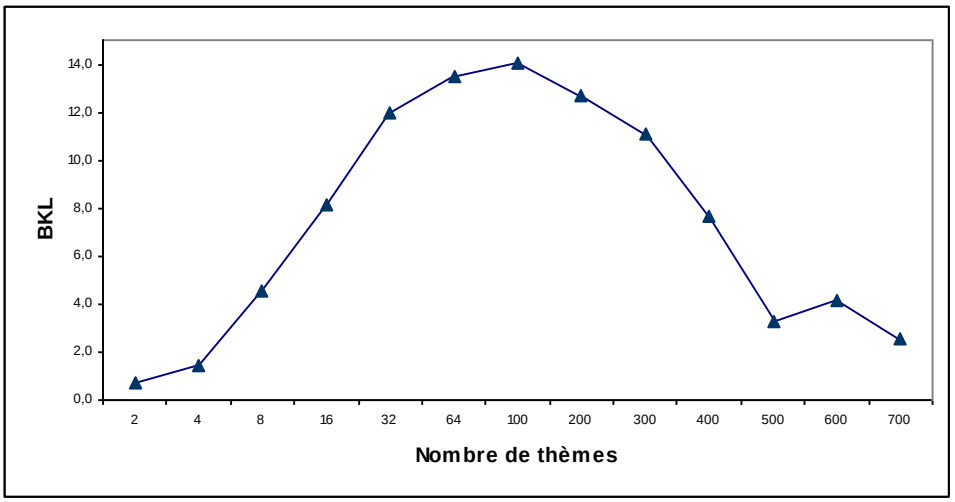


Figure 4-10. Mesure BKL du modèle LDA en fonction du nombre de thèmes.

Dans la Figure 4-10, nous présentons une évaluation préliminaire de la mesure *BKL* selon (4-20) en inférant LDA pour différent nombre de thèmes dans la collection *Ech-4000*. Le fait d'avoir le meilleur score *BKL* pour T=100 signifie que le modèle LDA_{100} était le plus stable et le moins confus. Mais surtout, que tout choix de 64<T<200 peut conduire à un apprentissage efficace du modèle LDA dans la collection *Ech-4000*. Nous verrons dans les chapitres suivants les résultats d'autres évaluations qui confirmeront ce constat.

7.5 Catégorisation dans l'espace des thèmes

Une autre méthode d'évaluation du modèle LDA consiste à mesurer la performance de catégorisation des documents représentés dans l'espace de thèmes. Après l'apprentissage du modèle LDA pour un nombre de thèmes variant entre 2 et 700, nous appliquons l'algorithme SVM avec une adaptation des mesures d'évaluation dans l'implémentation utilisée de

LIBSVM[24]. 3000 documents de la collection *Ech-4000* sont utilisés pour l'apprentissage des 8 catégories alors que les 1000 documents restants sont réservés pour le test. Nous évaluons la classification SVM par calcul de la macro-moyenne de précision et du rappel ; la *F-mesure* globale est estimée équitablement par F_1-score selon les équations (2-4) et (2-5) (voir Chapitre 2 :7). En plus, le ratio des vecteurs de support par rapport au nombre des exemples d'apprentissage (*rSV*) et donné comme indication de la capacité de généralisation (voir Chapitre 2 :5).

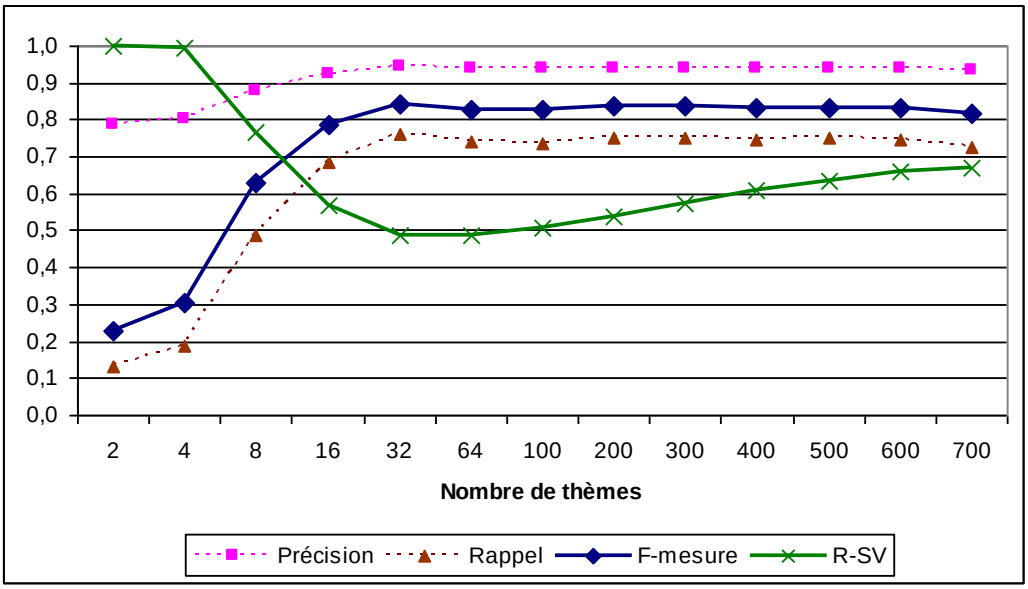


Figure 4-11. Performances de la catégorisation par SVM dans *Ech-4000*.

Les résultats d'évaluation préliminaire de la Figure 4-11 montrent que la classification, dans l'espace des thèmes de la collection *Ech-4000*, est plus performante (en termes de *F-Mesure*) pour $T \geq 32$. Cependant, la mesure R-SV devient relativement importante à partir de $T \geq 200$ (comme d'ailleurs pour $T < 32$). Ce qui peut réduire la capacité de généralisation du classifieur pour un nombre de thèmes assez important. En combinant les constatations sur la *F-Mesure* et le *rSV*, on peut conclure que l'apprentissage *SVM* est plus performant dans un espace de thèmes lorsque $32 \leq T \leq 100$. Nous examinons, dans les chapitres suivants, ces résultats par expérimentation du modèle LDA sur d'autres collections.

8 Conclusion

La modélisation par thème permet de prendre en charge certains aspects sémantiques du texte. Le modèle LSI (vu dans le chapitre précédent) tente de réaliser un objectif similaire en opérant une transformation matricielle par la technique de décomposition en valeurs singulières. La réduction de l'espace de description des documents constitue une forte motivation pour ce type de modèle. L'ambition de lever les problèmes de synonymie et de polysémie reste toujours acclamé par les partisans de l'approche vectorielle [Deerwester et al., 1990] [Dumais, 1993]. Néanmoins, cette approche manque de fondement théorique solide,

Partant d'une approche probabiliste, plus solide, le modèle de thème d'allocation latente de Dirichlet (LDA) propose de considérer une distribution multinomiale des thèmes sur les

[24] LIBSVM-2.89 est librement disponible à http://www.csie.ntu.edu.tw/~cjlin/libsvm+zip

mots. Il définit un cadre statistique élégant pour estimer la distribution des documents sur ces variables latentes apprises par le modèle (thèmes). A la différence du modèle pLSI, même les nouveaux documents non observés lors de l'apprentissage, peuvent être estimés.

L'espace des thèmes, généré par l'apprentissage LDA, a été utilisé efficacement dans les tâches de catégorisation multi-classe et de recherche ad-hoc. En plus, nous avons proposé un cadre théorique pour l'extension thématique des requêtes de recherche. L'analyse comparative, de plusieurs modèles sur deux corpus de références, nous a permis de prouver l'efficacité d'intégrer le modèle LDA dans la recherche ad-hoc.

Nous avons décrit les principales mesures d'évaluation du modèle LDA mais surtout, nous avons proposé une mesure empirique (*BKL*) pour déterminer le nombre de thèmes, donné en entrée, adéquat à l'apprentissage d'un modèle de thèmes LDA dont les thèmes soient stables et discriminants.

Bien que secondaire, l'étiquetage manuel des thèmes permet d'avoir une lecture sémantique du contenu de la collection en considération. La détection et l'association des mots les plus (lourds) dans un thème tracent des contours de contexte (ou de concepts) sur les mots. Ceci nécessite, en premier lieu, d'apprendre le modèle LDA sur une collection de "sac de mots". Encore, faut-il que ces mots soient représentatifs, distinctifs et non redondants par cause d'une flexion de genre, de nombre ou de temps. Ces problèmes liés aux aspects linguistiques doivent être donc préalablement résolus. Ceci fera l'objet du chapitre suivant portant sur le prétraitement linguistique.

Chapitre 5 :
TRAITEMENT LINGUISTIQUE EN RECHERCHE D'INFORMATION

1 Introduction

Nous venons d'explorer, dans les deux premiers chapitres, les fondements théoriques de la recherche d'information et les principaux modèles qui ont été proposés, depuis les années 1960 jusqu'à l'aire du Web mondial, pour représenter, indexer et retrouver les documents textuel. L'hypothèse du "sac de mots" commune dans ces modèles sous-entend qu'un document est caractérisé essentiellement par la présence (et parfois la fréquence) de ces mots. Aucune considération d'ordre sur les mots n'est prise en charge (sauf pour les modèles *k*-gramme *k*>1). La détection des mots peut être réalisée par le découpage de texte selon la ponctuation (marques de séparation) en vigueur. Cette phase de "tokenization" est essentielle pour détecter les graphèmes "mots" qui seront utilisés dans l'indexation du document.

La pertinence est estimée dans un processus d'appariement avec la requête reposant en premier lieu sur la correspondance des termes (dans ce cas les graphèmes[25] du texte brut). Même en supposant que la requête exprime fidèlement le besoin d'information de l'utilisateur, les systèmes de recherche d'information se trouvent confrontés à deux problèmes majeurs :

Le premier concerne la formulation (morphologiquement ou sémantiquement) diversifiée du même concept (ou du même lemme) ; ainsi, un document pertinent peut contenir des

[25] mots délimités par des ponctuations ou toute marque de séparation d'un texte.

termes sémantiquement proches de la requête (dérivation morphologique, synonymie, hyperonymie, …etc.). Le SRI ne peut pas donc proposer de tels documents ce qui conduit à une baisse du rappel.

Le deuxième concerne la polysémie des mots où le même terme peut exprimer plusieurs sens. Cette ambigüité engendre une baisse de précision puisqu'elle aboutisse à la sélection de certains documents non pertinents.

Ces problèmes sont dus essentiellement à l'ignorance des aspects linguistiques dans l'analyse et la représentation des documents et des requêtes. Pourtant il est évident que l'aspect approximatif dans les modèles de RI revient en partie à la complexité du langage naturel. En l'absence de toute information organisationnelle dans les documents non-structurés, il est donc impératif d'introduire des techniques de traitement automatique de la langue dans les SRI. Nous présentons dans ce chapitre les principales approches de traitement automatique du texte en vue d'une indexation dans un SRI.

2 Traitement automatique du langage naturel

Dès les années 1950, alors que l'informatique n'en était qu'à ses débuts, est apparue l'idée d'utiliser les ordinateurs pour traduire des textes d'une langue à l'autre. Puis assez rapidement, ce champ d'exploitation des langues s'est diversifié et peu à peu étendu à la vérification et à la correction orthographique, à l'indexation et au résumé de textes, …etc.

Ainsi deux disciplines à priori très différentes, l'informatique et la linguistique, se sont confrontées et rapprochées pour donner une naissance à une informatique linguistique ou une linguistique informatique. Les finalités de ce nouveau champ disciplinaire, le traitement automatique du langage naturel (TALN) (*en angl. Natural language processing NLP*), sont multiples : Tout d'abord, elles correspondent à la mise en place d'applications concrètes (indexation et accès à l'information, résumé de textes, traduction assistée par ordinateur, dialogue home-machine par exemple). Mais elles permettent aussi de confronter la linguistique, qui pendant longtemps demeura descriptive, à des exigences d'opérationnalité et de la modélisation informatique. Enfin, l'informatique linguistique est l'une des domaines privilégiés de l'informatique et de l'intelligence artificielle. Le langage, sa structure et son usage, étant parmi les objets naturels de modélisation informatique les plus intéressants.

Deux objectifs principaux ambitionnent le développement des outils de TALN : D'une part, concevoir des interfaces de plus en plus intuitives et ergonomiques. D'autre part, traiter (produire, lire, rechercher, classer, analyser, traduire) de manière plus "intelligente" les informations disponibles sous forme textuelle.

Notons ici que les applications de recherche d'information ne sont pas assez exigeantes en matière de traitement linguistique. Certains SRI ignoraient (ou presque) les aspects linguistiques dans le processus d'indexation[26]. Néanmoins, nous verrons que l'intégration des outils du TALN peut améliorer la performance de certaines tâches en RI. En particulier, deux domaines, qui sont en pleine expansion, sollicitent un prétraitement efficace et spécifique de la langue naturelle :

- La recherche intelligente basée sur la sémantique,
- La recherche d'information dans le texte non-anglais.

[26] Google introduisa partiellemnt les techniques de stemming vers la fin de l'année 2003.

Sans pour autant décrire en détail les principes et les méthodes du TALN, nous nous contentons de citer, dans ce qui suit, ceux qui peuvent intervenir dans l'analyse textuelle en vue d'une indexation dans un SRI.

3 Indexation et traitement linguistique

En recherche d'information, le processus d'indexation consiste à analyser chaque texte du fond documentaire afin de produire une liste de mots clés (descripteurs) qui seront utilisés ultérieurement dans le processus de recherche [Salton, 1971] [Rijsbergen, 1979]. Les mots d'indexation doivent être à la fois :

- Descriptifs : c.à.d. représentatifs du contenu du document,
- Discriminants : c.à.d. qu'ils doivent mettre en évidence ce qui différencie un document parmi tous les autres de la collection.

3.1 Objectifs du traitement linguistique

Du point de vue linguistique, l'indexation doit traduire le contenu conceptuel du texte. Elle permet de donner une représentation mieux structurée et si possible normalisée du contenu des textes avec plus ou moins de traitement. La première difficulté dans le processus d'indexation est alors de résoudre les problèmes linguistiques les plus visibles et fortement liés à la nature de la langue elle même :

- Une variation typographique propre à la langue peut nuire la correspondance des mots (majuscule/minuscule, accentuation, diacritiques, …etc.)
- Un mot à usage fréquent (mot vide) n'est généralement pas discriminant (articles, préposition, auxiliaires, …etc.).
- Une même entrée du dictionnaire (lemme ou racine) peut avoir plusieurs formes ; par exemple : travail, travaille, travaillent, …etc.
- Un même concept peut être exprimé par des mots différents (synonymie) ; par exemple : bicyclette, vélo, vélocipède.
- Un même mot peut représenter plusieurs concepts (polysémie) ; par exemple : le mot arbre n'aura pas la même signification en agriculture, en informatique ou en mécanique.
- Un concept peut être exprimé par mots composés ; par exemple : lave-vaisselle, Moyen-Orient, …etc.

Le prétraitement linguistique en RI s'intéresse principalement aux trois premières difficultés. Les trois derniers peuvent être pris en charge soit en se dotant de ressources linguistiques externes (thésaurus), soit automatiquement mais partiellement par des méthodes statistiques indirectes. Les modèles de thème, comme les modèles de langage, représentent une bonne alternative que nous analysons dans le présent travail.

3.2 Variation typographique et normalisation

Le problème de variation typographique peut être pris en charge dans une phase préliminaire de normalisation. Par exemple, mettre tous les mots en minuscule et supprimer les accents va pouvoir lever la sensibilité aux modes variés d'écriture dans les documents ou les requêtes utilisateurs.

La complexité de normalisation dépend de la langue elle-même. Alors que cette étape est assez "légère" dans le texte anglais et un peu moins dans les autres langues indoeuropéennes, elle peut être plus problématique dans les langues sémitiques telles que l'arabe. Nous verrons plus loin quelques aspects relatifs au prétraitement du texte arabe dans un processus d'indexation.

3.3 Mots vides et réduction de l'index

Nous avons évoqué en (Chapitre 3 :3.3) que la loi de *Zipf* nous indique qu'une bonne partie des textes est constituée des mots vides. L'observation fondamentale de *Zipf* réclamait que les mots utiles, dans une collection réelle de textes, sont rares et longs et que les mots vides sont plus fréquents et plus courts [Zipf, 1949].

Un "mot vide" est un mot qui ne doit pas être indexé. Ceci peut économiser jusqu'à 50% de la taille d'index et par conséquent, réduire la dimension d'un modèle de représentation de documents. Il est généralement admis que ces mots très fréquents ne sont pas à indexer, car ils ne sont pas informatifs, et ils augmentent énormément la taille de l'index relatif à un fond documentaire, ce qui est le cas courant. Néanmoins, certaines études prônent l'indexation des mots vides, comme pouvant être informatifs [Riloff, 1995]. Par exemple, si quelqu'un veut s'informer sur la fameuse expression de Hamlet "*to be or not to be*", il ne trouvera rien car tout simplement l'expression n'est constituée que des mots vides hors index.

En RI, il convient comme pré-indexation de supprimer les mots vides en utilisant un anti-dictionnaire (stop-list) préalablement construit. Il est évident que chaque langue dispose de son propre anti-dictionnaire. De plus, faut-il l'adapter à la nature de la collection elle-même et le domaine traité. L'analyse des situations dans lesquelles les mots vides influencent la performance d'un SRI reste toujours un sujet d'actualité où on cherche toujours l'impact des mots vide dans la description des documents [Dolamic et Savoy, 2010].

3.4 Notions de morphologie

La morphologie est une branche de la linguistique qui s'intéresse à la structure des mots [Polguère, 2003]. Cette structure est généralement construite à partir de morphèmes par flexionou par dérivation.

La flexion étudie les mécanismes d'affixation propres à chaque langue. La flexion modifie la catégorie grammaticale (genre, nombre, temps, …etc.) tout en restant dans le même lexème. Par exemple : lec*teur*, lec*trice* ; révis*erai*, révis*erions*.

La dérivation, quant à elle, apporte un changement au niveau sémantique et crée un nouveau lexème. Par exemple : lire, lecteur, lisible, *il*lisible ; atlantique, *trans*atlantique.

Le lemme est une forme canonique du mot qui représente une entrée du dictionnaire. En morphologie flexionnelle, le lemme constitue la forme unique qu'on peut obtenir par suppression des flexions possibles. Ce processus d'unification des formes flexionnelles est appelé "lemmatisation".

La racine est l'élément de base, irréductible, commun à toutes les représentations d'une même famille des mots relatifs à une langue ou famille de langues. Généralement verbale, la racine est une forme abstraite indiquant un champ sémantique et qui peut dériver par affixation sur différents lemmes. La racinisation est le processus qui consiste à normaliser des formes autour de leur racine.

Le stem (ou pseudo-racine) est une unité lexicale qui se rapproche de la notion de racine sans pour autant avoir une origine correcte. La pseudo-racinisation (ou stemming tel qu'il est utilisé en anglais) est un processus lexicale de "dés-affixation" qui consiste à supprimer, selon

une liste prédéfinie et un algorithme de priorité, les affixes les plus communs dans une langue donnée. Le résultat du stemming n'est pas nécessairement une racine mais peut avoir une forme rapprochée.

4 Approches de stemming

De la section précédente on peut conclure qu'un mot peut avoir plusieurs formes ayant un sens proche et dont la nuance n'est pas assez importante en RI. Il convient, ainsi, de représenter les diverses formes du mot par une seule entrée d'index. Nous abordons dans cette section les approches de prétraitement du texte en vue d'une indexation dans un SRI.

Les termes racinisation, segmentation (ou troncature), pseudo-racinisation (ou stemming) sont utilisés de façon différente, et parfois confuse, pour interpréter l'étape de prétraitement linguistique dans le processus d'indexation des textes. Le mot anglais *stemming* réfère à la procédure qui extrait d'un mot son stem (ou pseudo-racine). En français, il est généralement traduit par *racinisation* ou *radicalisation*[27].

Par ailleurs, la lemmatisation dont l'objectif diffère de la racinisation peut être utilisée comme prétraitement dans le même contexte. De plus, les traitements des variations flexionnelles (lemmatisation) et des variations dérivationnelles (racinisation) sont souvent combinés dans un seul processus. Par conséquent, nous préférons ici de dénommer ce prétraitement linguistique par le ***terme*** "***stemming***". Nous précisons le concept lorsqu'il s'agit de trouver :

- un lemme, par ***lemmatisation***,
- une racine, par ***racinisation***,
- un stem, par pseudo-racinisation ou stemming léger.

Deux objectifs majeurs peuvent être fixés pour la procédure de stemming dans un SRI :

- Unifier les mots ayant le même sens (ou un sens proche),
- Construire un index discriminant et réduit.

Le processus d'indexation qui suivra doit permettre une représentation efficace des documents afin de pouvoir les retrouver ou les classer de façon optimale.

4.1 Analyse morphologique

L'extraction de descripteurs lexicaux par analyse morphologique consiste à trouver une forme commune du vocabulaire selon les règles flexionnelles et dérivationnelles de la langue. Deux approches peuvent être considérées à cet effet : la racinisation et lemmatisation.

La lemmatisation est un processus qui consiste à reconnaître pour chaque mot sa forme de base en supprimant ses traits liés à la variation de catégorie grammaticale (genre, nombre, personne, …etc.). En considérant les variations flexionnelles, la forme unique qu'on doit retrouvée est le lemme. Il n'est pas suffisant de se contenter des règles flexionnelles surtout dans les langues hautement flexionnelles ou induisant une diversité de formes irrégulières. Par exemple en anglais, qui est la langue la plus investie dans ce domaine et n'est pas assez compliquée morphologiquement, les mots {*is*, *was*, *are*, *were*, … } doivent être indexé par un seul lemme "*be*".

[27] Un radical réfère à une racine parfois incomplète.

La racinisation s'intéresse aux variations dérivationnelles pour remonter à une forme unifiée plus abstraite (racine). Cette méthode regroupe des mots appartenant à un champ sémantique commun dans une même unité lexicale. En français par exemple les mots {*lecteur*, *lisible*, *illisible*} remontent à la même racine représentée par l'infinitif du verbe "*lire*". Pourtant, les trois mots représentent, chacun, un lemme distinct.

Du point de vue linguistique, la lemmatisation paraît comme une méthode convenable du moment où elle ramène les mots vers leur forme de lemme (ou lexème) résolvant en partie le problème d'ambigüité. Cependant, la racinisation paraît plus "agressive" que la lemmatisation surtout dans les langues hautement dérivationnelle. Bien que les mots regroupés restent généralement dans le même champ sémantique, le descripteur extrait par racinisation crée une confusion sémantique irréversible.

Par ailleurs, la détection du lemme est un processus qui nécessite une désambigüisation par le contexte. Ce qui dépasse le cadre d'une pré-indexation dans un SRI. Par exemple comment détecter le vrai lemme dans le mot lire ? Est-ce que c'est le verbe *lire* ou bien c'est la devise italienne *lire* ?

Le stemming par analyse morphologique est très coûteuse en matière de complexité de traitement et nécessité de ressources linguistiques externes. Sur le plan pratique, diverses études expérimentales, menées par la communauté de la recherche d'information, ont montré que l'analyse morphologique n'améliore pas la performance moyenne d'un SRI [Salton, 1989] [Hull, 1989].

Il est intéressant de noter que le texte anglais était l'objet d'étude dans la majorité des travaux relatifs aux analyseurs morphologiques et leur impact sur la performance d'un SRI. La langue anglaise s'articule autour d'une morphologie relativement légère et, par conséquent, le besoin de traiter intelligemment la morphologie n'est pas suffisamment justifié [Manning et Schütze, 2001].

4.2 Pseudo-racinisation

Une alternative moins coûteuse et, parfois, plus efficace consiste à réaliser l'extraction de descripteurs par une pseudo-racinisation. Bien que leur sorties ne soient toujours identiques, les termes *racinisation* et *stemming* sont souvent confondus en littérature. Selon notre étude, deux causes essentielles peuvent être derrière cette confusion terminologique : La première revient à la morphologie faible de l'anglais qui est la langue la plus investie dans ce domaine. La distinction entre *racine* et *stem* devient assez claire dans d'autres langues plus riches en morphologie flexionnelle et dérivationnelle. La deuxième est due à la nature elle-même du prétraitement désiré dans une indexation. Peu importe la validité lexicale du terme s'il sert à améliorer les performances d'un SRI.

Le stemming est vu, dans un sens, comme une procédure de fusion des mots dans une même entrée dite "stem" [Lennon et al., 1981]. Toutefois, certains le considère comme un outil de lemmatisation [Jurafsky et Martin, 2000]. Nous optons dans ce qui suit d'utiliser le terme *stemming léger* (ou pseudo-racinisation) pour interpréter une forme simplifiée de l'analyse morphologique dans un prétraitement d'indexation. Elle consiste à opérer de simples troncatures sur le mot en supprimant les affixes communs dans une langue [Manning et Schütze, 2001].

Les algorithmes de stemming les plus utilisés sont ceux de Lovins [Lovins, 1968] et de Porter [Porter, 1980]. Ce dernier a été révisé et recodé dans plusieurs implémentations. L'algorithme de Porter est devenu (ou presque) un référence du stemming dans les applications de RI pour le texte anglais. Il a été étendu plus tard dans *Snowball*[28] pour supporter d'autres langues européennes telles que le français, l'espagnol, le portugais, l'italien et certaines variantes de l'allemand.

4.3 Aspects multi-langages

Les méthodes d'analyse du texte en vue d'extraction de descripteurs ont été expérimentées en premier lieu dans le texte l'anglais. Les conclusions mitigés, et parfois contradictoires sur l'impact des méthodes de stemming sur les performances en RI, reviennent en partie à la morphologie simple de l'anglais. Plusieurs autres langues sont plus riches en flexion et en dérivation et nécessitent, par conséquent, une analyse morphologique spécifique [Manning et Schütze, 2001].

En effet, nombreux sont les travaux proposant et analysant les méthodes de stemming dans les langues européennes et de l'Asie de l'est. Que se soit par analyse morphologique ou par simple stemming, l'étude de l'impact du processus d'extraction de descripteurs sur les tâches de RI reste un sujet d'actualité [Savoy, 2006] [Jongejan et Dalianis, 2009] [Savoy, 2010]. Nous décrivons ultérieurement comment traiter ce problème dans le texte arabe.

5 Evaluation

Les facteurs significatifs de performance du processus d'extraction de descripteurs doivent inclure, mais de façon non-exhaustive, [Kraaij et Pohlmann, 1996] :

- Le type de l'algorithme,
- Les mesures d'évaluation de la tâche de RI,
- La nature du langage,
- La longueur de la requête et des documents.

Par nature, de telles méthodes augmentent le rappel dans une recherche mais diminue la précision [Krovetz, 1993] [Kraaij et Pohlmann, 1996] [Strzalkowski et al. 1999]. Par ailleurs, il n'est pas sûr que la meilleure performance d'une tâche de RI résulte seulement de la qualité de la méthode de stemming [Paice, 1996] [Frakes, 2003]. Nous décrivons dans ce qui suit, trois mesures d'évaluation indépendamment de la performance des tâches de RI.

5.1 Facteur de compression d'index

Le facteur de compression d'index (ICF) nous informe à quel point une collection de mots uniques puisse être réduite (compressée) par une procédure de stemming. L'idée part du fait que le stemmer le plus faible produit le facteur *ICF* le plus grand [Frakes, 2003]. Ce facteur peut être caluculé comme suit

N : le nombre des mots uniques avant le stemming,

S : le nombre des mots uniques après le stemming.

[28] Un cadre de développement des outils de stemming : http://snowball.tartarus.org/

$$ICF = \frac{(N - S)}{N} \quad \textbf{(5-1)}$$

Le facteur *ICF* a été introduit comme une mesure efficace pour évaluer les méthodes de stemming et leur capacité de compression. Ceci est concordant avec l'un des objectifs du processus de pré-indexation en RI.

Cependant, la réduction du vocabulaire ne signifie pas nécessairement un stemming idéal. En effet, une racinisation perfomante est celle qui effectue la meilleure unification du concept (par stem, racine ou lemme).

5.2 Mesures de sous et sur-stemming

Le ***sous-stemming*** représente l'insuffisance de fusion des mots morphologiquement et sémantiquement reliés. Cette insuffisance s'accumule lorsque deux mots, qui devraient avoir la même racine, ne le sont pas. Par exemple, si on considère une racinisation par troncature des suffixes (de flexion du temps) dans les verbes {*vais, allai, irai*}, nous n'aurons jamais la même racine du verbe *aller*.

Le sur-stemming réfère aux mots regroupés par la racinisation dans une seule forme alors qu'ils le ne devront pas. Par exemple, fusionner les mots *probe* et *probable* dans un seul stem va constituer une erreur de sur-stemming.

En utilisant un échantillon de *W* groupes de mots, ces deux mesures ont été proposées est et expérimentées dans [Paice, 1996]. Un groupe-concept contient des formes qui sont, à la fois sémantiquement et morphologiquement, reliées. Pour chaque groupe *g* contenant n_g mots, le nombre de pairs des mots distincts va définir le compte total des fusions désirées (DMT_g)[29] :

$$DMT_g = \frac{n_g(n_g - 1)}{2} \quad \textbf{(5-2)}$$

Du moment où un algorithme de stemming performant ne doit pas fusionner aucun mot membre d'un groupe *g* avec ceux des autres groupes, nous pouvons définir le compte total des fusions non-désirées (DNT_g)[30] comme suit :

$$DNT_g = \frac{n_g(W - n_g)}{2} \quad \textbf{(5-3)}$$

Lorsqu'on calcule les deux totaux des équations (5-2) et (5-3) pour obtient le compte global des fusions désirées (*GDMT*) et celui des fusions non-désirées (*GDNT*). Ainsi, on peut calculer les erreurs de stemming comme suit :

- L'index de fusion *CI* : représente la proportion des pairs de mots équivalents et correctement groupés (par rapport à *GDMT*).
- L'index de distinction *DI* : représente la proportion des pairs de mots non-équivalents et restant non groupés après stemming (par rapport à *GDNT*).

[29] Acronyme de "desired merged total"

[30] Acronyme de "desired non-merge total"

Et donc, les index de sous-stemming (*UI*) et de sur-stemming (*OI*) peuvent être définis par [Paice, 1996] :

$$UI = 1 - CI \quad \textbf{(5-4)}$$

$$OI = 1 - DI \quad \textbf{(5-5)}$$

Dans [Paice, 1996], les auteurs propose de calculer le ratio des quantités (5-4)et (5-5) comme mesure du poids d'un algorithme de stemming (*SW*) :

$$SW = OI / UI \quad \textbf{(5-6)}$$

Malgré l'avantage qu'apporte la compression de l'index, elle n'est utile qu'à un certain point. C'est l'objectif de la mesure d'erreur de stemming de *Paice*. Car en alourdissant le stemming pour réduire l'index, la fusion des concepts distincts devient de plus en plus fréquente. Dans cette situation, il est évident que le rappel augmente au détriment de la perte en précision [Frackes, 2003].

L'approche d'évaluation proposée par [Paice, 1996] nécessite un complément assez important. En effet, l'élaboration et la validation de l'ensemble de groupe-concepts constituent une tâche délicate dont le produit doit être approuvé et standardisé. Ces groupes-concepts ont été construits par jugement humain durant un examen minutieux de quelques échantillons de listes de mots [Paice, 1996] [Frackes, 2003].

6 Analyse du texte arabe

Contrairement aux langues Indo-Européens, l'Arabe appartient à la famille des langues sémitiques. Elle est écrite de droite à gauche et compte 28 lettres. Malgré que les populations arabes parlent plusieurs dialectes, il existe une seule forme écrite, dans la plupart des documents publiés, connue sous le nom de l'Arabe standard moderne et que nous la référençons tout simplement ici par l'arabe [Kadri et Nie, 2006].

6.1 Caractéristiques de la langue arabe

Avec une grande capacité de dérivation morphologique, l'arabe ne compte que trois catégories morphosyntaxiques (nom, verbe ou particule). Influencée par leur origine anglaise, plusieurs études relatives au traitement automatique du langage naturel utilisent d'autres catégories supplémentaires pour l'arabe, telles que les adverbes et les prépositions [Larkey et al., 2002] [Moukdad, 2006] [Tuerlinckx, 2004] [Moukdad, 2006].

6.1.1 Complexité morphologique

Tout mot arabe peut être restitué à sa racine selon l'une des fromes de dérivation. La racine en arabe est l'unité linguistique non-vocalisée (sans diacritiques) portant un champ sémantique déterminé. Elle est composée généralement de 3 lettres et rarement de 4 ou 5 [Tuerlinckx, 2004] [Kadri et Nie, 2006].

Moins abstrait qu'une racine mais complètement vocalisé, un lemme constitue l'entrée du dictionnaire arabe bien qu'il soit trié par racine. En plus de sa morphologie hautement dérivationnelle, l'arabe se distingue par la conjonction de deux caractéristiques, l'agglutination et la non-vocalisation, compliquant ainsi les méthodes d'analyse automatique.

Dans les langues sémitiques, la racine est l'élément principal à partir duquel différents mots peuvent être dérivés. La complexité morphologique en arabe combine la flexion et la dérivation de façon étroite. La flexion peut être obtenue par différentes forme d'affixation (préfixe, suffixe, infixe). L'exemple suivant illustre certaines formes irrégulières du pluriel :

- De la racine [**Elm**]علم[31] : le pluriel de [**Ei**l**m**, *science*]عِلم est [**Eu**l**u**m, *sciences*]عُلوم,
- De la racine [**ktb**]كتب : le pluriel de [**ki**t**A**b, *livre*]كِتاب est [**ku**t**u**b, *livres*]كُتُب.

La dérivation en arabe génère de nouveaux mots à partir d'une racine tout en restant dans le même champ sémantique. Par exemple, on peut dériver de la racine [**Elm**]علم les verbes [>a**E**o**l**a**m**, *notifier/informer*]أعلم et [{isota**E**o**l**a**m**, *enquêter*]إستعلم.

6.1.2 Agglutination

Dans le texte arabe, il est difficile d'identifier automatiquement une unité lexicale à partir d'un graphème. Le processus d'affixation morphologique devient plus compliqué lorsque des affixes supplémentaires sont collés au lemme. Un analyseur doit considérer quatre types d'affixes rattachables à la racine d'un mot arabe. Nous illustrons dans la Figure 5-1 un exemple d'une forme agglutinée, *[sayaEolamuwnahu]* سَيَعْلَمُونَهُ.

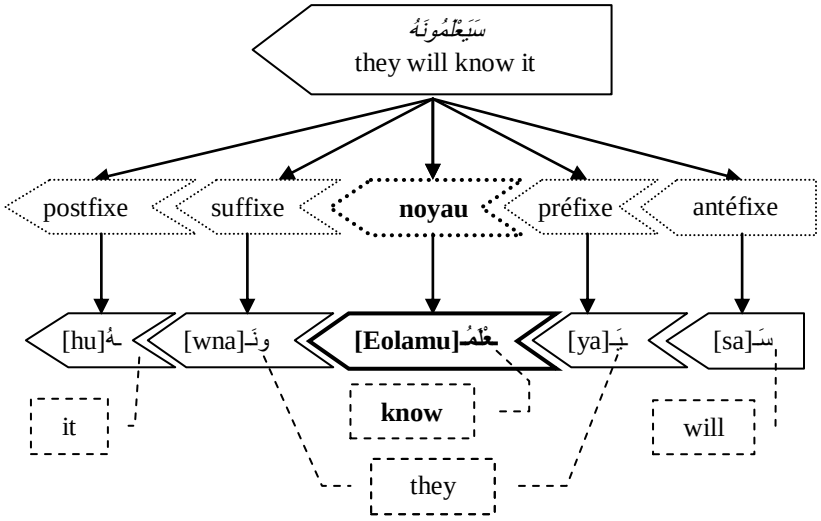


Figure 5-1. Analyse du graphème *[sayaEolamuwnahu]* سَيَعْلَمُونَهُ.

6.1.3 Non-vocalisation

En plus des trois voyelles de base (ا,و,ي), la vocalisation d'un mot arabe utilise les diacritiques (petites voyelles). Cependant, nous ne trouvons cette vocalisation que dans le texte coranique ou les documents didactiques. Cette situation accentue l'ambigüité des mots et oblige tout analyseur automatique de prêter plus d'attention à la morphologie et le contexte du mot. Le Tableau 5-1 présente un exemple d'un mot composé de 3 consonnes seulement mais dont la segmentation change radicalement à cause de l'absence des diacritiques.

Solution	Morphologie	Vocalisation	Anglais	Français
1	Nom	[basom] بَسْم	*Smiling*	*Sourire*
2	Verbe	[basam] بَسَم	*Smile*	*Sourire*
3	Prép. + nom	[bi]بـ + [somi] سْم	*In/by (the) name of*	*Au nom de*
4	Prép. + nom	[bi]بـ + [sam~] سَمّ	*By/with poison*	*Par/avec poison*

Tableau 5-1. Quatre solutions possibles pour la segmentation du mot [bsm] بسم.

[31] Selon la translitération de *Buckwalter* et en application des exigences de rédaction des articles scientifiques portant des termes non-latins.

Une telle ambigüité dans cette situation peut compliquer la détection de la bonne racine d'une forme agglutinée. Dans un texte non-vocalisé, l'analyse morphologique devient de plus en plus difficile comme illustré dans Figure 5-1 et Tableau 5-1. L'agglutination est une caractéristique qu'on trouve dans d'autres langues telles que le japonais et le finnois mais son adjonction avec le problème de non-vocalisation va l'aggraver davantage dans le texte arabe.

6.2 Travaux relatifs à l'analyse du texte arabe

La méthode de racinisation de *Khoja* figure parmi les approches efficaces pour le stemming du texte arabe [Khoja et Garside, 1999]. Basée sur des listes prédéfinies de racines et une analyse morphologique, l'algorithme de *Khoja* tente d'extraire une racine correcte d'un graphème arabe. Néanmoins, nous avons vu dans la section précédente qu'un mot arabe non-vocalisé peut être dérivé différemment de plusieurs racines. Bien le racineur *Khoja*[32] n'a pas été maintenu depuis sa première publication, il a été largement utilisé et analysé comme référence dans les travaux ultérieurs. A titre d'exemple, l'algorithme de stemming à base de lemme d'Al-Shammari avait inclut le racineur *Khoja* pour le traitement des verbes [Al-Shammari et Lin, 2008]. Les auteurs ont combiné le stemming léger, la racinisation et la consultation d'un dictionnaire pour le tester efficacement dans la tâche d'agrégation (clustering) [Al-Shammari, 2010]. Les expérimentations ont montré que l'algorithme d'Al-Shammari est plus performant, en termes de sur-stemming (voir section 5.2), que le racineur *Khoja* et d'autre méthodes de stemming léger [Al-Shammari, 2010].

Pour le stemming léger, plusieurs variantes ont été développées au début du troisième millénaire [Larkey et al., 2002]. En testant sur la collection AFP_ARB (voir Chapitre 2 :8.3.4), les auteurs déduisirent que le stemming léger est plus efficace que l'analyse morphologique dans la recherche multilingue. Ils réclament qu'il n'est pas essentiel d'avoir une forme correcte d'une racine pour un algorithme de stemming [Larkey et al., 2002]. Etonnement, les auteurs déclare dans un rapport technique que leurs résultats ont été obtenus sans aucune connaissance préalable de la langue l'arabe [Larkey et Connell, 2001]. Une étude ultérieure a confirmé le même constat en préférant le stemming léger pour les tâches de recherche dans le texte arabe [Moukdad, 2006].

Par ailleurs, le stemming léger ou l'utilisation des formes brutes avaient donnés, tous les deux, les mêmes performances dans l'analyse des thèmes du texte arabe [Brants et al., 2002]. Une étude récente sur la catégorisation du texte arabe a soulevé cette contradiction en littérature en tentant d'analyser une variété d'outils de stemming [Said et al., 2009]. Dans [Darwish et al., 2005], les auteurs avaient montré que l'utilisation du contexte dans le processus d'extraction des racines peut améliorer la performance des tâches de recherche. Néanmoins, l'analyse de contexte nécessite un coût de calcul trop élevé par rapport au stemming léger ou à la racinisation.

Similaire à l'approche de *Khoja* mais sans dictionnaire de racines, l'algorithme *ISRI*[33] basé sur le stemming léger a été proposé et efficacement expérimenté dans [Taghva et al., 2005]. Les auteurs conclurent que les listes prédéfinies de racines ne sont pas indispensables pour la racinisation de l'arabe dans un processus de RI.

Comparées à l'anglais, ou du moins à d'autres langues européennes, les recherches ne sont pas suffisamment investies dans le stemming du texte arabe pour les tâches de RI et

[32] Librement disponible sur http://zeus.cs.pacificu.edu/shereen/ArabicStemmerCode.zip

[33] Librement disponible sur http://nltk.googlecode.com/svn/trunk/doc/api/nltk.stem.isri-pysrc.html

l'extraction des connaissances [Taghva et al., 2005] [Said et al., 2009]. L'effort principal pour développer des systèmes de RI efficaces pour le texte arabe a été dépensé dans un cadre commercial. Les approches utilisées aussi bien que l'évaluation des performances ne sont pas publiés. Comme exemple significatif, nous citons le système *Siraj*[34] qui permet classer le texte arabe et d'extraire les entités nommées (personne, endroit, organisation, …etc.) avec satisfaction individuelle de l'utilisateur. Cependant, nous ne pouvons trouver aucune documentation technique pour la description ou l'évaluation du système.

A travers cet aperçu des approches de stemming du texte arabe, il est intéressant de souligner le fait que peu d'algorithmes ont reçu une évaluation standard [Said et al., 2009] [Al-Shammari, 2010]. Ceci est, peut être, parmi les aspects, et les causes en même temps, d'une telle contradiction en littérature.

Pourtant nous retenons, à cet effet, l'étude réalisée pour une évaluation rigoureuse des méthodes de stemming du texte arabe [Al-Shammari, 2010]. Afin d'estimer la performance des algorithmes de stemming arabe par les mesures de *Paice* (voir la section 5.2), l'auteur a conçu un échantillon de 419 mots répartis sur 81 groupes-concepts (près de cinq mots par groupe). En comparant aux algorithmes de *Khoja* et de stemming léger, elle montre que son lemmatiseur Al-Shammari a réduit considérablement les erreurs du sur-stemming. Cependant, aucune amélioration significative n'a été reportée pour le sous-stemming [Al-Shammari, 2010].

6.3 Méthodes de stemming du texte arabe

Afin fusionner efficacement les mots arabes, le traitement des aspects combinées (de flexions, de dérivation, d'agglutination et non-vocalisation) suit deux approches : soit par stemming léger, en supprimant des affixes communs, soit par analyse morphologique, en cherchant chaque noyau (racine ou lemme) selon un schéma déterminé.

6.4 Stemming léger

L'efficacité de cette approche dépend de la nature morphologique de la langue, du contenu des listes des affixes utilisé et de l'algorithme lui-même. Le stemming du texte arabe nécessite la prise en charge des formes ambigües et agglutinées impliquant plusieurs dérivations morphologiques. L'analyse d'une telle approche peut être revue dans [Larkey et al., 2002]. L'algorithme *ISRI* constitue une autre alternative pour le stemming léger sans disposer préalablement d'une liste de racines [Taghva et al., 2005]. Il utilise une liste étendue d'affixes avec des schémas de dérivation des plus fréquents pour extraire des racines. Néanmoins, la normalisation dans les mots non trouvés reste irréversible.

Ce type d'algorithmes peut effectivement traiter une majorité des cas les plus fréquents. Néanmoins, le mot correct peut être perdu dans d'autres situations. Par exemple, dans le mot [wafiy]وَفِي , quelqu'un peut lire deux prépositions agglutinées signifiant "et dans". Mais un autre va le lire comme un seul nom signifiant "fidèle".

6.5 Analyse morphologique

Selon le type de sortie désirée, on peut distinguer deux catégories d'analyse morphologique pour le texte arabe : le stemming à base racine et les stemming à base lemme. Le choix entre telle ou telle approche dépend de la nature de la tâche RI qui suivra.

[34] *Siraj* est un produit de la compagnie *Sakhr* accessible sur : http://siraj.sakhr.com/

Dans la première catégorie, l'algorithme *Khoja* a été proposé avec une liste de racines et de schémas afin d'extraire d'un mot une racine correcte [Khoja et Garside, 1999]. Cet algorithme permet de produire des racines abstraites ce qui réduit considérablement la dimension de l'espace des descripteurs des documents. Cependant, il conduit vers une confusion ennuyante de sens divergents à cause de la non-vocalisation. Par exemple, le mot [sayaEolamuwnahu]سَيُعَلِّمُونَهُ, cité dans la Figure 5-1, doit être dérivé de la racine [Elm] علم interprétant les verbes {*connaître*, *enseigner*}. Néanmoins, la racine isolée [Elm] علم peut signifier, comme troisième sens, le nom {*drapeau*} qui ne doit pas figurer dans cette situation.

En deuxième catégorie, l'algorithme à base de lemme d'*Al-Shammari* a été développé et comparé à celui de *Khoja* [Al-Shammari and Lin, 2008]. L'approche combine le stemming léger avec la racinisation *Khoja* pour la dés-affixation des verbes avant de traiter le reste des mots en exploitant un dictionnaire des verbes et des noms. Pour l'évaluation, les auteurs ont exploité 7000 documents collectés des ressources de presse en-ligne en plus d'un sous-ensemble (taille inconnue) du corpus d'apprentissage multilingue *ACE2004*[35] du clatalogue LDC. En plus de l'évaluation de la tâche d'agrégation, les auteurs ont utilisé une collection de 81 groupe-concepts pour mesurer le sous et sur-stemming de différents algorithmes.

L'approche suivie dans les travaux récents sur l'algorithme d'*Al-Shammari* trace un cadre astucieux pour la conception et l'évaluation des méthodes de stemming du texte arabe [Al-Shammari and Lin, 2008] [Al-Shammari, 2010]. Néanmoins, ni l'implémentation, ni les collections de test n'ont été mis à disposition de la communauté des chercheurs dans le domaine. L'utilisation d'une liste excédant les 2200 mots vides laisse des zones d'ombres sur le prétraitement de normalisation et la nature des textes utilisés dans la validation. Par ailleurs et comme montré dans Tableau 5-1, un graphème arabe admet avoir plusieurs possibilités de lemmatisation. Malheureusement, aucune des méthodes de stemming développées pour la recherche du texte arabe n'avait pris en charge cet aspect.

A cet effet, nous trouvons en littérature un ensemble de ressources lexicales, qui ont été développées en 2002 et raffinées en 2004, pour détecter les règles de flexion et de composition dans un mot arabe [Buckwalter, 2002]. L'analyseur morphologique de *Buckwalter*[36] a été incorporé plus tard dans le package *AraMorph*[37] pour la lemmatisation du texte arabe. Plusieurs solutions peuvent être proposées pour chaque mot en entrée. Partant de ces ressources linguistiques et sous certaines considérations, on peut développer un algorithme de stemming à base de lemme pour l'indexation du texte arabe dans un SRI. Cette approche sera décrite dans la section suivante.

7 Analyseur à base de lemme BBW

Nous proposons de développer un algorithme pour le stemming à base de lemme dans le texte arabe. L'analyseur, qui doit s'intégrer dans un processus d'indexation, exploite les ressources libres de la première version de l'analyseur de *Buckwalter* (la deuxième version étant payante et protégé). Notre algorithme, que nous proposons de le dénoter par *BBw* (en référence à *Brahmi-Buckwalter*), porte deux contributions principale : le traitement de normalisation et la sélection du stem par analyse morphologique.

[35] http://www.ldc.upenn.edu/Catalog/catalogEntry.jsp?catalogId=LDC2005T09

[36] http://www.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC2002L49

[37] http://www.nongnu.org/aramorph/english/index.html

7.1 Normalisation

Cette phase doit être opérée sur le texte brut en entrée pour l'aligner et supprimer toute perturbation typographique qui peut nuire à l'analyse morphologique. Afin de récupérer en sortie des graphèmes normalisés, nous appliquons les transformations suivantes sur le texte brut :

- Convertir le texte vers l'encodage UTF-8,
- Segmenter le texte selon la ponctuation standard,
- Supprimer les diacritiques et *tatweel* (ـ)
- Supprimer les lettres non arabes et les mots vides,
- Remplacer *alef* initiale avec *hamza* (أ ou إ) par *bar-alef* (ا),
- Remplacer *waw* finale ou *yeh* avec *hamza* (ؤ ou ئ) par *hamza* (ء),
- Remplacer *maddah* (ـٰ) ou *alef-waslah* (ٱ) par *bar-alef* (ا),
- Remplacer deux *bar-alef* (اا) par *alef-maddah* (آ),
- Remplacer *teh-marbuta* finale (ة) par *heh* (ه),
- Supprimer *yeh* finale (ي) lorsque le stem restant est valide.

7.2 Sélection du stem

Trois cas peuvent être recensés pour l'analyse d'un graphème arabe en entrée : (i) une solution selon un schéma unique de dérivation. (i) une multiplicité de solution pour le même graphème. (iii) aucune solution n'est détectée. Les actions que l'analyseur BBw doit entreprendre sont décrites ci-dessous :

Solution unique : l'analyseur *BBw* retient seulement le lemme non-vocalisé de la solution (sans affixes). Une solution qui ne contient ni verbe ni nom (c.a.d. composée de particules) sera ignorée et considérée comme un mot vide.

Solution multiple : l'analyseur *BBw* traite toutes les solutions proposées comme un ensemble de solutions uniques séparées en retenant les lemmes non-vocalisés. Notons ici que l'élimination des diacritiques des lemmes proposés peut les unifier et réduire par conséquent la multiplicité. Par exemple, quatre solutions vocalisées sont possibles pour le graphème *[bsm]* بسم dans le Tableau 5-1, mais l'analyseur *BBw* ne va produire que deux lemmes non-vocalisés {*[bsm]* بسم, *[sm]* سم}.

Aucune solution : lorsqu'aucune solution de lemmatisation ne puisse être trouvée, différentes raisons peuvent être soulevées : (i) le mot en entrée est erroné et il n'implique aucun lemme arabe. (ii) le mot correspond à un nom propre sans entrée valide dans le lexique arabe. (iii) le mot est correct mais son lemme n'est pas encore inclus dans la version actuelle de l'analyseur de *Buckwalter*.

A priori, nous n'avons aucune information préalable pour nous aider à détecter la cause d'un tel échec d'analyse. En plus, la multiplicité de solution ne peut être désambigüisée à ce stade de l'analyse. Nous proposons de développer et expérimenter trois variantes (*BBw0*, *BBw1*, *BBw2*) selon la manière de traiter de telles situations. Nous résumons dans Tableau 5-2 l'essentiel des actions à entreprendre par ces différentes variantes.

Analyseur	Solution unique	Solution multiple	Aucune solution
BBw0	1 lemme	tous les lemmes	Ignorer
BBw1	1 lemme	tous les lemmes et le mot en entrée	le mot en entrée
BBw2	1 lemme	tous les lemmes	le mot en entrée

Tableau 5-2. Description des résultats d'analyse par les variantes *BBwX* .

7.3 Mesure de l'ambigüité lexicale

Afin d'évaluer l'ambigüité relative engendrée par l'analyseur *BBw*, nous proposons d'utiliser une mesure simple. Partant d'un ensemble de documents *S* et un algorithme de stemming *R* et pour chaque mot en entrée (t_i), *R* donne (d_i) lemmes distincts (le degré de multiplicité individuel de l'analyse de (t_i). Notons par *L* le nombre total des mots de l'ensemble *S* (après suppression des mots vides) et L_R le nombre total de stems obtenus par l'algorithme *R*. Nous définissons alors le degré de confusion $C(S|R)$ comme suit :

$$C(S|R) = \frac{1}{L}\sum_{i=1}^{L} d_i = \frac{L_R}{L} \qquad \textbf{(5-7)}$$

En appliquant cette mesure pour l'algorithme *Khoja*, qui ne donne qu'une seule racine au plus, il est clair que pour toute collection *S*, $C(S|Khoja)=1$.

Toutefois, le degré de confusion relatif à l'analyseur *BBw* devrait augmenter ($C(S|BBw)\geq 1$). Puisque cette approche peut proposer plusieurs solutions pour le même mot.

Pour un lecteur arabe humain, l'ambigüité lexicale dans un texte brut peut être résolue par des considérations sémantiques guidées par le contexte. Cette confusion n'est pas produite par l'analyseur *BBw* puisqu'elle est réelle et revient à la nature elle-même de la langue arabe. Il faut noter ici qu'on ne dispose d'aucune connaissance préalable pour choisir un stem parmi les solutions multiples proposées par *BBw*. C'est la raison pour laquelle nous avons choisi de retenir tous les lemmes possibles pour représenter la forme en entrée de façon équitable. Nous estimons qu'un calcul statistique sur des collections réelles assez larges pourrait contribuer à lever cette ambigüité. Les modèles à sémantique latente peuvent être très utiles dans cette situation.

8 Conclusion

Ce chapitre a été consacré aux aspects linguistiques dans un processus de pré-indexation documentaire. Malgré l'histoire relativement longue des modèles de représentation des textes en RI, la littérature recensée accorde peu d'importance au prétraitement linguistique. L'hypothèse commune de "sac de mots" est devenue pratiquement "sac de stems" ou "sac de séquences de caractères". La dominance de l'anglais, dont la morphologique est relativement légère, dans le développement des modèles de RI en était probablement la cause principale.

Bien qu'elle reste toujours en première position, l'usage de la langue anglaise sur Internet n'a pas dépassé en 2011 les 27% avec un peu plus d'un demi-milliard d'internautes[38]. L'importance que d'autres langues commencent à gagner sur le Web, comme dans les entreprises et les institutions internationales, requiert une attention plus fine quant à sa prise en charge dans les méthodes d'indexation en RI.

[38] Source : *Internet World Stats* http://www.internetworldstats.com/stats7.htm (31/05/2011)

Les techniques pragmatiques de stemming léger sans analyse morphologique adéquate paraissent inadaptées aux tendances actuelles en recherche sémantique, pour satisfaire au mieux le besoin d'information. L'évaluation des méthodes de stemming étaient en majorité liées aux mesures de performance dans les tâches de RI subséquentes. Nous avons décrits dans ce chapitre d'autres mesures de stemming indépendamment de l'évaluation des modèles postérieurs.

Si certaines langues à légère morphologie offrent des évaluations performantes en RI, des études récentes sur d'autres langues plus complexes, telles que l'arabe, affichent des constats mitigées nécessitant d'investir davantage dans ce domaine. Une partie de ce chapitre était consacrée à l'analyse du texte arabe en vue d'une indexation documentaire efficace dans un SRI. Notre étude a décrit clairement les aspects morphologiques et lexicaux d'un mot arabe. Notre proposition relative à l'analyseur morphologique *BBw* est donnée dans ce contexte d'évaluer les modèles de thème pour la recherche d'information en arabe. Nous détaillons l'approche des expérimentations menées à cet effet et leurs résultats dans le chapitre suivant.

Chapitre 6 :
ANALYSE SEMANTIQUE DES TEXTES ARABES NON STRUCTURES

1 Introduction

La recherche d'information a connu des progrès significatifs couvrant l'indexation sémantique des textes et l'analyse multi-langage. Cependant, les développements relatifs à la recherche d'information en langue arabe, n'ont pas suivi la croissance extraordinaire de son utilisation sur le Web. En effet, les études recensées dans ce domaine, soit en mono-langage soit en textes parallèles, restent limitées. En particulier l'application de la catégorisation des textes par apprentissage automatique et la modélisation par thème, qui s'avèrent de bonnes démarches pour prendre en charge la sémantique imbriquée dans les textes, est quasi-absente pour les documents arabes et donc insuffisantes pour apprécier l'efficacité de ces approches dans les textes autres que l'anglais.

Le présent chapitre décrit les aspects pratiques de notre contribution pour analyser la sémantique du texte arabe en vue de l'indexer dans une application recherche d'information. Nous comparons notre analyseur *BBw* avec d'autres algorithmes de stemming dans trois corpus réels automatiquement extraits de la presse en-ligne. La modélisation par thème latents (LDA) est appliquée pour apprécier son efficacité d'indexation sémantique dans le texte arabe.

2 Construction des Collections

Vu le déficit enregistré dans les corpus libres des textes arabes, nous avons opté pour construire nos propres corpus en développant une application, *Stories-Crawler*[39], d'exploration spécifique du Web pour collecter des articles de différentes sources (voir Figure 6-1). Chaque site a été préalablement analysé pour détecter la structure adoptée pour l'archivage des articles. Plusieurs variantes algorithmiques ont été développées pour explorer et récupérer les documents des différentes catégories proposées par le site. Un fichier (*xml*) de graine de crawl a été conçu de façon adaptable à chaque type de source.

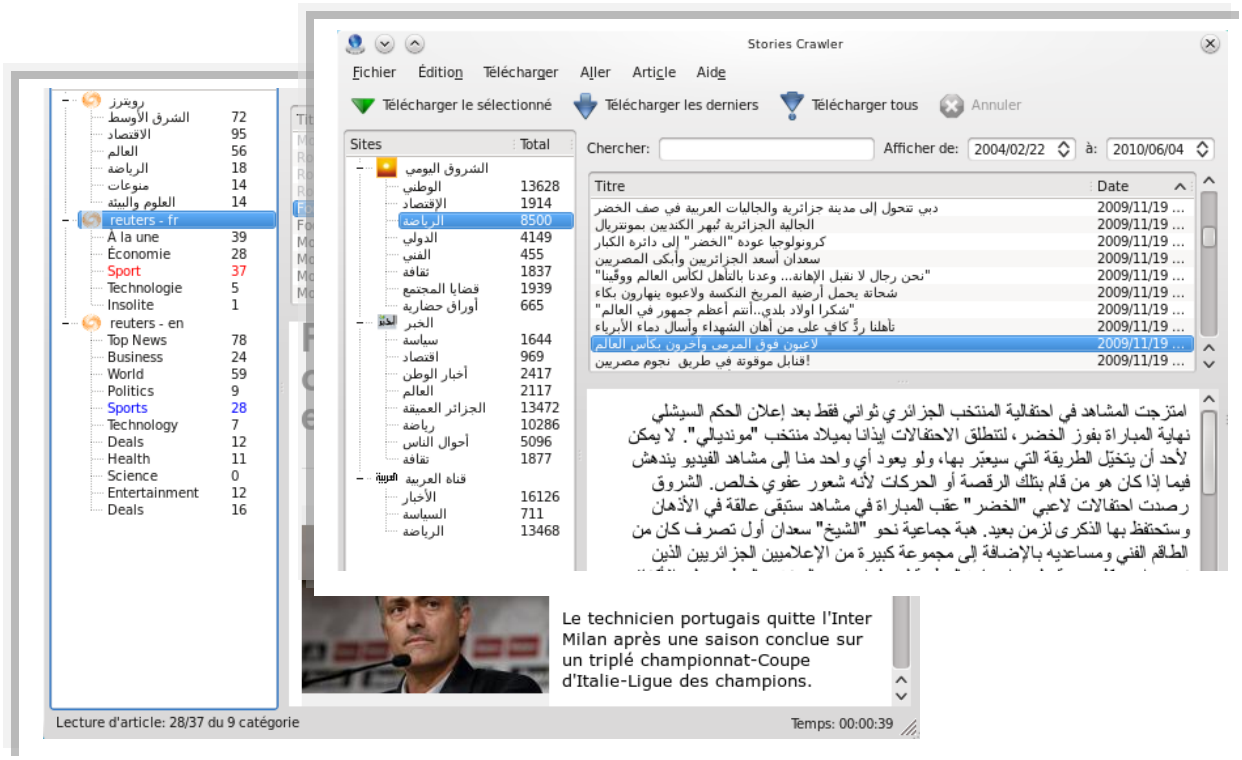


Figure 6-1. Fenêtres de crawling pour les sites de presse.

2.1 Description des corpus arabes

Dans cette étude, nous présentons trois collections d'articles de presse relatifs à la période 2007-2009 et extraits des sites *Echorouk*[40], *Reuters*[41] et *Xinhua*[42]. Une description sommaire, des trois collections, est présentée dans Tableau 6-1.

[39] Développé avec Amine Roukh et Abdelkadir Sadok à l'université de Mostaganem.

[40] Journal algérien avec une édition en-ligne sur http://www.echoroukonline.com/

[41] Agence de presse internationale avec une édition en-ligne sur http://ara.reuters.com/

[42] Agence de presse chinoise avec une édition en-ligne sur http://arabic.news.cn/

Caractéristique \ Collection	Ech-11k (Echorouk)	Rtr-41k (Reuters)	Xnh-36k (Xinhua)
# articles	11.313	41.251	36.696
# caractères	48.247.774	111.478.849	85.696.969
# mots	4.388.426	10.093.707	7.532.955
# mots arabes	3.341.465	7.892.348	6.097.652
Moyenne (#mot/article)	387,9	244,7	205,3
# catégories	8	6	8

Tableau 6-1. Description des trois collections d'articles de presse *d'Echorouk*, *Reuters* et *Xinhua*.

La collection *Ech-11k* contient 11.313 documents du quotidien *Echorouk* couvrant la période 2008-2009. Les articles du corpus sont répartis sur 8 catégories. Un sous-ensemble *Ech-4000* de 4.000 documents a été extrait pour les évaluations préliminaires.

Le corpus *Rtr-41k* contient 41.251 articles arabes, de l'agence de presse *Reuters*, relatifs à la période 2007-2008-2009 et sont étiquetés selon 6 catégories. Pour les évaluations préliminaires, nous utilisons le sous-ensemble *Rtr-5251* contenant 5.251 documents.

Le corpus *Xnh-36k* est constitué de 36.696 documents arabes, de l'agence de presse *Xinhua*, couvrant la période 2008-2009. Le corpus est réparti sur 8 catégories et un sous-ensemble de 4.500 articles est réservé pour les tests préliminaires.

Nous présentons dans Tableau 6-2 la description thématique des trios collections selon les catégories utilisés par les éditeurs.

	Source	Echorouk		Reuters		Xinhua	
	Catégorie \ Corpus	**Ech-11k**	**Ech-4000**	**Rtr-41k**	**Rtr-5251**	**Xnh-36k**	**Xnh-4500**
1	**Monde**	2.274	572	10.000	1.000	9.465	561
2	**Economie**	816	572	10.000	1.000	6.862	563
3	**Sport**	3.554	572	10.000	1.000	1.132	563
4	**Moyen-Orient**	-	-	10.000	1.000	9.822	563
5	**Science-Santé**	-	-	889	889	1.993	563
6	**Culture-Education**	566	566	-	-	1.508	562
7	**Algérie**	2.722	573	-	-	-	-
8	**Société**	808	572	-	-	-	-
9	**Art**	315	315	-	-	-	-
10	**Religion**	258	258	-	-	-	-
11	**Divertissement**	-	-	362	362	-	-
12	**Chine**	-	-	-	-	4.654	563
13	**Tourisme-Ecologie**	-	-	-	-	1.260	562
	Total	**11.313**	**4.000**	**41.251**	**5.251**	**36.696**	**4.500**

Tableau 6-2. Distribution des trios collections sur les différentes catégories.

2.2 Analyse statistique des corpus

Un corpus en soi ne peut rien faire sauf qu'il sert à stocker les éléments du langage utilisé [Manning et Schütze, 2001]. Nous présentons dans cette section une description analytique

des trois corpus en opérant certains tests statistiques. Dans cette phase, seul un prétraitement léger de normalisation est appliqué pour récupérer les graphèmes. Nous retenons tous les termes (pas de suppression de mots vides) sans aucun traitement linguistique.

2.2.1 Contribution du document dans un corpus

Afin d'apprécier la contribution d'un document dans le corpus, nous calculons le nombre de mots distincts rajoutés par chaque nouveau document. Nous présentons dans Tableau 6-3 une estimation préliminaire de la contribution des premiers 1000 documents.

	Echorouk		**Reuters**		**Xinhua**	
# docs	**# mots**	**# distincts**	**# mots**	**# distincts**	**# mots**	**# distincts**
10	3.880	2.147	2.790	1.489	1.396	815
20	9.725	4.722	5.895	2.833	3.092	1.581
30	14.044	6.308	9.308	4.170	4.576	2.267
40	17.716	7.524	11.797	4.980	6.813	3.102
50	20.415	8.375	13.607	5.480	9.187	3.935
60	22.646	9.019	15.088	5.845	11.162	4.528
70	25.445	9.910	17.103	6.419	13.060	5.088
80	31.459	11.427	20.785	7.294	15.032	5.648
90	37.096	12.775	23.725	7.981	17.082	6.139
100	40.085	13.583	25.779	8.442	19.262	6.704
200	85.327	22.321	50.707	13.283	38.496	10.878
400	175.633	34.960	103.065	20.051	76.243	16.910
600	260.141	44.173	152.195	25.191	111.288	21.360
800	357.794	53.090	201.179	29.504	146.228	25.285
1.000	443.736	59.683	250.526	33.189	179.691	28.662

Tableau 6-3. Contribution des premiers 1000 documents dans les 3 corpus.

Par ailleurs, nous présentons dans Figure 6-2 l'allure complète de la contribution des documents dans les 3 collections. Il est clair que l'évolution du nombre des mots distincts devient moins significative avec une taille assez importante du corpus. Même avec le corpus le moins volumineux *Ech-11k*, nous constatons que presque le même nombre de mots distincts est atteint. Ceci peut être expliqué par la richesse du vocabulaire de rédaction des auteurs du quotidien *Echorouk*.

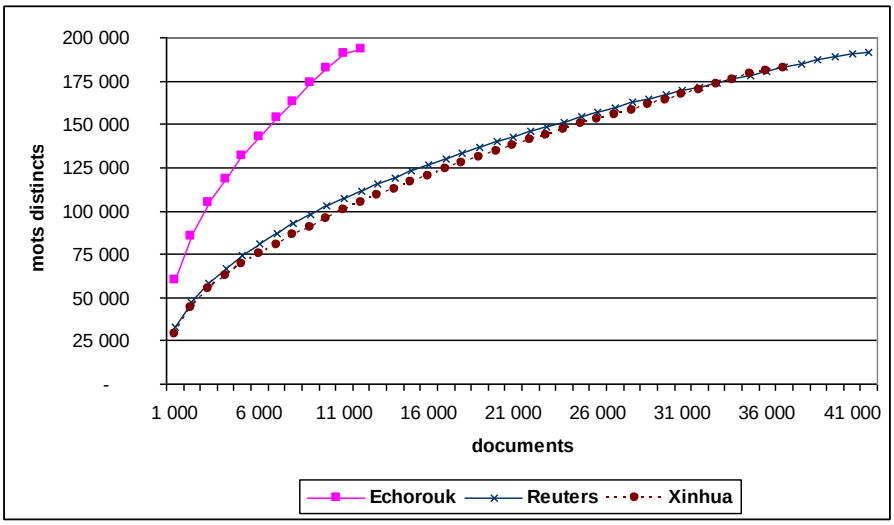


Figure 6-2. Contribution des documents dans la construction des 3 corpus.

2.2.2 Fréquence des mots et loi de Zipf

En calculant la fréquence des mots sur tous les documents d'une collection, nous pouvons détecter ceux qui sont les plus utilisés dans le langage des auteurs. Nous présentons dans Tableau 6-4 la distribution des 30 mots les plus fréquents dans chacun des 3 corpus. Le lecteur peut facilement détecter que la majorité des mots fréquents sont insignifiants et donc peuvent être considérés comme mots vides. Les mots significatifs (en gras dans le tableau) sont rares et respecte par conséquent l'hypothèse de Zipf. En plus, nous constatons que les mots les plus fréquents sont presque les mêmes dans les 3 corpus. Notons ici que le stemming des collections devrait unifier, par la suite, plusieurs mots ce qui augmenterait leur fréquence.

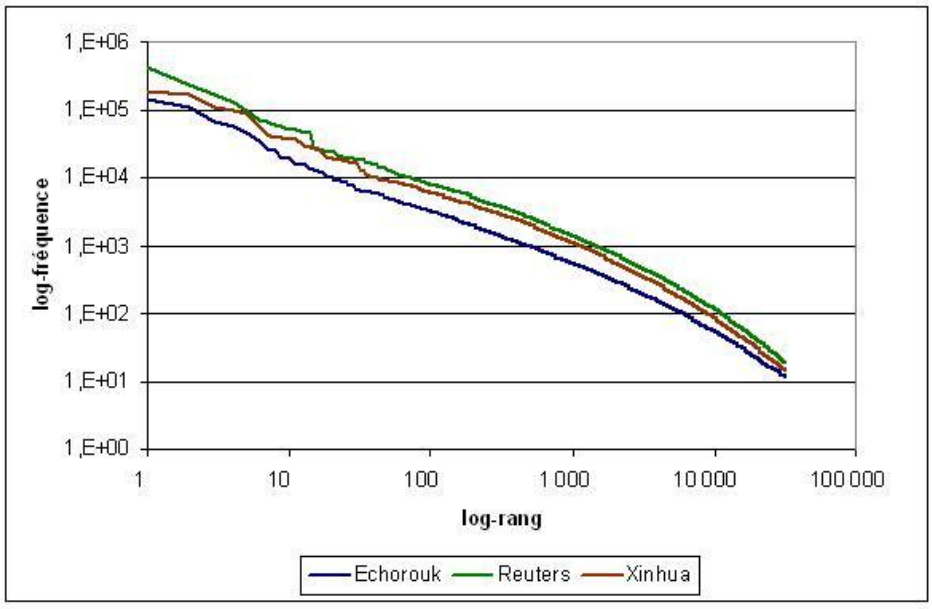


Figure 6-3. Courbe log-log de la fréquence en fonction du rang des 2^{16} mots les plus fréquents.

La loi de Zipf (voir Chapitre 3 :3.3) est vérifiée dans les 3 corpus. En traçant la courbe logarithmique de la fréquence des mots en fonction de leur rang (Figure 6-3), nous obtenons l'allure d'une droite de pente (-1).

Rang	Echorouk			Reuters			Xinhua		
	Mot	**Freq**	**taux**	**Mot**	**freq**	**taux**	**mot**	**freq**	**taux**
1	في	140.637	3,25	في	433.196	4,30	في	187.904	2,50
2	من	110.568	2,56	من	244.321	2,43	من	170.364	2,27
3	على	66.680	1,54	على	166.636	1,66	فى	115.609	1,54
4	أن	57.492	1,33	ان	126.501	1,26	على	103.094	1,37
5	إلى	45.818	1,06	الى	101.439	1,01	ان	91.036	1,21
6	التي	36.675	0,85	**يوم**	70.696	0,70	الى	55.868	0,74
7	عن	26.627	0,62	وقال	68.142	0,68	عن	41.552	0,55
8	الذي	26.208	0,61	التي	60.260	0,60	**شينخوا**	40.680	0,54
9	ما	19.040	0,44	أن	57.935	0,58	أن	39.931	0,53
10	مع	19.036	0,44	عن	51.580	0,51	إلى	38.553	0,51
11	بعد	15.458	0,36	الذي	49.473	0,49	**اليوم**	38.455	0,51
12	هذا	15.289	0,35	مع	49.022	0,49	مع	34.444	0,46
13	لا	14.845	0,34	**رويترز**	44.321	0,44	وقال	28.879	0,38
14	**الجزائر**	13.698	0,32	بعد	44.203	0,44	بين	28.152	0,38
15	حيث	13.509	0,31	**المتحدة**	28.846	0,29	**الصين**	27.603	0,37
16	هذه	12.938	0,30	هذا	25.920	0,26	التي	27.336	0,36
17	لم	11.819	0,27	انه	25.443	0,25	خلال	23.333	0,31
18	كان	10.751	0,25	ما	24.341	0,24	**العام**	20.193	0,27
19	بين	10.341	0,24	**عام**	24.147	0,24	ما	18.676	0,25
20	كل	10.287	0,24	لا	23.817	0,24	هذا	18.647	0,25
21	**الوطني**	9.027	0,21	بين	23.599	0,23	**عام**	18.598	0,25
22	كما	8.932	0,21	**العام**	21.267	0,21	بعد	18.344	0,24
23	أو	8.847	0,20	قال	20.425	0,20	**المتحدة**	17.897	0,24
24	قبل	8.702	0,20	قبل	20.383	0,20	**يوم**	17.586	0,23
25	خلال	8.657	0,20	**دولار**	19.706	0,20	**رئيس**	17.445	0,23
26	قد	7.656	0,18	لكن	19.544	0,19	إن	17.031	0,23
27	ذلك	7.518	0,17	**الماضي**	19.490	0,19	**الرئيس**	16.118	0,21
28	وهو	7.495	0,17	لم	18.784	0,19	هذه	15.927	0,21
29	أنه	7.178	0,17	**الرئيس**	18.364	0,18	انه	15.881	0,21
30	غير	6.524	0,15	وقالت	18.359	0,18	الذي	15.771	0,21

Tableau 6-4. Distribution des 30 mots les plus fréquents dans la presse arabe.

3 Analyse lexicale dans le texte arabe

Pour le prétraitement lexical des trois corpus, nous appliquons les trois variantes de notre analyseur morphologique (*BBw0, BBw1 et BBw2*) telles quelles sont décrites dans (Chapitre 5 :7). Pour la comparaison, nous utilisons l'algorithme *ISRI* (Chapitre 5 :6.4) du stemming léger proposé par [Taghva et al., 2005] ainsi que deux variantes de l'algorithme *Khoja* de racinisation (Chapitre 5 :6.5). La première, *Khoja0*, représente la version originale telle qu'elle est décrite dans [Khoja et Garside, 1999] en ne gardant que les racines reconnues. La seconde variante, *Khoja1*, permet de rajouter les mots non reconnus dans le vocabulaire d'indexation. Ceci nous permettra de considérer les mots non inclus dans le dictionnaire de *Khoja*, conçu depuis 1999. Par ailleurs nous utilisons le texte brut (après normalisation) sans stemming comme référence dans les tests préliminaires. Afin d'aligner les différentes comparaisons, nous utilisons la même liste des mots vides dans les différents algorithmes. Notre liste et constituée de 575 mots à supprimer à la fin de la phase de normalisation.

3.1 Dimension du vocabulaire

Nous présentons, d'abord en Tableau 6-11, la dimension des vocabulaires relatifs à chaque algorithme de prétraitement. Comme constat principal, l'algorithme *ISRI* génère le vocabulaire le plus volumineux dans les trois corpus et qu'il ne réduit le vocabulaire brut que d'environ 80%. Par opposition, l'analyse morphologique par *Khoja* et *BBw*, réduit considérablement la dimension de l'espace des termes en unifiant les mots ayant la même racine. En ne retenant que les termes reconnus dans leurs dictionnaires, les variantes *Khoja0* et *BBw0*, génèrent les dictionnaires les plus faibles.

	Brut	**ISRI**	**Khoja0**	**Khoja1**	**BBw0**	**BBw1**	**BBw2**
Ech-11k	179.225	144.442	3.172	22.411	17.558	60.148	48.789
Rtr-41k	183.510	150.977	3.153	38.644	17.020	73.382	62.843
Xnh-36k	175.410	140.192	3.089	10.712	11.352	25.835	22.456

Tableau 6-5. La dimension des vocabulaires selon différents algorithmes de stemming.

En comparant à l'algorithme de racinisation de *Khoja0*, il est clair que l'analyseur morphologique *BBw0* génère un dictionnaire plus étendu. Trois causes peuvent expliquer ce décalage :

1. La racinisation est plus agressive que la lemmatisation, puisqu'une racine est plus abstraite qu'un lemme,
2. Les ressources lexicales de *Khoja* sont moins riches que celles de *Buckwalter* (*BBw*),
3. L'analyseur *BBw* peut produire plus qu'une entrée d'index pour un mot.

Cependant, nous constatons un autre écart considérable en retenant les mots non reconnus par les variantes d'analyse morphologiques (*Khoja1*, *BBw1* et *BBw2*). Ceci peut être expliqué par le manque dans les ressources linguistiques utilisées par *Khoja* et *Buckwalter* et qui doivent être périodiquement mises à jour. Nous avons prévu la rencontre d'une telle limite en proposant d'analyser les trois variantes *BBwX* lors de la discussion des causes d'*aucune solution* de lemmatisation dans (Chapitre 5 :7.2). Afin d'estimer la perte causé par cette insuffisance, nous calculons le taux des mots non reconnu par rapport à tous les mots arabes contenus dans les trois corpus. Les résultats présentés dans Figure 6-4 montre que les ressources linguistiques utilisées par *BBw* sont plus complètes que celles de *Khoja*.

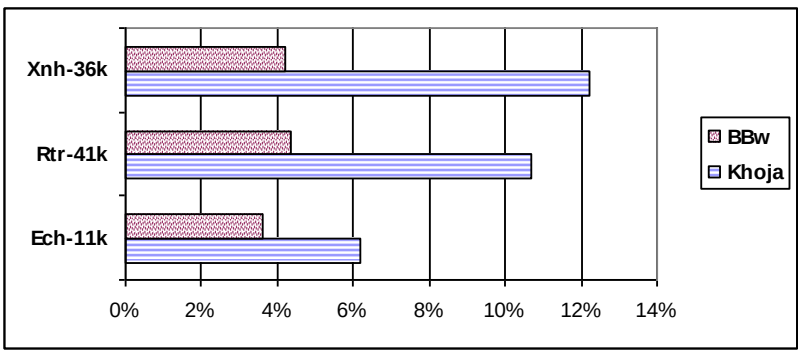


Figure 6-4. Taux des mots non reconnus par les analyseurs *Khoja* et *BBw*.

3.2 Ambigüité lexicale

Nous complétons notre analyse par l'appréciation de l'ambigüité lexicale dans la presse arabe. Notons que seul notre analyseur *BBw* qui prenne en compte les solutions multiples de lemmatisation d'un graphème. Nous commençons par analyser la multiplicité des solutions de lemmatisation avec l'algorithme *BBw2*. Nous constatons dans Tableau 6-6 que la multiplicité des résultats d'analyse peut atteindre 6 solutions possibles mais que la majorité (~90%) des mots arabes utilisés dans les trois corpus fournissent une solution unique.

Corpus	**Ech-11k**		**Rtr-41k**		**Xnh-36k**	
Multiplicité	**# mots**	**%**	**# mots**	**%**	**# mots**	**%**
1	3.021.314	90,42	7.092.343	89,86	5.532.178	90,73
2	287.168	8,59	694.572	8,80	495.599	8,13
3	30.615	0,92	95.202	1,21	64.117	1,05
4	2.174	0,07	9.179	0,12	5.612	0,09
5	194	0,01	1.051	0,01	146	0,00
6	0	0,00	1	0,00	0	0,00

Tableau 6-6. Analyse de la multiplicité de solution avec l'analyseur *BBw2*.

Par ailleurs et selon la mesure définie par l'équation (5-7) dans (Chapitre 5 :7.3), nous estimons l'ambigüité lexicale dans le langage des collections étudiées par le calcul du degré de confusion $C(S|BBwX)$ avec (S = *Ech-11k*, *Rtr-41k*, *Xnh-36k*). Pour chaque variante *BBwX*, nous présentons dans Tableau 6-7 le degré de confusion (*Conf-deg*) et le nombre maximal de solutions multiples engendrées pour un mot donné (*Max-mul*).

Corpus	**Ech-11k**		**Rtr-41k**		**Xnh-36k**	
Analyseur	**Conf-deg**	**Max-mul**	**Conf-deg**	**Max-mul**	**Conf-deg**	**Max-mul**
BBw0	1,118	5	1,129	6	1,121	5
BBw1	1,146	6	1,155	6	1,141	6
BBw2	1,106	5	1,116	5	1,105	5

Tableau 6-7. Degré de confusion dans le langage de trois collections arabes.

Les résultats de cette analyse nous indiquent une réalité impressionnante :

Bien que le lexique arabe induise une haute complexité morphologique (flexionnelle et dérivationnelle) augmentée par l'agglutination et la non-vocalisation, la proportion des mots ambigus dans un texte réel n'excède pas les 10%. En plus, le degré de confusion relevé des articles de presse en utilisant l'analyseur *BBw* ne dépasse pas 1,16 au pire des cas.

Ceci indique que malgré que notre approche *BBw* de stemming inclue toutes les solutions multiples d'un graphème arabe, elle préserve la richesse sémantique du texte arabe sans ambigüité lexicale pénalisante.

3.3 Evaluation des méthodes de stemming

Nous proposons, dans cette section, d'évaluer les méthodes de stemming du texte arabe selon les mesures décrites dans (Chapitre 5 :5) en utilisant nos trois collections. Nous rappelons que pour des raisons d'équité des évaluations, la liste des mots vides utilisée dans notre analyseur *BBw* a été généralisée dans les autres algorithmes.

3.3.1 Facteur de compression d'index

Comme définie dans l'équation (5-1), nous calculons le facteur de compression d'index (*ICF*) des six variantes algorithmiques sur les trois corpus (*Ech-11k*, *Rtr-41k*, *Xnh-36k*). Nous listons les résultats de calcul de l'*ICF* dans Tableau 6-8 mais pour une lecture généralisée, nous utilisons la moyenne estimée sur les trois corpus dans Figure 6-5. Nous constatons que les analyseurs morphologiques (*Khoja0* et *BBw0*) ont réalisé la meilleure compression. Ceci est dû surtout au fait que ces deux variantes ne retiennent que les formes reconnues dans leurs dictionnaires.

Algorithme **Corpus**	**ISRI**	**Khoja0**	**Khoja1**	**BBw0**	**BBw1**	**BBw2**
Ech-11k	0,194	**0,982**	0,875	**0,902**	0,664	0,728
Rtr-41k	0,177	**0,983**	0,789	**0,907**	0,600	0,658
Xnh-36k	0,201	**0,982**	0,939	**0,935**	0,853	0,872

Tableau 6-8. Comparaison du facteur *ICF* du stemming de 3 corpus arabes.

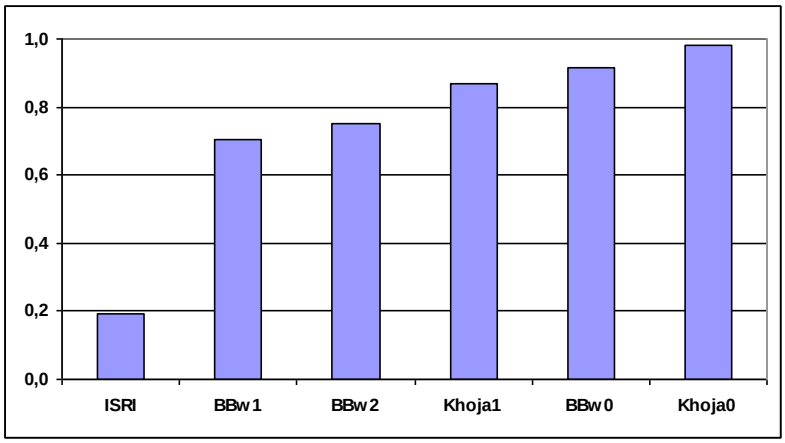


Figure 6-5. Estimation de la moyenne du facteur *ICF* pour 6 algorithmes de stemming arabe.

Même en considérant les mots non reconnus dans les autres variantes (*Khoja1*, *BBw1* et *BBw2*), le facteur de compression reste assez élevé par rapport à celui du stemming léger avec *ISRI*. Par ailleurs, nous enregistrons une performance légèrement élevée des variantes de *KhojaX* par rapport à celles de *BBwX*. L'explication évidente d'un tel phénomène réside dans la nature elle-même de l'analyse morphologique qui génère avec *Khoja* des formes plus abstraites (racine) que celles produites par *BBwX* (lemme non vocalisé).

3.3.2 Erreurs de stemming

Afin d'appliquer les mesures de *Paice* pour estimer les erreurs de stemming, il est impératif d'avoir un ensemble assez représentatif de groupes-concepts. La seule étude que nous avons identifiée pour évaluer les erreurs de stemming dans le texte arabe est celle citée dans [Al-Shammari, 2010]. Malheureusement, ni les corpus de test ni l'ensemble des groupes-concepts utilisés dans l'étude d'Al-Shammari ne sont disponibles. Rappelons que l'auteur a conçu un échantillon modéré de 81 groupes-concepts où chaque groupe contient environ 5 mots sémantiquement et morphologiquement liés.

Dans notre étude, nous avons conçu un ensemble plus large à partir des collections réelles. Notre approche consistait à sélectionner les (89.260) titres des articles de presse dans les trois corpus (*Ech-11k*, *Rtr-41k*, *Xnh-36k*). Après la suppression des mots vides et un

groupement automatique préliminaire, nous avons retenu les groupes contenant au moins 10 mots au moins. Ensuite nous avons révisé les groupes résultants minutieusement avec un expert arabe avant de retenir un ensemble de 13.142 mots répartis sur 689 groupes. Nous présentons dans Tableau 6-9 un extrait de ces groupes.

Mot	**Taille du groupe**	**Groupe-concept**
بنك (Banque)	15	بنك وبنك ببنك بنكا بالبنك بنكية بنكان لبنكين بنكين البنكين البنكان البنكية كبنك بنكي البنكي
بيع (Vente)	22	بيع والبيع لبيع البيع يبيع للبيع وبيع لبيعها ببيع بيعت ويبيع بيعة ستبيع يبيعون يبيعه بيعها سيبيع بيعه أبيع وبيعهم بيعي ببيعه
سبب (Cause)	12	سبب بسبب السبب سببته سببها سببه سببا يسبب سببت بسببه وسبب وسببها
سعر (Prix)	15	سعر السعر لسعر وسعر سعرية بسعر يسعر سعرا السعري بالسعر السعرية للسعر سعره سعرها سعري
سفينة (Bateau)	15	السفينة سفينة سفينتين بسفينة سفينتي للسفينة سفينتها بالسفينة سفينتهم السفينتين والسفينة لسفينتين لسفينة وسفينة سفينتان

Tableau 6-9. Extrait de l'ensemble des groupes-concepts arabes.

En plus de l'algorithme *ISRI*, nous limitons notre évaluation sur les analyseurs morphologiques *Khoja0* et *BBw0*. L'évaluation des autres variantes étant équivalente puisque toutes les entrées de l'ensemble des groupes-concepts sont reconnues dans les ressources linguistiques de *Khoja* et *BBw*. Les résultats d'évaluation de *Paice*, calculés selon les équations définies dans (Chapitre 5 :5.2), sont présentés dans Tableau 6-10.

Algorithme	**OI (sur-stemming)**	**UI (sous-stemming)**	**SW (poids)**
ISRI	$0{,}019 \times 10^{-3}$	0,968	$0{,}019 \times 10^{-3}$
Khoja	$1{,}452 \times 10^{-3}$	0,098	$14{,}87 \times 10^{-3}$
BBw	$0{,}006 \times 10^{-3}$	0,060	$0{,}096 \times 10^{-3}$

Tableau 6-10. Evaluation de *Paice* pour trois méthodes de stemming arabe.

Les résultats montrent qu'avec notre analyseur *BBw* on obtient les plus faibles indices d'erreurs de stemming (*OI* et *UI*). La performance est significative en comparant aux indices d'erreurs obtenus pour *Khoja* et *ISRI*. Rappelons que l'analyseur *BBw* peut générer des solutions multiples pour les formes ambigües. Théoriquement, ce fait augmente les erreurs de sous-stemming en dispersant un mot sur deux groupes-concepts au moins. Néanmoins, les résultats empiriques prouvent la robustesse de notre approche en améliorant la distinction des sens d'une même forme.

4 Analyse sémantique dans la presse arabe

Nous illustrons dans cette section quelques résultats de la modélisation par thèmes latents des articles de presse dans les trois corpus étudiés. Le prétraitement lexical des articles a été réalisé par analyse morphologique selon la variante *BBw2*. Le modèle LDA a été inféré ensuite par échantillonnage de Gibbs pout différents nombre de thèmes.

4.1 Détermination du nombre de thèmes

L'apprentissage du modèle LDA nécessite le choix préalable d'un nombre de thèmes adapté à la nature de la collection et aux objectifs de l'analyse elle-même. Nous avons vu dans (Chapitre 4 :7.1) que choix est généralement effectué de façon arbitraire et qu'il n'existe pas de méthode formelle approuvée pour le détecter. Seule la performance des tâches, telles que la recherche ou la classification, utilisant le modèle généré peut nous indiquer le nombre

approprié de thème. Nous testons dans cette section la mesure combinée *BKL*, basée sur la divergence de *Kullback-Leibler*, telle qu'elle a été définie dans (Chapitre 4 :7.4).

A cet effet, nous utilisons les trois corpus (*Ech-4000*, *Rtr-5251*, *Xnh-4500*) prétraités par deux algorithmes de stemming (*Khoja0* et *BBw0*). Toutes les expérimentations sont réalisées pour 13 valeurs, du nombre de thèmes, réparties sur la plage [2, 700]. Ceci implique que s'il existe un nombre approprié de thèmes T^* pour l'apprentissage d'un modèle LDA dans une collection donnée, et que nous trouvons la mesure *BKL* optimale pour une valeur T_i, nous déduisons que $T_{i-1} < T^* < T_{i+1}$.

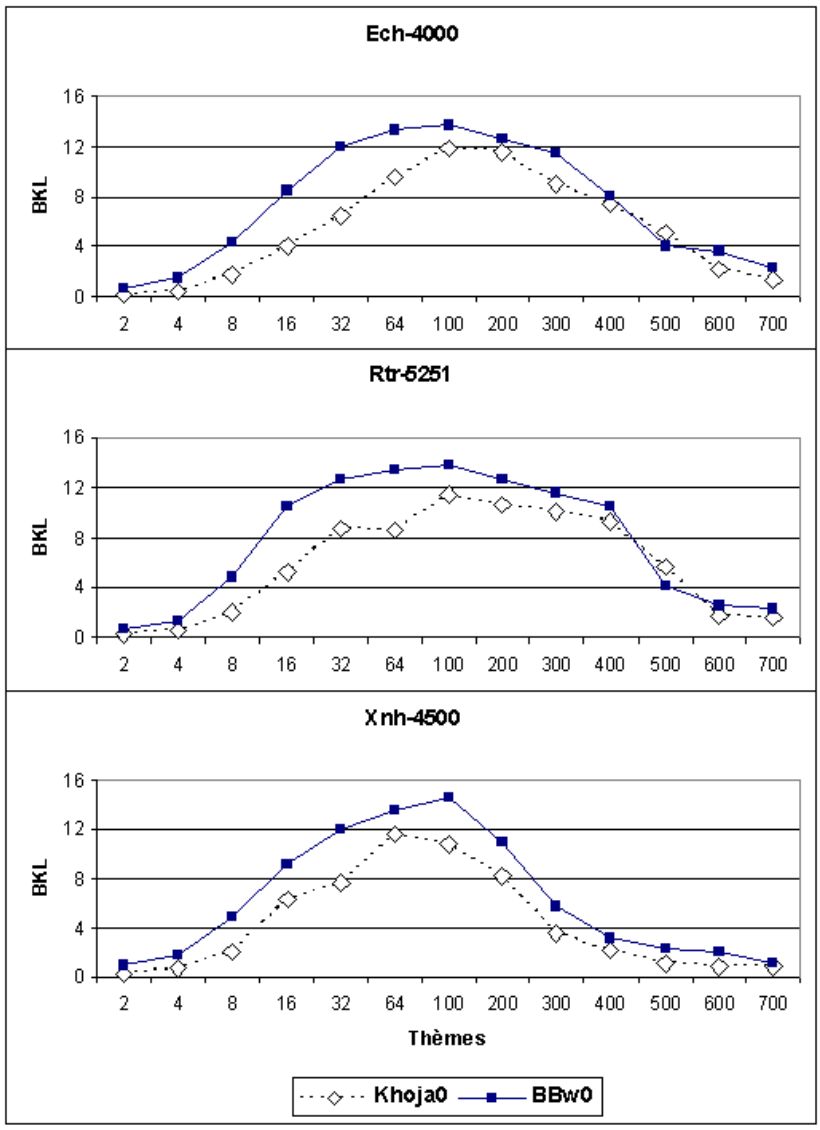


Figure 6-6. Mesure *BKL* en fonction du nombre des thèmes.

Les résultats des expérimentations sont tracés dans Figure 6-6, où la mesure *BKL* est évalué pour chaque nombre de thèmes et pour deux espaces de mots dans chaque corpus. Nous constatons tout d'abord qu'en utilisant l'analyse du texte arabe par lemme (*BBw0*), la mesure *BKL* du modèle LDA est un peu plus améliorée que celle avec la racinisation (*Khoja0*). Sachant que cette mesure interprète un compromis entre la stabilité des thèmes inférés et leur capacité discriminative, nous pouvons déduire que le stemming du texte arabe par lemme est plus adapté à la modélisation par thème que celui de la racinisation.

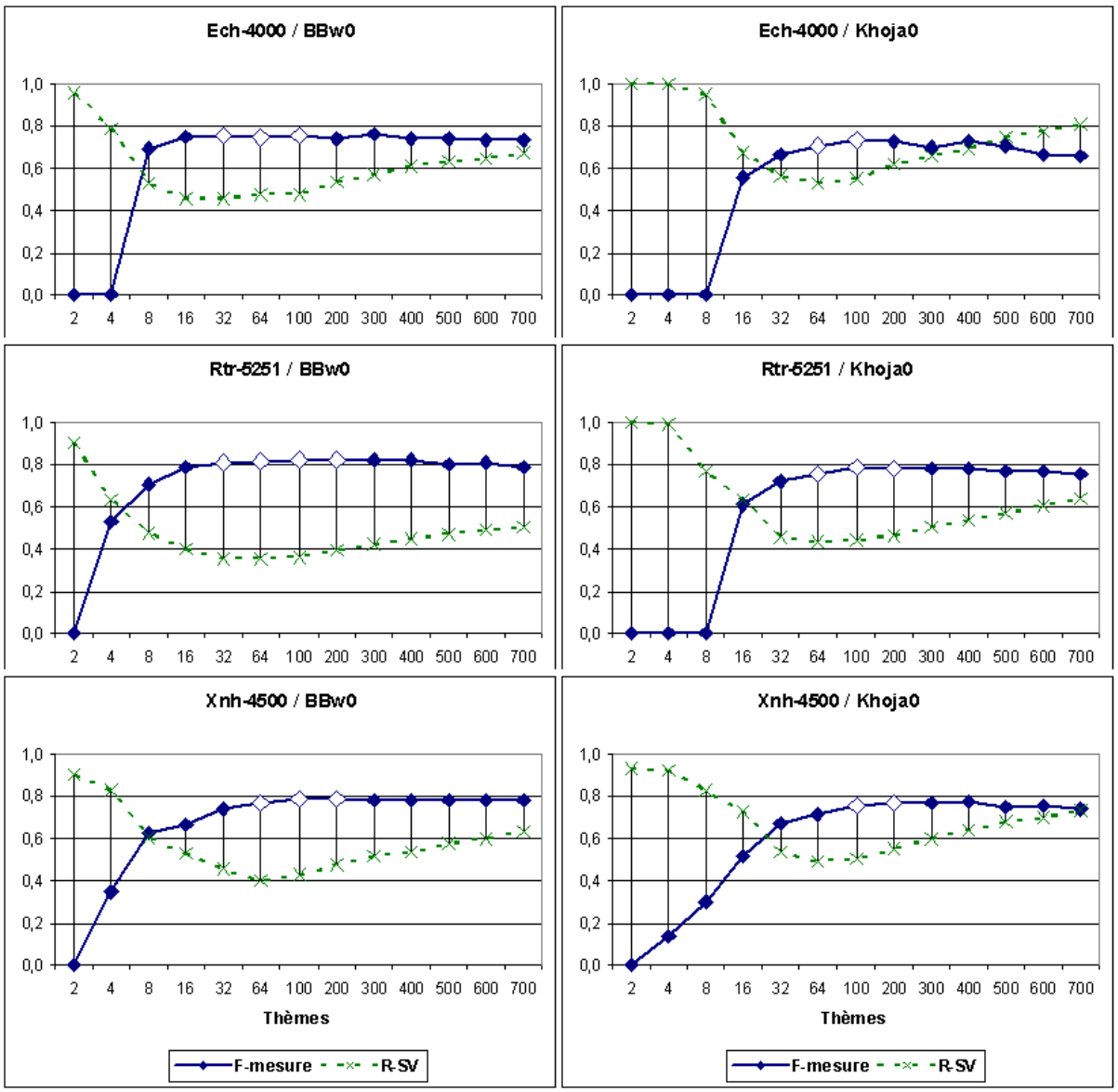


Figure 6-7. Variation des performances de classification en fonction du nombre des thèmes.

Par ailleurs, la mesure *BKL* nous renseigne sur la plage appropriée du nombre de thèmes à apprendre, par LDA, dans chaque corpus. Cette plage, qui varie selon la nature du corpus lui-même et la méthode de stemming appliquée, est confirmée par le calcul des performances de la classification SVM (voir Figure 6-7). Dans cette dernière nous reportons la performance de classification dans l'espace des thèmes en termes de *F-mesure* calculée par macro-précision et macro-rappel selon la description dans Chapitre 2 :7.4.1). La mesure *r-SV* nous indique une meilleure capacité de généralisation lorsqu'elle est minimale. Notons que cette mesure comme critère de préférence entre des valeurs élevées et rapprochées de *F-Mesure* (voir Chapitre 2 :7.4.3).

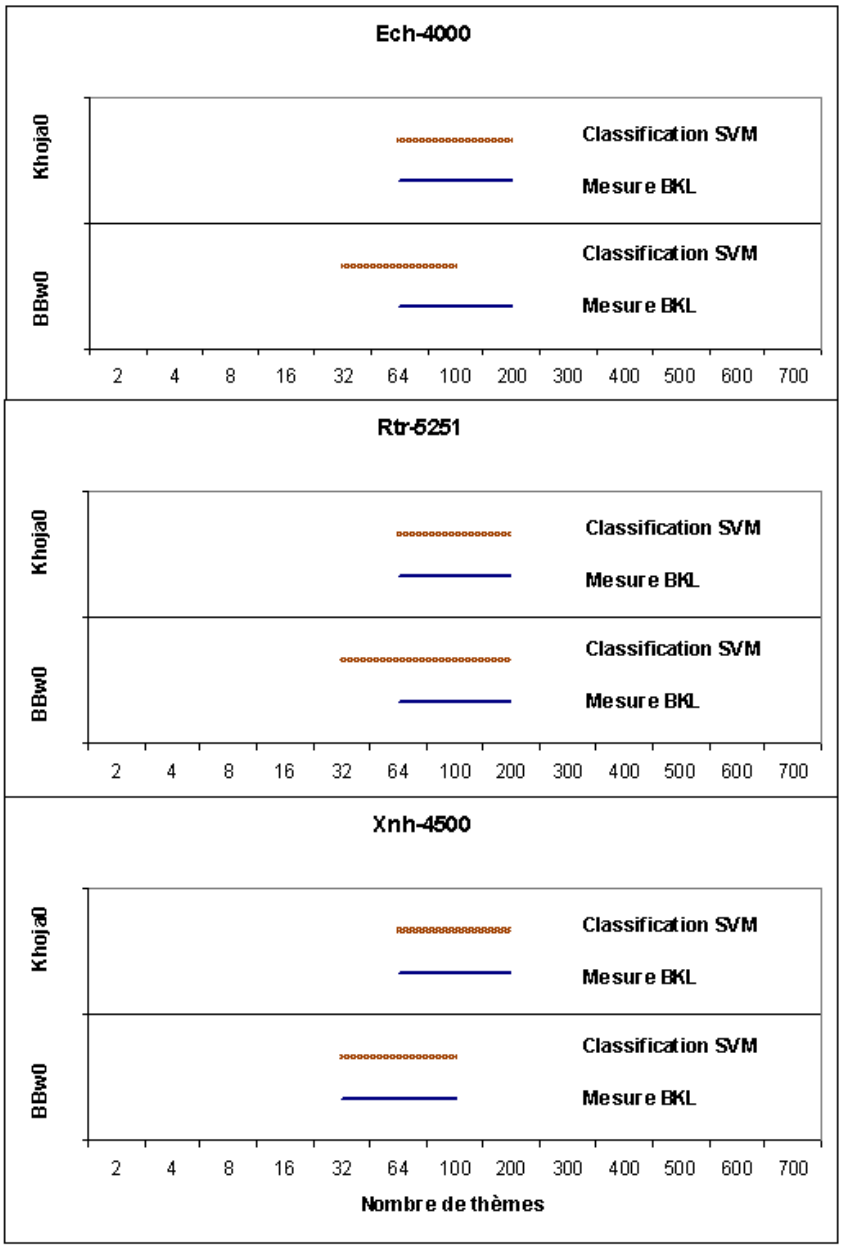


Figure 6-8. Comparaison des mesures de performance (*BKL* / SVM) pour la détermination du nombre de thèmes dans la modélisation LDA.

Nous récapitulons dans Figure 6-8 un graphe comparatif entre les performances obtenues par notre mesure combinée *BKL* et celles de la classification SVM pour déterminer l'intervalle approprié du nombre de thèmes dans la modélisation LDA. Une lecture rapide dans ce graphe nous montre qu'il existe une concordance presque complète entre les deux mesures d'évaluation des performances d'une modélisation par thème. On peut déduire que la stabilité et la capacité discriminative d'un modèle LDA améliore la classification par SVM dans l'espace des thèmes.

4.2 Analyse par matrice de confusion

Du moment où le modèle LDA attribue pour chaque document de la collection une distribution de thèmes, nous pouvons calculer la distribution d'une catégorie en cumulant et en normalisant la distribution de tous les articles de la même catégorie. Nous pouvons

concevoir ainsi une matrice de confusion thème×catégorie qui donne la distribution groupée de chaque catégorie sur les thèmes inférés par l'apprentissage LDA. Dans (Tableau 6-11, Tableau 6-12, Tableau 6-13) nous regroupons la distribution des articles sur les (T=8) thèmes latents selon leurs catégories d'édition originale.

Catégorie **Thème**	**Algérie**	**Economie**	**Sport**	**Monde**	**Art**	**Culture**	**Société**	**Religion**
Economie	**18,8**	**62,6**	1,9	4,6	3,7	4,2	**15,3**	4,3
Affaires judiciaires	**28,5**	3,9	1,8	7,8	6,3	3,5	**28,8**	4,5
Sport local	1,4	1,0	**45,3**	0,9	2,9	1,5	1,2	1,0
Moyen-Orient	5,7	4,4	1,5	**66,9**	5,5	6,5	1,3	**12,2**
Equipe algérienne de football	1,0	1,2	**35,2**	1,3	2,7	1,9	0,8	0,9
Religion	**10,8**	3,2	6,6	9,0	**27,7**	**20,7**	**38,4**	**46,2**
Art et culture	3,8	2,6	2,5	2,8	**48,2**	**51,2**	3,5	**24,1**
Politique algérienne	**30,0**	**21,0**	5,1	6,6	3,2	**10,5**	**10,7**	7,0

Tableau 6-11. Distribution des catégories d'articles *d'Echorouk* sur 8 thèmes latents.

Catégorie **Thème**	**Moyen-Orient**	**Economie**	**Monde**	**Sport**	**Divertissement**	**Science**
Actions militaires (USA, Iraq, Afghanistan)	**25,4**	1,8	21,1	0,8	2,8	2,0
Football	0,6	0,6	0,9	**52,3**	4,9	1,2
Economie	3,0	**75,2**	2,8	1,4	5,4	5,9
Santé et science	**12,3**	5,6	**17,4**	4,0	**50,3**	**59,6**
Nucléaire iranien	**12,4**	**11,4**	**28,9**	3,3	6,9	**26,0**
Politique Iranienne	**21,0**	3,8	**26,9**	3,2	**22,2**	4,1
Tennis	0,3	0,5	0,7	**34,4**	1,6	0,5
Moyen-Orient	**25,0**	1,1	1,4	0,6	6,0	0,8

Tableau 6-12. Distribution des catégories d'articles de *Reuters* sur 8 thèmes latents.

Catégorie **Thème**	**Chine**	**Culture-Education**	**Economie**	**Moyen-Orient**	**Science**	**Sport**	**Tourisme Ecologie**	**Monde**
Moyen-Orient	1,5	3,3	2,1	**32,9**	1,7	1,6	1,8	3,2
Guerre contre le terrorisme	6,0	2,1	0,8	**15,5**	2,1	1,6	1,1	**26,2**
Economie	6,4	5,9	**55,1**	1,9	8,0	1,4	**19,1**	2,6
Nucléaire iranien	8,0	5,2	7,7	7,1	7,7	5,5	**10,6**	**39,8**
Football	4,2	**38,1**	2,0	8,1	3,7	**76,1**	8,2	2,3
Coopération chinoise	**40,4**	**24,6**	**19,8**	5,4	9,5	4,9	**31,7**	8,3
Chine interne	**31,9**	**11,5**	6,6	3,9	**22,1**	3,4	**24,0**	**15,1**
Région arabe	1,6	9,2	6,0	**25,2**	**45,3**	5,6	3,4	2,5

Tableau 6-13. Distribution des catégories d'articles de *Xinhua* sur 8 thèmes latents.

Notons que les intitulés des thèmes sont attribués par appréciation humaine selon les termes pertinents dans chaque distribution de thème. En fixant un seuil minimal de 10%, nous pouvons constater, par exemple, que les thèmes principaux développés dans la catégorie *Sport* de *Reuters* concernent le *Football* et le *Tennis*. Le sujet principal développé dans la catégorie *Monde* d'*Echorouk* concerne le *Moyen-Orient*.

Nous pouvons approfondir notre analyse thématique en réalisant un apprentissage LDA pour un nombre de thèmes plus élevé. Cette analyse est limitée aux articles du corpus et concerne seulement la période 2007-2009. A titre d'exemple, en fixant le nombre de thèmes à (T=16), nous avons obtenu, par apprentissage du modèle LDA dans les trois collections étudiées, les thèmes listés dans Tableau 6-14.

Thème	**Ech-11k**	**Rtr-41k**	**Xnh-36k**
1	Economie algérienne	Marchés boursiers	Culture
2	Affaires judiciaires	Palestine	USA et ONU
3	Algérie interne	Santé	Iraq
4	Affaires internationales d'Algérie	Football	Nucléaire iranien
5	Affaires sécuritaires	Iran	Economie
6	Football algérien	Economie	Affaires d'anti-terrorisme
7	Iraq et Iran	Somalie and Soudan	Palestine
8	Palestine	Jeux olympiques de Pékin	Politique de la chine
9	Equipe algérienne de football	Pétrole and gaz	Football
10	Art et culture	Affaires judiciaires	Economie chinoise
11	Termes généraux	Football européen	Marchés boursier
12	Santé	Actions militaires en Iraq et Afghanistan	Région arabe
13	Education	Tennis	Séisme en chine
14	Organismes de sport	Nucléaire iranien	Transport et voyage
15	Islam	Moyen-Orient	Santé
16	Projets gouvernementaux	Politique des USA	Coopération chinoise-internationale

Tableau 6-14. Liste des 16 thèmes latents dans la presse arabe durant 2007-2009.

Nous constatons que les thèmes générés deviennent plus fins et relèvent des principaux sujets traités par la presse durant la période 2007, 2008 et 2009. Néanmoins, ce qui peut être plus intéressant est de calculer la matrice de confusion thème×catégorie afin d'apprécier la distribution des 16 thèmes sur les catégories publiées par l'éditeur. Nous présentons dans Tableau 6-15 la distribution des catégories de *Reuters* sur les principaux thèmes.

Catégorie **Thème**	**Moyen-Orient**	**Economie**	**Monde**	**Sport**	**Divertissement**	**Science**
Marchés boursiers	0,6	**30,8**	1,1	0,5	1,7	2,0
Palestine	**18,1**	0,6	0,8	0,4	3,5	0,5
Santé	7,0	2,3	**11,6**	1,7	**30,4**	**50,0**
Football	0,3	0,3	0,3	**25,0**	1,0	0,5
Iran	3,2	0,7	8,9	0,6	1,9	1,1
Economie	2,0	**29,4**	1,2	1,0	4,4	2,1
Somalie and Soudan	**13,5**	1,0	5,5	0,5	1,4	2,0
Jeux olympiques de Pékin	0,4	1,1	2,2	**10,7**	**13,3**	2,8
Pétrole and gaz	1,8	**18,0**	1,7	0,4	1,7	5,8
Affaires judiciaires	**11,5**	1,6	**10,5**	1,5	**19,6**	3,0
Football européen	0,6	0,6	2,2	**23,4**	2,5	1,1
Actions militaires en Iraq et Afghanistan	**12,4**	0,7	**15,6**	0,5	2,3	1,4
Tennis	0,2	0,4	0,4	**29,8**	0,7	0,4
Nucléaire iranien	3,7	1,7	**10,8**	0,6	1,1	2,1
Moyen-Orient	**11,9**	2,1	9,4	0,8	5,1	1,2
Politique des USA	**12,7**	8,8	**17,7**	2,5	9,5	**24,0**

Tableau 6-15. Distribution des catégories d'articles de *Reuters* sur 16 thèmes latents.

En ne considérant que les valeurs supérieures à 10%, nous découvrons par une simple lecture que les thèmes pertinents en *sport* concernaient le *football*, le *tennis* et les *jeux olympiques de Pékin*.

4.3 Analyse du contexte d'un terme

Il est intéressant de détecter les différents contextes d'un terme donné dans une collection de texte. Avec un apprentissage LDA sur une collection représentative du langage, cette analyse permet de désambigüiser un terme qui porte plus qu'un sens. Nous commençons par un exemple dans, Tableau 6-16, qui liste les thèmes où figure le terme (*[mAl]* مال**)** avec une probabilité supérieure à 1%. Pour un modèle LDA à 100 thèmes appris des articles d'*Echorouk*, nous avons pu détecter 4 thèmes latents fortement liés au terme (*[mAl]* مال).

En plus, il est possible de tracer une relation sémantique contextuelle entre les termes les plus pertinents dans la même distribution de thème. A titre d'exemple, nous montrons dans Figure 6-9 les différents contextes liés au terme ([mAl, *money*]**مال**) en fixant un seuil de probabilité minimal de 1,5%. La probabilité attribuée à un terme donné pour chacun des thèmes est inscrite en pourcentage sur la liaison entre chaque nœud de thème et ces termes pertinents. Ceci peut nous permet de tracer un graphe sémantique autour de chaque terme du langage.

Thème 9 **Valeurs monétaires**	**Thème 42** **Opérations d'Al-Qaida**	**Thème 89** **Institutions financières**	**Thème 100** **Projets gouvernementaux**
[mlywn] مليون	[jmAE] جماع	**[mAl] مال**	[wzyr] وزير
[mAl] مال	[tnZym] تنظيم	[bnk] بنك	[Hkwm] حكوم
[qym] قيم	[slf] سلف	[jzA}r] جزائر	[m$rwE] مشروع
[mlyAr] مليار	[mslH] مسلح	[AqtSAd] اقتصاد	**[mAl] مال**
[mblg] مبلغ	**[mAl] مال**	[bnwk] بنوك	[jdyd] جديد
[mnH] منح	[dEw] دعو	[lbn] لبن	[m&ss] مؤسس
[dfE] دفع	[Ebd] عبد	[qrwD] قروض	[wTn] وطن
[dj] دج	[qAEd] قاعد	[EAlm] عالم	[wzAr] وزار
[qdr] قدر	[qtAl] قتال	[Azm] ازم	[dwl] دول
[AmwAl] اموال	[sAbq] سابق	[dwl] دول	[qTAE] قطاع
[mlAyyn] ملايين	[HTAb] حطاب	[m&ss] مؤسس	[m$AryE] مشاريع
[Awrw] اورو	[AxtTAf] اختطاف	[Eml] عمل	[Emwm] عموم
مقابل [mqAbl]	[qyAd] قياد	[mlyAr] مليار	[AjrA'] اجراء
[HsAb] حساب	[Eml] عمل	[Hkwm] حكوم	[Eml] عمل
[mbAlg] مبالغ	[ArhAb] ارهاب	[AstvmAr] استثمار	[Dm]ضم
[Srf] صرف	[drwdkAl] درودكال	[t>myn] تأمين	[qAnwn]قانون

Tableau 6-16. Exemple de thèmes latents parmi 100 appris du corpus *Ech-11k*.

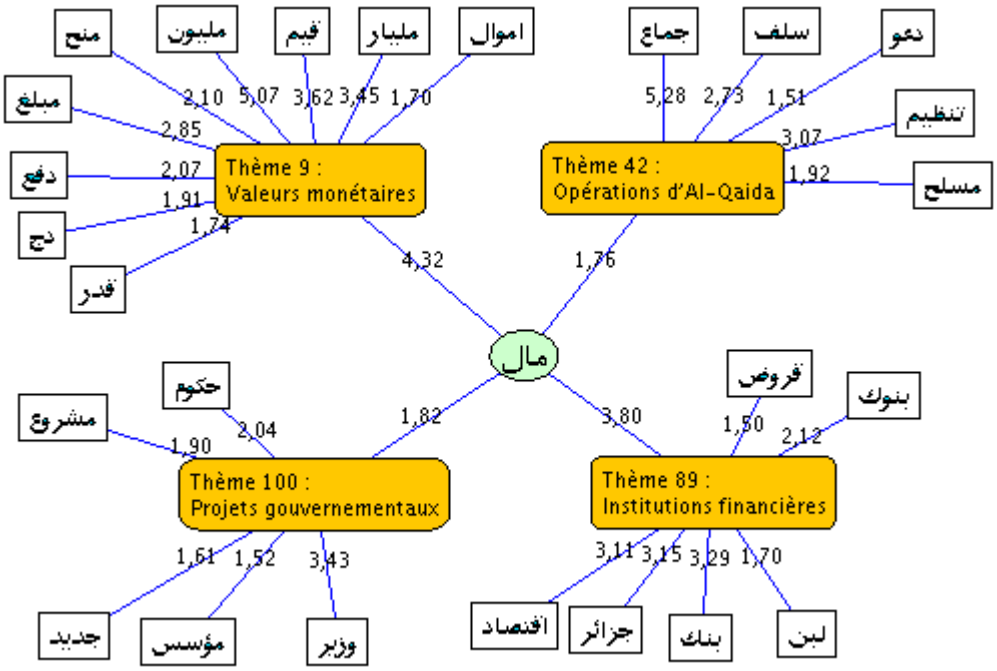


Figure 6-9. Graphe contextuel du terme (*[mAl]*مال) dans le corpus *Ech-11k* .

De manière générale, nous pouvons analyser le sens d'un terme par l'extraction de son contexte (thème dont la probabilité du terme dépasse un certain seuil). Dans (**Tableau 6-17**)

et par un apprentissage LDA sur 100 thèmes, nous présentons les différents contextes découverts pour le terme (*[slAm]*سلام) et (*[mAl]* مال) dans les trois corpus.

Terme	**Ech-11k**	**Rtr-41k**	**Xnh-36k**
[slAm] سلام	Ligue arabe	Acteurs politiques en Somalie	Missions de l'ONU
	Dialogue palestinien	Prix Nobel	Efforts de la Jordanie pour la paie
		Négociations du Moyen-Orient	Négociations du Moyen-Orient
		Evénements au Soudan	Mines de charbon en Australie
			Evénements au Soudan
[mAl] مال	Valeurs monétaires	Investissements des pays du golfe arabe	Crise financière
	Opérations d'Al-Qaida	Opérations financières	Opérations financières
	Institutions financières	Crise financière	Projets d'énergie
	Projets gouvernementaux	Investissements gouvernementaux	

Tableau 6-17. Les thèmes relatifs aux termes (*[slAm]*سلام) et (*[mAl]* مال) dans les trois corpus.

Bien que le terme *[slAm]*سلام porte deux sens (*salutation* et *paie*), l'apprentissage LDA nous renseigne des différents contextes qui lui sont associés dans les trois corpus. Tous les thèmes LDA, présentés dans Tableau 6-17, concernent le sens *paie* sauf pour le thème "*Mines de charbon en Australie*". En revenant sur le texte brut des articles associés à ce thème, nous trouvons qu'il s'agit plutôt du terme, *[slAmh]*سلامه, mais dont l'analyse *BBw* lui a donné deux solutions : la première (*sa paie*) qui s'unifie dans sa base avec l'entrée (*[slAm]*سلام, *paie*) alors que la deuxième (*[slAmh]*سلامه, sécurité) concorde fortement avec le thème relatif aux "*mesures de sécurité dans les mines de charbon*".

5 Catégorisation des articles de presse arabe

Nous utilisons pour la tâche de catégorisation des textes la méthode SVM dans une classification multi-classe [Vapnik, 1995]. Nous avons adapté l'implémentation du package *LIBSVM*[43] afin de calculer les mesures de précision et de rappel pour chaque classe. En plus, nous retenons le ratio des vecteurs de support pour chaque apprentissage (*rSV*) et nous calculons le taux de reconnaissance en adoptant une validation croisée à 5-fold (voir Chapitre 2 :7.4). Nous nous sommes contentés par le simple noyau linéaire en fixant le paramètre de coût à 10.

Les trois collections des articles de presse sont utilisées à cet effet (comme décrites dans Tableau 6-2). Nous commençons par des sous-ensembles réduits (*Ech-4000, Rtr-5251, Xnh-4500*) pour l'évaluation préliminaire avant de passer aux collections complètes (*Ech-11k, Rtr-41k, Xnh-36k*).

[43] LIBSVM-2.89 est librement disponible à : http://www.csie.ntu.edu.tw/~cjlin/libsvm+zip

5.1 Classification dans l'espace des termes

Pour la définition de l'espace des caractéristiques (par terme) des articles de chaque collection, nous appliquons d'abord un prétraitement de stemming. Six algorithmes sont utilisés à cet effet en plus du texte brut normalisé. Deux mesures sont appliquées pour la pondération des termes dans chacun des trois corpus : *TF* et *TF.IDF* (voir Chapitre 3 :3.2).

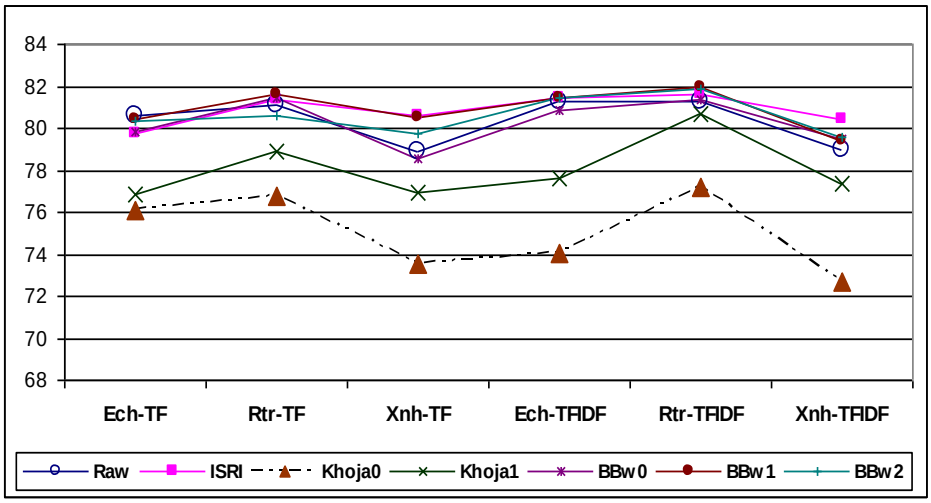


Figure 6-10. Performance de classification dans l'espace des termes de 3 corpus (*Ech-4000*, *Rtr-5251*, *Xnh-4500*).

Comme constat principal, nous retenons des résultats de classification dans Figure 6-10 que les variantes de stemming *Khoja* donnent les plus faibles taux de reconnaissance. En produisant des racines abstraites mais avec un lexique incomplet, l'algorithme *Khoja0* est le moins convenable à la catégorisation du texte arabe dans l'espace des termes. Même en rajoutant les mots non reconnus, l'amélioration de la classification n'est pas assez significative avec *Khoja1*. Cependant, nous constatons que les analyseurs *BBwX* améliorent tous la classification avec une dimension raisonnable de l'espace des caractéristiques.

Par ailleurs et bien que le stemming léger par *ISRI* produise un vocabulaire très large, il semble assez performant dans une classification SVM. Le même constat peut être enregistré en utilisant directement un texte brut.

Ceci paraît comme une contradiction chez les linguistes qui s'attendaient à une nette amélioration des performances par tout prétraitement qui unifie, plus ou moins, les graphèmes morphologiquement et sémantiquement liées. Rappelons que ce phénomène a été déjà signalé en revenant aux études relatives à l'analyse automatique du texte arabe (voir Chapitre 5 :6.2). Néanmoins, nous soulevons à ce stade de notre étude les remarques suivantes :

- Les résultats obtenus, dans cette étude ou ailleurs, sont partiels et ne traitent qu'une tâche particulière en RI mais, surtout, elles utilisent des corpus n'englobant pas toute la richesse du langage arabe [Larkey et al., 2002] [Brants et al., 2002] [Darwish et al., 2005] [Moukdad, 2006] [Said et al., 2009].
- Il est prouvé, pour l'anglais et d'autres langues, qu'il suffit pour les tâches de RI (recherche ad-hoc, classification, …etc.) d'effectuer un stemming léger sans pour autant trouver la forme correcte du mot [Manning et Schütze, 2001]. Ceci n'est pas le cas avec les tâches qui prennent en considération autres aspects syntaxiques ou sémantiques (résumé, traduction, recherche sémantique, …etc.). Dépassant le cadre de cette thèse, une large investigation pour l'analyse de l'impact du stemming sur la

performance de telles tâches est nécessaire. La présente étude tente d'en apporter quelques réponses dans les sections suivantes.

- Même si l'utilisation du texte brut peut aboutir à une performance en RI équivalente ou proche de celle obtenue par un prétraitement de stemming, la dimension élevée du vocabulaire généré peut compliquer l'inférence du modèle d'indexation dans un fond documentaire aussi large que réel tel que le Web.

5.2 Classification dans l'espace des thèmes

Avant d'appliquer la catégorisation, nous calculons la distribution des documents de chaque collection sur les thèmes latents. Nous varions le nombre de thèmes pour l'apprentissage du modèle LDA entre 2 et 700. Ceci nous permettra de détecter le nombre approprié de thèmes à utiliser pour chaque situation (source de collection et algorithme de stemming). Ensuite nous calculons, par validation croisée, le taux de reconnaissance de classification SVM dans l'espace des thèmes de chaque corpus. Nous reportons, dans Tableau 6-18 les performances de classification pour les trois collections (*Ech-4000*, *Rtr-5251*, *Xnh-4500*) selon six variantes de stemming. En parcourant les résultats du Tableau 6-18, il est intéressant de noter les constatations suivantes :

- ***Entre algorithmes de stemming :*** Comme pour la classification dans l'espace des termes, la catégorisation dans l'espace des thèmes a eu des performances aussi meilleures avec un texte brut qu'avec une analyse morphologique par lemme (*BBwX*). Les variantes *KhojaX* ont conduit à des performances moins bonnes par rapport aux autres algorithmes de stemming. Cependant, l'analyse morphologique par *BBwX* permet de réaliser une performance satisfaisante de classification même avec un nombre de thèmes réduit ($4<T<64$).
- ***Entre thème et terme :*** Bien que la classification soit réalisée dans un espace de thèmes trop réduit, le taux de classification a enregistré une amélioration générale par rapport à celui dans l'espace des termes (*TF* et *TF.IDF*).
- ***Entre nombres de thèmes :*** Nous pouvons détecter sommairement une plage de nombre de thèmes entre 64 et 600 avec laquelle le taux de classification est maximal pour chaque situation (valeurs en gras dans Tableau 6-18). Ceci peut nous renseigner du paramétrage adéquat dans l'apprentissage du modèle LDA afin de définir efficacement l'espace de caractéristiques (thèmes).

Afin d'analyser de façon plus profonde les performances de classification, nous proposons d'évaluer la catégorisation de la totalité de chaque corpus en utilisant la validation croisée par *5-fold* pour calculer le taux de reconnaissance. La classification SVM est appliquée dans l'espace de thèmes résultant de la modélisation LDA pour T= 64, 100, 200. En plus, nous comparons ces résultats à celles de la classification dans l'espace des termes pondérés par simple fréquence (*TF*) ou par normalisation avec la fréquence inversée globale (*TF.IDF*). Les trois corpus (*Ech-11k*, *Rtr-41k*, *Xnh-36k*) sont prétraités par deux variantes d'analyse morphologique incluant les mots non reconnus (*Khoja1* et *BBw2*).

Corpus	#Thèmes	Brut	ISRI	Khoja0	Khoja1	BBw0	BBw1	BBw2
Ech-4000	**4**	21,9	20,8	17,3	56,4	56,2	60,3	56,9
	8	43,6	41,1	33,9	72,0	73,6	73,0	74,5
	16	70,9	70,4	63,8	75,9	77,8	78,0	78,7
	32	76,3	77,4	71,7	76,5	79,0	78,9	79,1
	64	79,2	79,4	75,4	78,7	79,4	79,7	80,2
	100	80,2	80,3	**77,3**	79,2	**80,9**	80,0	80,8
	200	80,9	**81,2**	77,2	**79,6**	80,6	80,6	81,1
	300	**81,8**	80,4	76,9	78,2	80,3	80,8	**81,3**
	400	80,9	80,5	76,7	79,2	80,5	80,5	80,7
	500	81,3	80,7	77,3	77,5	80,0	**80,9**	81,0
	600	80,9	80,6	76,0	77,7	80,3	80,9	80,3
	700	80,6	79,7	75,6	76,4	79,3	80,1	79,6
Rtr-5251	**4**	30,4	32,0	22,5	64,1	65,8	62,5	66,7
	8	56,8	61,3	51,3	66,4	75,3	74,8	74,7
	16	75,2	76,9	66,9	77,8	80,8	80,9	79,0
	32	80,0	81,1	77,5	82,3	83,8	83,3	84,2
	64	84,2	84,4	80,4	83,9	85,6	84,8	84,6
	100	84,3	**85,5**	82,0	84,6	**86,0**	85,6	85,3
	200	**85,7**	85,2	82,8	84,9	85,6	**85,8**	**85,8**
	300	85,6	84,9	**83,5**	84,7	86,0	85,7	85,7
	400	84,9	84,9	83,2	**85,0**	85,5	85,7	85,6
	500	84,2	84,5	82,4	84,4	85,5	85,2	85,5
	600	84,4	84,7	81,8	84,9	85,2	85,1	85,4
	700	84,3	84,5	81,6	83,7	85,2	85,4	85,4
Xnh-4500	**4**	19,1	19,9	16,3	38,6	41,6	40,4	44,8
	8	36,9	41,0	33,1	60,7	63,5	60,6	59,7
	16	56,9	62,0	52,9	67,0	68,8	67,1	68,9
	32	67,5	70,1	68,9	74,6	75,3	75,0	74,2
	64	75,8	77,7	74,3	78,4	80,3	81,0	79,9
	100	78,8	77,5	75,6	79,8	81,9	81,6	80,6
	200	81,3	**82,1**	78,7	80,8	82,8	**83,1**	82,2
	300	**82,0**	81,5	79,5	81,1	**82,9**	82,5	**83,2**
	400	81,0	81,2	**79,8**	**81,3**	82,5	82,3	82,4
	500	80,5	81,0	78,9	81,3	82,5	81,8	82,3
	600	81,3	80,9	78,2	80,9	81,8	81,7	81,9
	700	81,2	81,4	78,5	79,3	81,8	82,4	82,4

Tableau 6-18. Performances de classification dans l'espace des thèmes.

Corpus	Stemming	LDA64	LDA100	LDA200	TF	TF.IDF
Ech-11k	**Khoja1**	84,31	84,40	84,83	82,52	82,55
	BBw2	85,39	85,70	**85,89**	85,33	85,63
Rtr-41k	**Khoja1**	88,45	89,59	90,72	85,85	85,68
	BBw2	90,32	91,15	**91,16**	87,43	87,52
Xnh-36k	**Khoja1**	79,49	80,77	81,98	75,86	75,51
	BBw2	80,56	82,50	**82,80**	78,41	78,31

Tableau 6-19. Evaluation de la catégorisation dans différents espaces de thèmes et de termes.

Les résultats présentés dans Tableau 6-19 montrent une nette amélioration de la catégorisation dans l'espace des thèmes par rapport à celle utilisant directement la pondération des termes (*TF* et *TF.IDF*).

6 Conclusion

Le présent chapitre était consacré aux aspects pratiques de l'analyse automatique du texte arabe. Un effort considérable a été fourni pour concevoir une plate-forme d'expérimentation aux différentes applications de RI. Trois collections ont été conçues à cette fin en développant un explorateur du Web (crawler) pour extraire automatiquement les articles de presse de trois éditeurs arabes (*Echorouk*, *Reuters* et *Xinhua*). Sur quelques millions de mots, chacun des trois corpus utilise un vocabulaire avoisinant les 180.000 mots distincts et dont la distribution de fréquence respecte la loi de Zipf.

Par ailleurs, l'application de notre analyseur morphologique par *BBw* nous a révélée que l'ambigüité lexicale dans le texte arabe affecte 10% environ du langage de la presse arabe. Une mesure du degré de confusion a été proposée pour apprécier l'effet de retenir tous les lemmes possibles pour l'espace d'indexation. Les expérimentations ont montré que notre approche a pu préserver la richesse sémantique sans pour autant compromettre la caractérisation lexicale des textes arabes. En plus, l'évaluation de *Paice* prouve l'efficacité de l'analyse par lemme en comparant à la racinisation (par *Khoja*) ou au stemming léger (par *ISRI*).

L'apprentissage du modèle LDA dans les trois corpus nous a permis d'analyser la sémantique des articles de la presse arabe. Tout d'abord, le calcul de la matrice de confusion thème×catégorie constitue un outil de synthèse astucieux permettant de détecter les sujets pertinents dans chaque catégorie publiée par l'éditeur. En plus, nous avons proposé une approche pour analyser les différents contextes d'un terme à partir des distributions des thèmes inférés. En fixant un seuil minimal de probabilité, il est possible de tracer un graphe sémantique autour de chaque terme et effectuer, par conséquent, une désambigüisation contextuelle automatique. Par ailleurs, les expérimentations, de la catégorisation dans des collections réelles du texte arabe, ont prouvé l'efficacité de l'indexation dans l'espace réduit des thèmes. La sélection du nombre de thèmes adéquat pour l'apprentissage LDA est réalisée par la nouvelle mesure *BKL* dont les différentes expérimentations ont montré une large concordance avec les performances obtenues par la classification SVM.

Chapitre 7 :
CONCLUSION GENERALE

1 Contexte

La satisfaction du besoin d'information dans un environnement aussi étendu et diversifié que le Web constitue une tâche vitale pour tout système d'information accessible sur le réseau mondiale. La recherche intelligente englobe l'ensemble des approches et des techniques développées dans le cadre de la recherche d'information moderne pour faciliter davantage l'accès à l'information. Cette "intelligence" peut être concrétisée par le développement d'interfaces interactives et intuitives mettant en œuvre des techniques de visualisation avancées. Une autre approche "intelligente" consiste à adapter la recherche au profil particulier de l'utilisateur. Le contexte d'une recherche peut être caractérisé par différents aspects (personnel, social, professionnel, spatial, temporel, …etc.).

Par ailleurs, la prise en compte de la sémantique du contenu représente une autre approche en s'intéressant à l'analyse et l'indexation des objets eux-mêmes de la recherche. Une première solution pour appliquer la recherche sémantique sur le Web consiste à l'organiser autour de connaissances conceptuelles (thésaurus ou ontologie). Cette solution nécessite la disponibilité préalable de ces ressources dont la création et la maintenance s'avèrent trop coûteuses. Une autre solution propose de concevoir un système d'annotation qui sera enrichi par des experts ou par une masse d'utilisateurs dans un processus collaboratif. Néanmoins, le succès d'une telle approche dans sa dimension sociale nécessite un cumul considérable, aussi équilibré que diversifié d'annotations du contenu en-ligne.

La nature non-structurée de la majorité des documents sur la toile mondiale relativise l'apport des méthodes d'annotation manuelle ou celles basées sur les ontologies. L'indexation sémantique des textes non-structurés représente une approche efficace pour développer une recherche intelligente dans un fond documentaire aussi évolutif et diversifié que le Web. En particulier, les modèles de thème constituent une alternative judicieuse pour indexer de

manière complètement automatique le texte des pages Web ou des bibliothèques électroniques. L'indexation sémantique par l'allocation latente de Dirichlet (LDA) se base sur une approche probabiliste pour calculer les distributions des thèmes latents à travers les termes composant les textes d'une collection.

2 Synthèse

Les travaux présentés dans ce manuscrit s'inscrivent dans ce contexte en s'intéressant à la faisabilité d'appliquer un modèle de thème dans l'indexation sémantique des textes non-structurés pour certaines tâches de RI. Du moment où la modélisation par LDA attribue pour chaque document une distribution de thèmes (descripteurs latents), nous avons pu représenter différents corpus réels de textes dans un espace de thèmes préalablement appris par échantillonnage de Gibbs. Les modèles générés ont été utilisés efficacement dans la classification par SVM ou même dans une recherche ad-hoc. En plus de la réduction considérable de dimensionnalité, les performances de classification ou de recherche étaient équivalentes et même parfois améliorées par rapport à celles appliquées dans l'espace des termes. En particulier, nous avons proposé et validé un cadre théorique pour l'extension thématique des requêtes de recherche. Cette approche permet d'améliorer la performance d'une recherche ad-hoc par enrichissement de la requête initiale ou, tout simplement, par suggestion d'affinement de la requête dans une saisie interactive. Les évaluations préliminaires, obtenues par analyse comparative de plusieurs modèles de recherche, sont prometteuses et ouvrent de nouvelles perspectives pour la prise en charge du besoin d'information dans la recherche ad-hoc.

Cependant, l'apprentissage du modèle LDA requiert l'introduction arbitraire du nombre de thèmes. Généralement, ce nombre est choisi en littérature entre 30 et 300 selon le volume et la richesse de la collection en question. Même si nous trouvons des propositions théoriques, pour déterminer un nombre convenable de thèmes pour l'inférence LDA, peu sont les travaux expérimentaux qui l'ont mises en œuvre. De notre part, nous avons développé une mesure combinée basée sur la divergence de *Kullback-Leibler* pour sélectionner ce nombre de thèmes. Notre idée partait du fait que les thèmes inférés doivent être, autant que possible, stables (sur plusieurs inférences) et distincts (dans le même modèle) en même temps. La modélisation LDA d'une panoplie de collections selon un large intervalle de nombre de thèmes nous a permis d'apprécier l'efficacité de cette nouvelle mesure *BKL*. La concordance quasi-totale avec les performances de classification valide notre hypothèse de départ : Un modèle LDA optimal doit établir le meilleur compromis entre la stabilité des thèmes et leur capacité discriminative.

Par ailleurs, la modélisation par thème offre plusieurs issues pour l'analyse sémantique d'un texte. En fixant un seuil de probabilité minimal dans la distribution des thèmes, nous avons proposé de définir, pour chaque terme, les contextes qui lui sont rattachés. Un graphe de sémantique contextuelle peut être tracé autour de chaque terme du vocabulaire de la collection. Quelques exemples de mots arabes, qui incarnent plusieurs sens, ont été analysés. De plus, dans une collection de documents étiquetés, la modélisation LDA nous permet de détecter les thèmes pertinents dans chaque catégorie. Par calcul de la matrice de confusion terme×thème, nous avons proposé une lecture simplifiée des principaux sujets discutés dans la presse électronique de plusieurs éditeurs durant la période 2007-2009.

Afin d'approfondir notre étude sur les modèles de thèmes, nous nous sommes intéressés à l'analyse de l'impact des aspects linguistiques dans l'indexation sémantique des textes. Vu qu'une majorité des travaux en RI a été réalisée sur le texte anglais, l'hypothèse dominante affirmait qu'il suffisait pour un SRI d'appliquer un prétraitement par stemming léger. La

nature morphologique relativement simple de l'anglais soutenait cette hypothèse et, par conséquent, elle ne justifiait pas le supplément de coût des analyseurs linguistiques. Cependant, la vérification de cette hypothèse sur le texte non-anglais n'est pas assez claire pour deux raisons : La première réside dans la richesse morphologique (dérivationnelle et flexionnelle) dans d'autres langues. Dans ce cas, il n'est pas évident qu'un processus de dés-affixation par stemming léger puisse unifier la majorité des formes du langage. La deuxième cause est liée à l'objectif lui-même de l'indexation sémantique où un prétraitement linguistique impropre peut embrouiller les aspects sémantiques du texte original.

En particulier, l'arabe, notre langue nationale qui inscrit la plus haute croissance dans son utilisation sur le Web, enregistre un déficit énorme dans les études relatives aux applications de recherche d'information moderne. A cet effet, nous avons jugé utile de développer des outils linguistiques pour l'analyse lexicale du texte arabe. Ceci nous a permis d'évaluer la modélisation par thème et l'indexation sémantique dans une langue s'articulant autour d'une morphologie aussi riche que l'arabe.

Nous avons introduit une nouvelle méthode de stemming à base de lemme non vocalisé et indexant toute les solutions possibles de lemmatisation. Basé sur l'analyseur morphologique de *Buckwalter*, notre algorithme *BBw* a été développé et expérimenté sur trois corpus réels de presse arabe (*Echorouk*, *Reuters* et *Xinhua*). Une mesure du degré de confusion a été proposée pour apprécier l'utilité de retenir tous les lemmes possibles pour l'espace d'indexation. Les expérimentations ont montré que notre approche a pu préserver la richesse sémantique sans pour autant compromettre la caractérisation lexicale du texte arabe.

En plus de l'algorithme *BBw*, nous avons conçu un ensemble de groupes-concepts pour l'évaluation des méthodes de stemming arabe selon les mesures de *Paice*. Par ailleurs, nous avons réalisé et testé trois collections linguistiques arabes, à base d'articles de presse relatifs à la période 2007-2010. Toutes ces ressources seront mises à la disposition des chercheurs dans le domaine afin de contribuer dans les efforts de promotion de l'analyse automatique du texte arabe.

Il est intéressant de noter qu'il est difficile de comprendre comment peut-on apprécier les aspects sémantiques dans le texte arabe sans connaissance linguistique suffisante [Larkey et al., 2002] [Larkey et al., 2004]. Notre étude montre qu'un développement significatif de la recherche d'information arabe ne puisse être réalisé sans collaboration étroite entre informaticiens et linguistiques arabes.

Nos travaux, relatifs à l'analyse sémantique du texte arabe, ont été présentés et débattus avec deux communautés de chercheurs. La première représentait les linguistes intéressés par le traitement automatique du langage naturel. Une communication a été présentée dans le 4ème colloque international en traductologie et traitement automatique de la langue [Brahmi et Benyettou, 2010]. La deuxième concernait les spécialistes arabes en informatiques. Un article rédigé en arabe a été accepté dans la 7ème conférence de l'informatique en arabe [Brahmi et al., 2011]. Dans les deux papiers, nous avons expliqué notre approche d'analyse, du texte arabe, par lemme non vocalisé en vue d'une catégorisation ou une modélisation par thèmes latents. Par ailleurs, notre approche de catégorisation et de modélisation par thème pour l'analyse sémantique du texte arabe a été appréciée par une communauté spécialisée en recherche d'information moderne [Brahmi et al., 2012].

3 Perspectives

Nos travaux définissent une base, tant théorique que pratique, de la modélisation par thème pour l'indexation sémantique des textes non-structurés. Les perspectives qu'on peut tracées de cette étude sont :

- Intégrer la modélisation par thème dans une application finale de recherche intelligente ad-hoc sur le Web ou dans des bibliothèques électroniques. L'approche d'extension thématique de la requête peut être intégrée dans deux niveaux : dans le calcul de la pertinence et dans l'interface interactive d'acquisition.
- Reformuler le modèle génératif LDA afin de prendre en charge la multiplicité des solutions de lemmatisation pour un mot observé dans un document. L'application sur des corpus réels, dont la langue pose une ambigüité lexicale, permettra d'apprécier l'efficacité d'une telle approche dans différentes tâches de RI.
- Construire des modèles thèmes typiques sur des collections de textes approuvées et conséquentes. Les thèmes latents générés serviront d'index sémantique de références pour les pages Web dans une navigation intelligente. L'utilisation de la source des articles de l'encyclopédie libre *Wikipedia* représente une solution intéressante dans cette démarche.
- Développer un analyseur morphologique du texte arabe dont le lexique doit être enrichi périodiquement en collaboration avec des linguistes arabes. La promotion des études relatives à l'analyse automatique du texte arabe n'est possible que par la réalisation de ressources libres mises à la disposition des chercheurs dans le domaine. En plus des trois corpus et des variantes algorithmiques développées dans cette étude, il est nécessaire d'élaborer un catalogue en-ligne regroupant les différentes ressources linguistiques arabes (algorithmes, corpus, documentation).

Bibliographie

A

1. Abdelali A., Cowie J., & Soliman H. S. (2005). Building a modern standard Arabic corpus. In workshop on computational modeling of lexical acquisition. Croatia, pp. 1-7.

2. Al-Shammari, E. & Lin, J. (2008). A novel Arabic lemmatization algorithm. In Proc. Workshop on Analytics for Noisy Unstructured Data, Singapore, pp. 113-118.

3. Al-Shammari, E. (2010). Lemmatizing, stemming, and query expansion method and system. US Patent 20100082333, Apr 2010.

4. Acid, S., de Campos, L., Fernandez, J., & Huete, J. (2003). An information retrieval model based on simple bayesian networks. International Journal of Intelligent Systems. Vol. 18, No. 2, pp. 251–265.

5. Allan, J., Aslam, J., Belkin., N., Buckley, C., Callan, J., Croft, B., Dumais, S., Fuhr, N., Harman, D., Harper, D.J., Hiemstra, D., Hofmann, T., Hovy, E., Kraaij. W., Lafferty, J., Lavrenko, V., Lewis, D., Liddy, L., Manmatha. R., McCallum, A., Ponte, J., Prager, J., Radev, D., Resnik, P., Robertson, S., Rosenfeld, R., Roukos, S., Sanderson, M., Schwartz, R., Singhal, A., Smeaton, A., Turtle, H., Voorhees, E., Weischedel, R., Xu, J., & Zhai, C.X. (2003). Challenges in information retrieval and language modeling: report of a workshop held at the center for intelligent information retrieval, University of Massachusetts Amherst, September 2002, ACM SIGIR Forum, Vol. 37, No. 1, pp. 31-47.

6. Azzopardi, L., Girolami, M. & Rijsbergen, K.V. (2003). Investigating the relationship between language model perplexity and IR precision-recall measures. In Proc. SIGIR'03.

B

7. Baeza-Yates, R.A, & Ribeiro-Neto, B.A.(1999). Modern Information Retrieval. Addison-Wesley / ACM Press NY, USA.

8. Bennett, G., Scholer, F., & Uitdenbogerd, A. L. (2008). A Comparative Study of Probabalistic and Language Models for Information Retrieval. ADC'08: 65-74.

9. Blei, D.M., Ng, A.Y., & Jordan, M.I. (2003). Latent Dirichlet allocation. Journal of Machine Learning Research, 3, pp. 993–1022.

10. Berners-Lee, T. (1998). Semantic Web roadmap. http://www.w3.org/DesignIssues/ Semantic.html

11. Biber, D. (1993). Representativeness in corpus design. Literary and Linguistic Computing. Vol. 8, No. 4. pp. 243-257.

12. Bonny, P., & Garnier, A. (2008). Le marché mondial du Text Mining. Veille Magazine, Février 2008.

13. Bordogna, G., Carrara, P., & Pasi, G. (1991). Query term weights as constraints in fuzzy information retrieval. Information Processing and Management, Vol. 27, No. 1, pp. 15-26.

14. Bordogna, G., & Pasi G. (2001). Modeling Vagueness in Information Retrieval. Lecture Notes in Computer Science, Vol. 1980, pp. 207-241.

15. Boughanem, M., Chrisment, C., & Soulé-Dupuy, C. (1999). Query Modification Based on Relevance Back-Propagation in an Ad hoc Environment. Inf. Process. Manage. Vol. 35, No. 2, pp. 121-139.

16. Boughanem, M., Loiseau, Y., & Prade, H. (2005). Rank-Ordering Documents According to Their Relevance in Information Retrieval Using Refinements of Ordered-Weighted Aggregations. Adaptive Multimedia Retrieval, pp. 44-54.

17. Brants, T., Chen, F., & Farahat, A. (2002). Arabic Document Topic Analysis. LREC-2002 Workshop on Arabic Language Resources and Evaluation, (Las Palmas, Spain).

18. Brahmi, A. (2005). Les Méthodes d'Apprentissage à base de Kernels pour la Recherche d'Information. Mémoire de magister, USTO-MB-Oran,.

19. Brahmi, A., & Ech-Cherif, A. (2005). Regularized Classifiers for Information Retrieval. Lecture Notes in Computer Science: Advances in Artificial Intelligence, Vol. 3501, pp. 427-431.

20. Brahmi, A. & Benyettou, A. (2010). Analyse de la Sémantique Latente dans les Textes Arabes pour la Recherche d'Information Intelligente. 4ème Colloque International en Traductologie et TAL, Oran, 7-9 novembre 2010.

21. Brahmi, A., Ech-Cherif, A., & Benyettou., A. (2011). Latent semantic analysis in Arabic texts with morphological analysis and topic modeling. Seventh International Computing Confrence in Arabic (ICCA'2011), Riadh-Saoudi Arabia, 31 may-2 june 2011.
عبدالرزاق براهمي، أحمد الشريف، عبدالقادر بن يطو. تحليل الدلالات الكامنة في النصوص العربية بواسطة التحليل الصرفي والنمذجة بالمواضيع. الدورة السابعة للمؤتمر الدولي لعلوم وهندسة الحاسوب. 31 مايو- 2 يونيو 2011. الرياض، المملكة العربية السعودية.

22. Brahmi, A., Ech-Cherif, A., & Benyettou., A. (2012). Arabic texts analysis for topic modeling evaluation. Information Retrieval, Vol. 15, No. 1, pp. 33-53. DOI : http://dx.doi.org/10.1007/s10791-011-9171-y.

23. Brin, S., & Page, L. (1998). The anatomy of a large-scale hypertextual Web search engine. Computer Networks and ISDN Systems, Vol. 30(1-7), pp. 107-117.

24. Broder, A. (2002). A taxonomy of web search. SIGIR Forum, Vol. 36(2), pp. 3-10.

25. Buntine, W.L. (1994). Operations for learning with graphical models. Journal of Artificial Intelligence Research Vol. 2, pp. 159-225.

26. Burges, C. J. C. (1998). A tutorial on support vector machines for pattern recognition. Data Mining and Knowledge Discovery, Vol. 2, pp. 121-167.

27. Bush, V. (1945). As we may think. In The Atlantic Monthly, July 1945, Trad. Ch. Monnatte.

28. Büttcher, S., Clarke, C.L.A., & Cormack, G.V. (2010). Information Retrieval: Implementing and Evaluating Search Engines. MIT Press, Cambridge, Mass..

C

29. Canfora, G., & Cerulo, L. (2004). A taxonomy of information retrieval models and tools. CIT. Journal of computing and information technology, Vol. 12, no3, pp. 175-194.

30. Chemudugunta, C., Smyth, P., & Steyvers, M. (2008). Combining concept hierarchies and statistical topic models. In Proceeding of the 17th ACM conference on Information and knowledge management, pp. 1469-1470, , NY, USA.

31. Cho, J., Garcia-Molina, H., & Page, L. (1998). Efficient crawling through URL ordering. Computer Networks, Vol. 30(1-7), pp. 161-172.

32. Ciravegna, F., & Chapman, S. (2005). Mining the Semantic Web: Requirements for Machine Learning. Machine Learning for the Semantic Web Dagstuhl Seminar 05071, Dagstuhl, DE.

33. Cleverdon., C. W. (1967). The Cranfield tests on index language devices. Aslib Proceedings, Vol. 19, pp. 173–192.

34. Cohen, P. R., & Kjeldsen, R. (1987). Information Retrieval by constrained spreading activation in semantic networks, In Information Processing and Management, Vol. 23(4), pp. 255-268.

35. Crestani, F., & Pasi., G. (1999). Soft Information Retrieval: Applications of Fuzzy Set Theory and Neural Networks. Neuro-Fuzzy Techniques for Intelligent Information Systems In Neuro-Fuzzy Techniques for Intelligent Information Systems, pp. 287-315.

36. Crestani, F., De-Campos, L.M., Fernandez-Luna, J.M., & Huete, J.F. (2003). A multi-layered bayesian network model for structured document retrieval. In Proc. of the 7th European Conference on Symbolic and Quantitative Approaches to Reasoning with Uncertainty (ECSQARU), pp. 74–86.

37. Croft W. B., (1981). Document representation in probabilistic models of information retrieval. In Journal of the American Society for Information Science, pp. 451-457.

D

38. Deerwester, S.C., Dumais, S.T., Landauer, T.K., Furnas, G.W., & Harshman, R.A. (1990). Indexing by latent semantic analysis. Journal of the American Society of Information Science, Vol. 41(6), pp. 391–407.

39. Dempster, A. P., Laird, N. M. & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. Journal of the Royal Statistical Society. Series B (Methodological), Vol. 39(1), pp. 1–38.

40. Dellschaft, K., & Staab, S. (2006). On How to Perform a Gold Standard Based Evaluation of Ontology Learning. International Semantic Web Conference'2006, pp. 228-241.

41. Dewey, M. (1920). DC Beginnings. Library Journal, 45, 152. In Comaromi, Op. Cit., pp. 6.

42. Dewey, M., Couture-Lafleur, R., & Mitchell, J.S. (2005). Classification décimale Dewey abrégée et index". Éditions ASTED, Montréal.

43. Dolamic, L., & Savoy, J. (2010). When stopword lists make the difference. JASIST 61(1), pp. 200-203.

44. Dumais, S. T., (1993). Latent semantic indexing (LSI) and TREC-2. In Proc. TREC, pp. 105–115.

E

45. Efthimiadis E. (2000). Interactive query expansion : a user based evaluation in relevance feedback environment. Journal of the American Society for Information Science, Vol. 51, no 11, pp. 989-1003.

46. Evans D., (2007). Corpus building and investigation for the Humanities. University of Nottingham http://www.corpus.bham.ac.uk/corpus-building.shtml

F

47. Frakes, W. B. (2003). Strength and similarity of affix removal stemming algorithms. In SIGIR Forum, Vol. 37, No. 1., pp. 26-30.

48. Friedman, N., & Moises, G. (1996). Building classifiers using Bayesian networks. In Proc. National Conference on Artificial Intelligence, pp. 1277–1284.

G

49. Gilks, W. R., Richardson, S. & Spiegelhalter, D. J. (1996). Markov Chain Monte Carlo in Practice. Chapman & Hall, New York.

50. Go-Globe, (2011). What happens in the Internet in 60 seconds. http://www.go-globe.com/. Published in sept. 2011.

51. Goweder, A. & De-Roeck, A. (2001). Assessment of a significant Arabic corpus. Presented at the Arabic NLP Workshop at ACL/EACL 2001, Toulouse, France, 2001.

52. Griffiths, T. L., & Steyvers, M. (2004). Finding scientific topics. Proc Natl. Acad. Sci. USA, 101 Suppl 1, pp. 5228–5235.

53. Griffiths, T. L., Steyvers, M., Blei, D. M., & Tenenbaum, J. B. (2005) Integrating topics and syntax. In Advances in Neural Information Processing 17. Cambridge, MA: MIT Press.

54. Grolier, E. (1970). Quelques travaux récents en matière de classification encyclopédique. BBF, no. 3, pp. 99-126.

H

55. Handschuh S., & Staab, S. (2002). Authoring and Annotation of Web Pages in CREAM. WWW'2002, pp. 462-473.

56. Heflin, J., & Hendler J. (2000). Searching the Web with SHOE. AAAI-2000 Workshop.

57. Heinrich, G. (2008). Parameter estimation for text analysis. Technical report, University of Leipzig.

58. Hofmann, T. (1999). Probabilistic latent semantic indexing. Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval, pp. 50-57.

59. Howard, R.A., & Matheson, J. E. (1981). Influence diagrams. In Howard, R. A., and Matheson, J. (Eds.), The Principles and Applications of Decision Analysis, pp. 720–762.

60. Hu, J., Fang, L., Cao, Y., Zeng, H.-J., Li, H., Yang, Q., & Chen, Z. (2008). Enhancing text clustering by leveraging Wikipedia semantics. In SIGIR '08: Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval, NY, USA, pp. 179-186.

61. Hull, D. (1998). Stemming algorithms - A case study for detailed evaluation. Journal of the American Society for Information Science, Vol 47, No. 1, pp. 70-84.

62. Hwang, M., Kong, H., Baek, S., Hwang, K., & Kim, P. (2007). The techniques for the ontology-based information retrieval. In The 9th International Conference on Advanced Communication Technology (IEEE Cat.No.07EX1671), Piscataway, NJ, USA, Vol. 2, pp. 1365-1369.

I

63. Indrawan, M., Srinivasan, B., Wilson, C., & Redpath, R. (1998). Optimising bayesian belief networks : the case study of information retrieval systems. In Proc. of the IEEE Conference on System, Man and Cybernetics, pp. 2273–2278.

64. Ingwersen, P. (1992). Information Retrieval Interaction. Taylor Graham, London.

65. Iwata T., Yamada, T., & Ueda, N., (2008). Probabilistic latent semantic visualization: topic model for visualizing documents. In KDD ’08: Proceeding of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 363–371, New York, USA.

J

66. Jagaralamudi, J., & Daumé, H. (2010). Extracting multilingual topics from unaligned corpora. In Proceedings of the European Conference on Information Retrieval (ECIR), Milton Keynes, UK, 2010.

67. Jensen, F.V. (2001). Bayesian Networks and Decision Graphs. IEEE Computer Society Press, NY.

68. Jones, K.S, & Willett, P. (1997). Readings in Information Retrieval. Morgan Kaufmann Publishers, September 1997, San Francisco, (ISBN 1-55860-454-5), pages 587.

69. Jongejan, B., & Dalianis, H. (2009). Automatic training of lemmatization rules that handle morphological changes in pre-, in- and suffixes alike. ACL/AFNLP’2009, pp. 145-153.

70. Jurafsky, D, & Martin, J. (2000). Speech and Language Processing. Prentice Hall, Upper Saddle River NJ.

K

71. Kadri, Y., & Nie, J. (2006). Effective Stemming for Arabic Information Retrieval. The Challenge of Arabic for NLP/MT, International Conf. at the British Computer Society (BCS), pp. 68-74, London, UK.

72. Kipp, M.E.I., & Campbell, D.G. (2010). Searching with tags: Do tags help users find things? Knowledge Organization, Vol. 37(4), pp. 239-255.

73. Kraaij, W., & Pohlmann, R. (1996). Viewing stemming as recall enhancement. In: Proceedings, 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '96) Zurich, pp. 40-48.

74. Kraft, D. H., Bordogna, G. and Pasi, G. An extended fuzzy linguistic approach to generalize Boolean information retrieval, Journal of Information Sciences - Applications, 2(3), 1995.

75. Krovetz, R. Viewing morphology as an inference process. In: Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (1993) 191-202

76. Kuropka, D. (2004). Modelle zur Repräsentation natürlichsprachlicher Dokumente. Ontologie-basiertes Information-Filtering und -Retrieval mit relationalen Datenbanken. “Models for the representation of natural language documents. Ontology-based information filtering and retrieval

with relational databases" Advances in Information Systems and Management Science, Bd. 10. ISBN 978-3-8325-0514-1, 264 pages.

77. Kwok, K.L. (1995). A network approach to probabilistic information retrieval. ACM transactions on information systems, pp. 324-353.

L

78. Lafferty, J.D., & Zhai, C.X. (2001). Document Language Models, Query Models, and Risk Minimization for Information Retrieval. SIGIR'2001, pp. 111-119.

79. Lancaster, F.W. (1968). Information Retrieval Systems: Characteristics, Testing and Evaluation, Wiley, New York.

80. Larkey, L. S. & Connell, M. E. (2001). Arabic information retrieval at UMass in TREC-10. In TREC 2001, pp. 562-570. Gaithersburg, Maryland, USA.

81. Larkey, L. S., Ballesteros, L., & Connell, M. E. (2002). Improving stemming for Arabic information retrieval: Light stemming and co-occurrence analysis. In Proceedings of SIGIR'2002, pp. 275-282, Tampere, Finland.

82. Larkey, L.S., Feng F., Connell M. E., & Lavrenko V. (2004). Language-specific models in multilingual topic tracking. In Proceedings of SIGIR 2004, pp. 402-409, Sheffield, UK.

83. Lavrenko, V. & Croft, W.B. (2001). Relevance-based language models. In Proceedings of the ACM SIGIR International Conference on Research and Development in Information Retrieval, pp. 120–127, New Orleans, LA.

84. Lawrence, L, & Giles, C.L. (2000). Accessibility of information on the Web. Intelligence, Vol. 11(1): pp. 32-39.

85. Lennon, M., Pierce, D.C., & Willett, P (1981). An evaluation of some conflation algorithms. Journal of Information Science, 3 pp. 177-183.

86. Lloret, E., & Palomar, M. (2010). Challenging Issues of Automatic Summarization: Relevance Detection and Quality-based Evaluation. Informatica (Slovenia) Vol. 34(1): pp. 29-35.

87. Lovins, J.B. (1968). Development of a stemming algorithm. Translation and Computational Linguistics, Vol. 11: pp. 22-31.

88. Luhn, H. P. (1957). A statistical approach to mechanized encoding and searching of literary information. IBM Journal of Research and Development.

M

89. Manning, C.D., Raghavan, P., & Schütze., P. (2008). Introduction to information retrieval. Cambridge University Press: I-XXI, pp. 1-482.

90. Manning, C.D., & Schütze., P. (2001). Foundations of statistical natural language processing. MIT Press 2001, pp. 1-680.

91. Maron, M.E. (1965). Mechanized documentation: The logic behind a probabilistic interpretation. In: Statistical Association Methods for Mechanized Documentation (Edited by Stevens et al.) National Bureau of Standards, Washington, pp. 9-13.

92. Menczer, F., Pant, G., & Srinivasan, P. (2004). Topical web crawlers: Evaluating adaptive algorithms. ACM Trans. Internet Techn. 4(4): 378-419.

93. Miller, G.A., Beckwith, R., Fellbaum, C. D., Gross, D., & Miller., K. (1990). WordNet: An online lexical database. International Journal of Lexicography, Vol. 3, pp. 235-244.

94. Moukdad, H. (2006). Stemming and root-based approaches to the retrieval of Arabic documents on the Web. Webology, 3(1), article 22.

N

95. Najork, N., & Wiener, J.L. (2001). Breadth-first crawling yields high-quality pages. WWW 2001: pp. 114-118

96. Nallapati, R., & Cohen, W. (2008). Link-pLSA-LDA: A new unsupervised model for topics and influence of blogs. In Proceedings of the International Conference on Weblogs and Social Media (ICWSM).

97. Newman, D., Hagedorn, K., Chemudugunta, C., & Smyth, P. (2007). Subject metadata enrichment using statistical topic models. In JCDL '07: Proceedings of the 7th ACM/IEEE-CS joint conference on Digital libraries, pp. 366-375, New York, NY, USA. ACM Press.

98. Nie, J.Y., (1990). Un modèle logique général pour les systèmes de recherche d'informations : application au prototype RIME. Phd thesis University Joseph-Fourier – Grenoble.

O

99. Oard, D. W., & Gey, F. (2002). The TREC-2002 Arabic/English CLIR Track. In TREC2002 notebook, pp. 81-93.

P

100. Paice, C. D. (1996). Method for evaluation of stemming algorithms based on error counting. Journal of the American Society for Information Science, Vol. 47, No. 8., pp. 632–649.

101. Paliouras, G. (2005). On the Need to Bootstrap Ontology Learning with Extraction Grammar Learning. ICCS 2005: pp. 119-135.

102. Pearl, J. (1988). Probabilistic reasoning in intelligent systems: Networks of plausible Inference. Morgan Kaufman Publishers, Inc., San Mateo, CA, 2nd Edition.

103. Polguère, A. (2003). Lexicologie et sémantique lexicale ; Notions fondamentales. Les Presses de l'Université de Montréal, collection Paramètres, Montréal, 260p.

104. Ponte, J. M. & Croft, W. B. (1998). A language modeling approach to information retrieval. In Proceedings of the ACM SIGIR International Conference on Research and Development in Information Retrieval, pp. 275–281, Melbourne, Australia.

105. Popescul, A., Ungar, L., Pennock, D. & Lawrence, S. (2001). Probabilistic models for unified collaborative and content-based recommendation in sparse-data environments. In Uncertainty in Artificial Intelligence, Proceedings of the Seventeenth Conference.

106. Popov B., Kiryakov A., Kirilov A., Manov D., Ognyanoff D., & Goranov M. (2003). KIM – Semantic Annotation Platform, 2nd International Semantic Web Conference (ISWC2003), LNAI Vol. 2870, pp. 834-849, Springer-Verlag.

107. Porteous, I., Newman, D., Ihler, A., Asuncion, A., Smyth, P., & Welling, M. (2008). Fast collapsed Gibbs sampling for latent Dirichlet allocation. In KDD '08: Proceeding of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining, pp. 569-577, New York, NY, USA.

108. Porter, M. F. (1980). An algorithm for suffix stripping. Program 14:130-137.

R

109. Ratinov, L. & Roth, D. (2009). Design challenges and misconceptions in named entity recognition. In CoNLL '09: Proceedings of the Thirteenth Conference on Computational Natural Language Learning, pp. 147-155.

110. Ribeiro-Neto, B. & Muntz., R.R. (1996). A belief network model for IR. In Proc. of the International ACM-SIGIR Conference, pp. 253–260.

111. Rijsbergen, K.V. (1979). Information Retrieval. Butterworths, London.

112. Rijsbergen, K.V. (1986). A non-classical logic for Information Retrieval. The Computer Journal, Vol. 29, No. 6, pp. 481-485.

113. Riloff, E. (1995). Little Words Can Make a Big Difference for Text Classification. In Proceedings of {SIGIR}-95, 18th {ACM} International Conference on Research and Development in Information Retrieval : 130-136.

114. Robertson, S.E. & Sparck-Jones, K., (1976). Relevance weighting of search terms, Journal of the American Society for Information Science, 27, pp. 129-146.

115. Robertson, S.E. (1977). The probabilistic ranking principle in IR. Journal of Documentation, 33:294–304.

116. Robertson, S. E., & Walker, S. (1994). Some simple effective approximations to the 2-poisson model for probabilistic weighted retrieval. In Proceedings of the ACM SIGIR International Conference on Research and Development in Information Retrieval, pages 232–241, Dublin, Irelan.

117. Robertson, S. E., & Walker, S. (1999). Okapi/Keenbow at TREC-8. In Proceedings of the Fourteenth Text REtrieval Conference (TREC-8). NIST Special Publication.

118. Rocchio J. (1971) "Relevance feedback in information retrieval", The SMART retrieval system-experiments in automatic document processing, Prentice Hall Inc, pp. 313-323.

S

119. Said, D., Wanas, N., Darwish, N., & Hegazy, N. (2009). A Study of Text Preprocessing Tools for Arabic Text Classification. In Proceedings of the 2nd International Conference on Arabic Language Resources and Tools, pp. 230-236, Cairo, Egypt.

120. Salton, G. (1968): Automatic Information Organization and Retrieval. New York: McGraw-Hill.

121. Salton, G. (1971). The SMART Retrieval System—Experiments in Automatic Document Retrieval. Prentice Hall Inc., Englewood Cliffs, NJ.

122. Salton, G., Wong, A., & Yang, C. S. (1975). A vector space model for automatic indexing. Communications of the ACM, 18(11):613-620.

123. Salton, G., & McGill, M. J. (1983). Introduction to Modern Information Retrieval. McGraw Hill Book Co., New York.

124. Salton G., Fox E. A., & Wu, H. (1983). "Extended Boolean information retrieval", Communications of the ACM, 26(12): pp. 1022-1036.

125. Salton, G. (1989). Automatic Text Processing: The Transformation, Analysis, and Retrieval of Information by Computer. Reading, MA: Addison Wesley.

126. Savoy, J. (2006). Light stemming approaches for the French, Portuguese, German and Hungarian languages. SAC 2006: pp. 1031-1035.

127. Savoy, J. (2010). Stemming Strategies for European Languages. IICS 2010: pp. 545-557.

128. Sebastiani, F. (2005). Text categorization. In Alessandro Zanasi (ed.), *Text Mining and its Applications*, WIT Press, Southampton, UK, pp. 109-129.

129. Shaw, W.M. & Wood, J.B. & Wood, R.E. & Tibbo, H.R. (1991). The Cystic Fibrosis Database: Content and Research Opportunities. LISR 13, pp. 347-366.

130. Singhal, A. (2001). Modern Information Retrieval: A Brief Overview, Bulletin of the IEEE Computer Society Technical Committee on Data Engineering, Vol. 24, pp. 35-43.

131. Shirky, C. (2005). Ontology is overrated: Categories, links, and tags. Shirky.com. http://shirky.com/writings/ontology_overrated.html.

132. Song, F., & Croft, W. B. (1999). A general language model for information retrieval. In CIKM '99: Proceedings of the Eighth International Conference on Information and Knowledge Management, pp. 316–321, New York, NY, USA.

133. Sparrow, B., Liu, J., & Wegner, D.M. (2011). Google Effects on Memory: Cognitive Consequences of Having Information at Our Fingertips. Science, Vol. 333, No. 6043, pp. 776-778.

134. Spousta, M. (2006). Web as a Corpus. WDS'06 Proceedings of Contributed Papers, Part I, pp. 179–184.

135. Steyvers, M., & Griffiths, T. (2007). Probabilistic Topic Models. Handbook of Latent Semantic Analysis. Lawrence Erlbaum Associates.

136. Strzalkowski, T., Lin, F., Wang, J., & Perez-Carballo, J. (1999). Evaluating Natural Language Processing Techniques in Information Retrieval: A TREC Perspective. In: Strzalkowski, T.(ed.) Natural Language Information Retrieval. Kluwer, Dordrecht.

137. Su, X., & Khoshgoftaar, T.M. (2009). A survey of collaborative filtering techniques, Adv. in Artif. Intell., Vol. 2009, pp. 1-19.

T

138. Taghva, K., Elkoury, R., & Coombs, J. (2005). Arabic Stemming without a root dictionary. In Proceedings of the International Conference on Information Technology: Coding and Computing, Vol. 01, pp. 152-157.

139. Tsatsaronis, G., & Panagiotopoulou, V. (2009). A Generalized Vector Space Model for Text Retrieval Based on Semantic Relatedness. EACL (Student Research Workshop), pp. 70-78.

140. Tuerlinckx, L., "La lemmatisation de l'arabe non classique," In JADT 2004, 7e Journées internationales d'Analyse statistique des Données Textuelles, pp. 1069-1078, 2004.

141. Turtle, H.R. & Croft, W.B. (1990). Inference networks for document retrieval. In Proc. of the International ACM-SIGIR Conference, pages 1–24.

V

142. Vapnik V. N. (1995). The nature of statistical learning theory. Springer-Verlag New York, NY.

W

143. Wallach, H.M. (2006). Topic modeling: beyond bag-of-words. ICML 2006: 977-984.

144. Wang, P., & Domeniconi, C. (2008). Building Semantic Kernels for Text Classification using Wikipedia. In KDD'08, Las Vegas, Nevada, USA.

145. Wang, W., Barnaghi, P.M., & Bargiela, A. (2008). Search with Meanings:An Overview of Semantic Search Systems. International journal of Communications of SIWN, Vol. 3, pp. 76-82.

146. Wei, X., & Croft, W.B., (2006). LDA-based document models for ad-hoc retrieval. SIGIR'06, pp. 178-185.

147. Wong, S.K.M., Ziarko, W, & Wong, P.C.N. (1985). Generalized Vector Space Model in Information Retrieval. SIGIR 1985: 18-25.

148. Wong, S.K.M., Ziarko, W., Raghavan, V.V., & Wong., P.C.N. (1987). On modeling of information retrieval concepts in vector spaces. ACM Transactions on Database Systems, 12(2), pp. 299–321.

X

149. Xu, J. Fraser, A., & Weischedel M. R. (2001). TREC 2001 Cross-lingual Retrieval at BBN NIST Text REtrieval Conference TREC10 Proceedings, Gaithersburg, MD, pp. 68-77.

Y

150. Yi, X., & Allan, J. (2009). A Comparative Study of Utilizing Topic Models for Information Retrieval. Advances in Information Retrieval, Vol. 5478, pp. 29-41.

Z

151. Zadeh, L. A. (1965). Fuzzy sets. Information and Control, Vol. 8, no. 3, pp. 338-353.

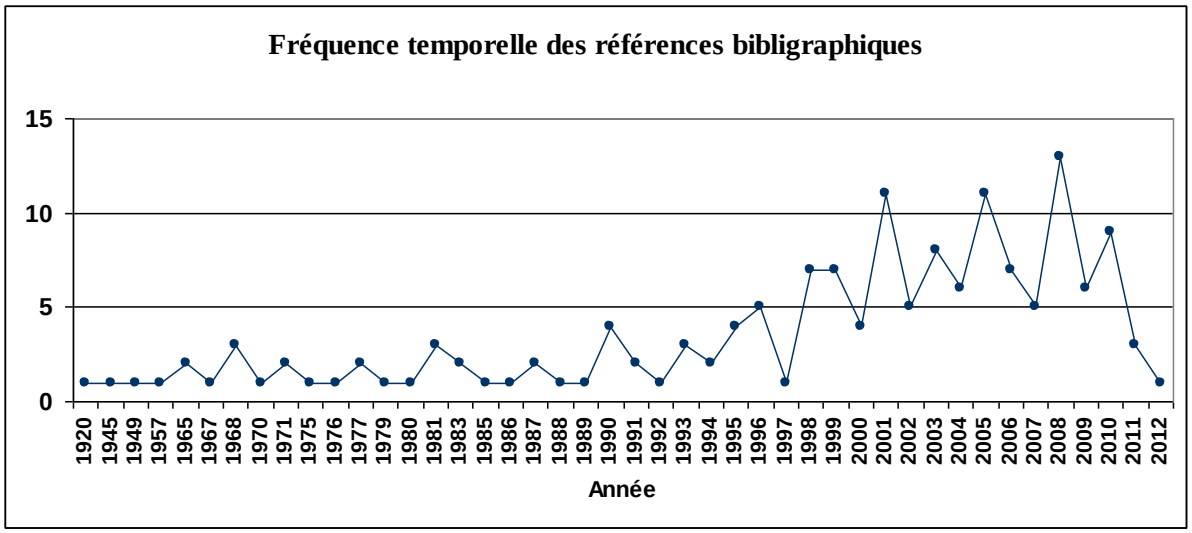
152. Zhang, D., Mei, Q., & Zhai C. (2010). Cross-Lingual Latent Topic Extraction," In Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics (ACL'10), pp. 1128-1137., Uppsala, Sweden, 2010.

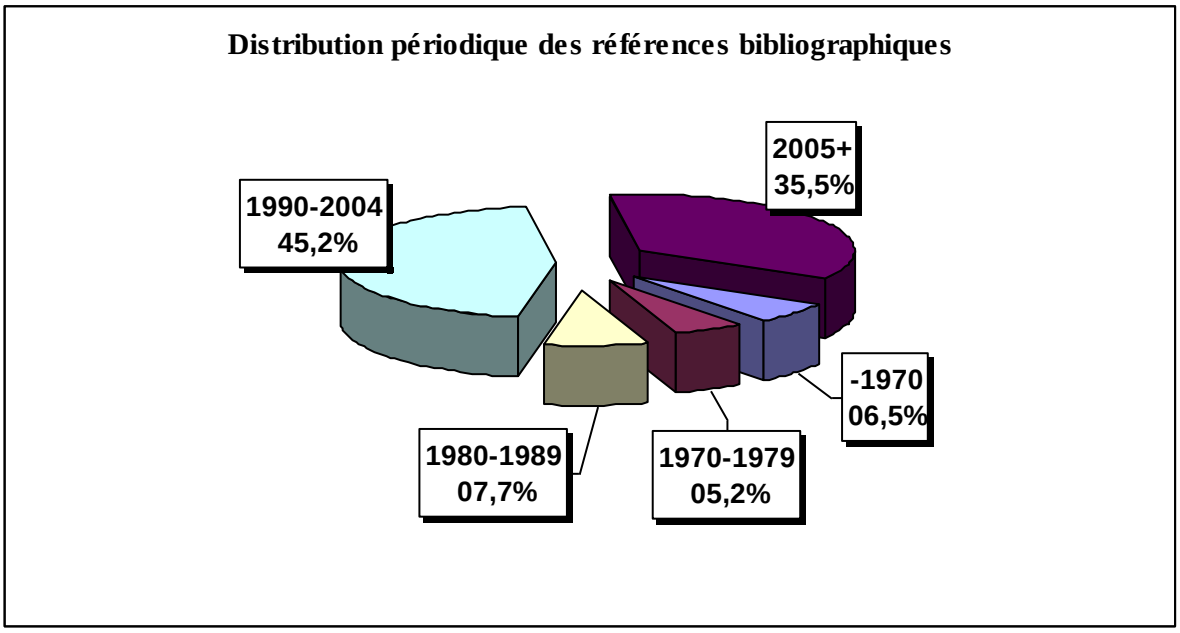
153. Zhou, X., Zhang, X., & Hu, X. (2007). Dragon Toolkit: Incorporating Auto-learned Semantic Knowledge into Large-Scale Text Retrieval and Mining. In proceedings of the 19th IEEE International Conference on Tools with Artificial Intelligence (ICTAI), October 29-31, 2007, Patras, Greece.

154. Zipf G.K. (1949). Human Behavior and the Principle of Least Effort. Harper. (Réédition 1966).

155. Zuccon, G., Azzopardi, L., & Rijsbergen, K.V. (2008). A Formalization of Logical Imaging for Information Retrieval Using Quantum Theory. DEXA Workshops 2008: pp. 3-8.

Analyse temporelle des références bibliographiques (155)





Index

Résumé :

Depuis sa promotion au grand public au début des années 1990, le Web a connu une croissance extraordinaire aussi bien dans son contenu que dans son utilisation. Malheureusement, la nature non-structurée, des larges volumes d'information disponibles sur la toile mondiale, a rendu de plus en plus difficile de cibler et retrouver l'information pertinente. Dans les systèmes classiques de recherche d'information, basés sur les mots-clés, les utilisateurs trouvent souvent des difficultés à exprimer leur besoin d'information. Parmi les nouvelles approches, qui ont été proposés pour promouvoir la recherche intelligente d'information, celle introduisant la dimension sémantique dans la modélisation des documents.

La recherche sémantique sur le Web peut être réalisée selon trois approches principales : (i) Organiser la recherche (indexation de documents et/ou analyse de requêtes) autour de connaissances conceptuelles (thésaurus ou ontologie), (ii) Utiliser un système d'annotation documenté par des experts ou une masse d'utilisateurs pour promouvoir la recherche collaborative, (iii) Développer des méthodes d'indexation sémantique des textes non-structurés. C'est dans cette dernière approche que la présente étude s'inscrit en essayant d'analyser les modèles de thèmes suivant trois axes d'investigation :

1. Quelle est la faisabilité d'utiliser un modèle de thème comme approche d'indexation sémantique des textes pour les tâches de recherche d'information ?
2. Comment évaluer et interpréter le modèle de thème pour l'analyse sémantique du contenu d'une collection ?
3. Dans quelle mesure peut-on appliquer les modèles de thème dans le texte non-structuré non-anglais (l'arabe comme exemple d'étude) ?

Comme contribution majeure dans cette étude, il intéressant de citer :

1. L'analyse et l'évaluation du modèle d'allocation latente de Dirichlet dans les tâches de recherche et de catégorisation des textes sur des corpus réels.
2. La proposition d'une nouvelle mesure, à base de la divergence de Kullback-Leibler, pour le paramétrage de l'apprentissage des thèmes dans une collection donnée.
3. Le développement d'un nouvel algorithme de stemming à base de lemme pour l'analyse et l'indexation du texte arabe.
4. L'élaboration de trois collections arabes, à base d'articles de presse relatifs à la période 2007-2010, pour les expérimentations de tâches de la recherche d'information.

Par ailleurs, les modèles de documents générés, par l'allocation latente de Dirichlet dans un des espaces réduits de thèmes, ont été utilisés efficacement dans la catégorisation des textes et la recherche ad-hoc. En plus, nos travaux ont montré l'efficacité de considérer les aspects morphologiques et les variations typographiques dans l'indexation sémantique des langues hautement flexionnelles telles que l'arabe.

Mots clé :

Recherche d'information, indexation sémantique, modèle de thème, catégorisation des textes, analyse du texte arabe, mesures d'évaluation, collections de test.