Date : 09/08/2006

Les défis de la catégorisation automatique utilisant les systèmes de classification de bibliothèque

Kwan Yi
Ecole des bibliothèques et sciences de l'information
Université du Kentucky
Etats-Unis

Traduction : Annie Milhaud Directrice des Ressources Documentaires, Bayard
Annie.Milhaud@bayard-presse.com

Meeting:	**97 Information Technology with Audiovisual and Multimedia and National Libraries (part 2)**
Simultaneous Interpretation:	No

WORLD LIBRARY AND INFORMATION CONGRESS: 72ND IFLA GENERAL CONFERENCE AND COUNCIL
20-24 August 2006, Seoul, Korea
http://www.ifla.org/IV/ifla72/index.htm

Résumé

Dans l'environnement traditionnel des bibliothèques, un système majeur de classification de bibliothèque ***a*** *pendant longtemps* ***été*** *le cadre standard de classification des sources d'information, et la catégorisation textuelle (TC) est en train de devenir un outil populaire et attractif pour organiser l'information numérique. Cette communication donne* ***un aperçu*** *général des études et projets de catégorisation textuelle (TC) déjà réalisés sur les systèmes majeurs de classification de bibliothèque et résume un débat sur les défis de la recherche en catégorisation textuelle.*

1. INTRODUCTION

L'accroissement énorme de la quantité d'information et de ressources numériques disponibles et la demande d'outils de recherche permettant de gérer la surcharge d'information ont conduit à un intérêt pour les tâches de catégorisation automatique **et donné** l'espoir de réduire le travail humain de façon significative voire de le remplacer dans une proportion limitée. Plusieurs projets de recherche et d'études associées sur la faisabilité de la Classification de la Bibliothèque du Congrès (LCC) et la Classification Décimale Dewey (DCC) comme cadre de classification pour la catégorisation automatique d'information numérique **ont été réalisés**.

Une des approches principales pour organiser l'information est de classer l'information recueillie selon un ensemble prédéterminé de catégories et de retrouver l'information pertinente en parcourant la liste des catégories utilisées. C'est une manière traditionnelle de classifier et de localiser les sujets de bibliothèque **fondée** sur les classifications de bibliothèque. Passées de mode, elles **ont été** remises au goût du jour par l'environnement numérique avec les annuaires thématiques et les annuaires sur le Web. Toutefois, le difficile enjeu de cette approche est dans le manque d'une classification de référence. Adopter une classification de bibliothèque semble être prometteur pour combler l'écart, **d'autant que** s'y ajoutent la popularité pratique en tant que classification et le fondement théorique et systématique. Si l'on considère les publications récentes, le nouveau programme de recherche semble être une zone de recherche prometteuse pour l'organisation et la recherche d'information numérique.

Alors qu'un grand nombre d'études et d'applications associées ou **consacrées à ce thème** ont fleuri, **comme le démontre** la richesse de projets achevés ou en cours de bibliothèques numériques, et les applications réussies d'apprentissage machine dans le domaine de la recherche d'information, la situation a atteint un degré de maturité suffisant pour arriver à un point ou il est approprié d'évaluer les progrès accomplis dans l'évolution de la catégorisation automatique et d'identifier les défis en cours pour établir le programme des recherches pour les années à venir.

La section 2 décrit le contexte de la classification textuelle. La section 3 **fait un** survol des études récentes et des projets portant sur la catégorisation textuelle utilisant les classifications de bibliothèque. La section 4 discute les enjeux courants dans l'organisation de l'information automatique et la section 5 **établit** des conclusions.

2. CONTEXTE

2.1 Comprendre la catégorisation textuelle

En tant que champ émergé relativement récemment dans la recherche d'information, la catégorisation textuelle (TC) consiste à catégoriser les documents selon un ensemble prédéfini de catégories sans assistance humaine. Cette tâche est tout à fait similaire à une tâche de catalogage par sujets dans les bibliothèques traditionnelles, mais très différente si elle est automatique, plutôt que d'être accomplie à la main par des professionnels. La catégorisation textuelle (TC) est devenue plus intéressante que jamais au fur et à mesure que le besoin d'outils d'organisation de l'information pour faire face à une grande quantité d'information numérique se fait plus pressant.

La catégorisation textuelle (TC) **permet d'indexer** des documents **par** les catégories les plus pertinentes prises dans un ensemble de catégories candidates. Les trois composants primaires sont impliqués dans le processus de catégorisation. Le premier composant est le corpus des objets à catégoriser, qui sont des documents textuels (Prenons $D = \{d_1,d_2,\Lambda,d_n\}$comme ensemble de documents). Le second composant est l'ensemble des catégories cibles considérées (Prenons $C = \{c_1,c_2,\Lambda,c_n\}$ comme ensemble de catégories cibles). Le troisième composant est un algorithme de mapping qui agit comme classificateur. Un algorithme de mapping peut être **présenté** comme une fonction prenant un document comme une entrée et produisant une décision binaire si le document tombe dans une catégorie donnée. La fonction est décrite avec $F : d_{i \to \{0,1\}}$ ou $1 \le i \le n$ et $1 \le j \le m$. Le résultat 1 signifie que le document est interprété **comme devant** tomber dans la catégorie considérée, et ne pas tomber dans la catégorie si le résultat est 0 . En conséquence, la réalisation d'une fonction de mapping F et la qualité déterminent la performance de la catégorisation textuelle en même temps que la fonction sert à mesurer la pertinence de l'attribution d'une catégorie à un document.

Les tâches de TC peuvent être divisées en différents types, selon le nombre de catégories et le nombre de noms de catégories. Si l'on considère seulement deux catégories, par exemple si la valeur de *m* dans la classe *C* est égale à 2, une telle TC est dite être une tâche de catégorisation *binaire*. Avec plus de deux catégories, c'est-à-dire que la valeur de *m* est supérieure à 2, on parle de catégorisation à *classes multiples*. De même, si chaque document est associé à seulement une catégorie, on parle de catégorisation *binaire*. Dans les catégorisations multiples, il y a au moins une catégorie associée à chaque document.

2.2 Apprentissage machine dans les applications de TC

La préoccupation principale dans la recherche en TC **est de savoir** comment une *machine* (on se réfère généralement à un système informatique, mais le terme *machine* est utilisé par convention) acquiert la connaissance nécessaire supposée pour une catégorisation correcte. L'approche la plus répandue pour ce problème est l'apprentissage machine (ML). Le canevas général d'apprentissage de ML est que la machine acquiert la connaissance à partir d'expérience précédente. On peut décrire un canevas de ML comme un processus systématique comportant quatre composants (Kubat, Bratko, et Michalski 1999) : *expériences, connaissance du contexte, algorithme d'apprentissage, et connaissance de la cible Les expériences* comme données du canevas sont censées relayer ce qui doit être appris comme *connaissance cible* appelée connaissance *explicite. L'algorithme d'apprentissage*, boite noire du canevas ML, représente la méthode d'apprentissage de la connaissance. *La connaissance cible* est la réalisation de ce qu'un modèle de ML apprend de la combinaison de connaissance explicite et implicite. Par exemple, si l'on considère les échecs comme une tâche d'apprentissage, une séquence des positions de l'échiquier changées en pratique peuvent être des exemples pour la tâche échecs et les règles générales des échecs seraient la connaissance du contexte. L'approche ML est similaire jusqu'à un certain point au processus d'apprentissage chez l'être humain. De même que les êtres humains acquièrent la connaissance en lisant des textes, une machine de ML acquiert aussi la connaissance d'un sujet ou d'une catégorie à partir de documents présélectionnés par des spécialistes du domaine. Une telle collection de documents est nommée un ensemble d' *entraînement* qu'il soit implicite ou explicite.

Les techniques ML ont été appliquées à la catégorisation de différents types de documents : documents **sur la** santé (Larkey and Croft 1996), données botaniques (Cui, Heidorn, et Zhang 2002), documents juridiques (Thompson 2001), documents issus du web (Chakrabarti, Dom, et Indyk 1998). On **a aussi** recherché d'autres types de catégories, différents du sujet ou de la

matière, **soit, par exemple**, des documents éditoriaux, des rapports, des articles de synthèse, des **articles de recherche**, et des **pages d'accueil de** site (Lee et Myaeng 2002), des essais portant sur plusieurs disciplines (Larkey 1998), un filtrage des messages de spam (Hidalgo, Lopez, et Sanz 2000).

2.3 Périmètre de la catégorisation textuelle

Le périmètre de la discussion en TC **concerne généralement les points suivants** :

- TC et *regroupemen*t sont des mécanismes similaires mais différents en matière de catégorisation de documents. Le concept de *regroupement* est apparenté à celui de TC, **puisqu'i**l rassemble plusieurs documents. Un regroupement est défini comme *"le groupe de documents qui répondent à un ensemble de propriétés communes"* (Baeza-Yates et Ribeiro-Neto 1999). **Lors d'un** regroupement, on ne prend pas en considération tout ensemble explicite de catégories. A l'inverse, on recherche des caractères communs inconnus. Par conséquent, en TC, le degré de similarité entre un document et une catégorie cible est mesuré pour voir à quel point le document est pertinent par rapport à la catégorie. **Lors d'un** regroupement, les documents similaires ne sont pas mesurés **par rapport** à des catégories cibles, mais **par rapport** à d'autres documents.

- Les éléments non textuels (autres indices que le texte) ne sont pas pris en compte

- La TC est une catégorisation **fondée** sur le contenu ; elle n'est fondée ni sur les métadonnées ni sur l'information structurée. La TC et la catégorisation de documents issus du web sont distincts dans l'usage de métadonnées autres que les contenus web, tels que liens hypertexte, et données structurées, dans les classifications sur le web.

3. SURVOL DES PROJETS ET ETUDES DE CATEGORISATION TEXTUELLE AVEC CLASSIFICATIONS DE BIBLIOTHEQUE

Nous passerons en revue quelques tentatives de catégorisation automatique de documents numériques utilisant les principales classifications de bibliothèque.

L'un des premiers travaux en catégorisation automatisée utilisant des classification de bibliothèque est **celui de** (Larson 1992) **où** un ensemble de fiches MARC **a été** catégorisé dans la classe Z (bibliographie et bibliothéconomie) de la LCC, fondée sur le titre et sujet avec 30 471 fiches MARC pour l'apprentissage et 286 fiches MARC pour les tests. On supposait que ce travail aiderait les bibliothécaires à déterminer des cotes pertinentes pour des éléments non classés en fournissant une liste de cotes potentielles **fondées** sur le titre et le sujet. Le travail le plus récent lié à celui de Larson est **celui de** (Frank et Paynter 2004). Leur travail vise à attribuer des cotes LCC à des métadonnées de ressources Internet en utilisant la LCC et la liste de sujets de la bibliothèque du Congrès (LCSH). Le classificateur apprend en utilisant 800 000 fiches du catalogue de la bibliothèque et est testé sur un ensemble indépendant de 50 000 fiches. La précision de catégorisation de ce classificateur va de 55% à 80%.

Dans les sections suivantes, on passera en revue un certain nombre de projets de TC ou les classifications traditionnelles de bibliothèque **ont été** adoptées comme base de catégorisation de documents numériques.

3.1 Pharos

Pharos est un prototype de système d'information agrégeant des sources diverses en contenu et en format, dérivé du projet de bibliothèque numérique Alexandria (Dolin, Agrawal, et El Abbadi 1999). Comme prototype initial à Pharos, **on a implémenté** un système de catégorisation automatique f**ondé** sur la LCC **pour** créer les profils d'information numérique hétérogènes. Dans ce projet, la technique d'indexation sémantique latente **a été** utilisée pour catégoriser des groupes de discussion et des entrées de catalogue utilisant la LCC. **On n'a utilisé** comme ensemble de données d'entraînement, 1,5 millions d'entrées du catalogue de la bibliothèque de l'Université de Californie Santa-Barbara, on **a procédé** à l'extraction des champs titre, sujet et cote LCC. **A titre d'opération** spécifique, les données titre et sujet sont vues comme description d'une catégorie spécifique représentée par une cote LCC. Une telle relation entre une cote LCC et ses descripteurs forme une donnée d'apprentissage pour le système de catégorisation. On **a catégorisé** 7214 entrées MARC dans les 21 classes majeures de la LCC, et les résultats expérimentaux **ont donné** une médiane moyenne de 13.0 ±3,9 et une moyenne de 76± 19 pour à peu près 4 200 classes. Dans une autre expérience avec des articles tirés de 2500 groupes de discussion sur Usenet, la précision de catégorisation pour l'expérience n'est pas rapportée car les articles concernés n'étaient pas pré classifiés.

3.2 Scorpion

Scorpion est un projet de recherche mené par l'OCLC (Online Computer Library Center) de 1996 à 1999 **pour** développer une méthode automatisée d'identification des classes DDC de documents numériques (Shafer 2001). Pour sa catégorisation automatique, Scorpion **a utilisé** une méthode de regroupement. **Pour** un document d'entrée, Scorpion mesure ses similarités **avec** les regroupements pré définis (correspondants aux classes DDC) et considère le regroupement le plus proche comme l'emplacement le plus probable pour le document d'entrée. On utilise un décompte de termes comme mesure de similarité. Pour l'évaluation, on a utilisé une collection de notices bibliographiques de ressources Internet indexées par des humains avec les classes DDC. Malheureusement, malgré cela, les résultats détaillés de l'expérience **n'ont pas été** révélés, vraisemblablement parce qu'il **n'a pas été** possible d'établir une comparaison appropriée car l'indexation humaine avec les classes DDC reposait seulement sur des expressions décrivant les ressources Internet. Leurs conclusions ont **confirmé** que la catégorisation automatique ne peut pas remplacer la classification manuelle, mais qu'elle peut fournir une solution économique pour soutenir les catalogueurs humains.

3.3 DESIRE (Koch et Ardö 2000)

Le projet DESIRE, commencé en 1996, est un projet international à grande échelle financé par l'Union Européenne dans le but de construire un portail thématique pour les ressources sur les sujets d'ingénierie. Dans une expérience, des documents Internet **étaien**t automatiquement catégorisés dans la classification Information Ingénierie (EI), **fondée** sur une correspondance terme à terme. Dans l'évaluation du système avec à peu près 100 pages Web, la précision de la catégorisation automatique **a été** comparée aux décisions de catégorisation de l'équipe. On **a découvert** globalement que 60% des catégorisations automatiques étaient correctement ou plus finement appariées aux décisions humaines. Avec la collaboration de l'OCLC, les mêmes données sur l'ingénierie **ont été** catégorisées avec la DDC. Plus **précisém**ent, quelques notices LCSH **ont été** ajoutées en plus du plein texte. Il n'y eut pas de rapport sur une évaluation de la catégorisation basée sur la DDC.

3.4 La Bibliothèque Internet de Wolverhampton (Jenkins et al.)

La Bibliothèque Internet de Wolverhampton (WWLib)[1] est un projet de moteur de recherche pour des documents britanniques **qui** utilise la DDC pour organiser les documents collectés. Un trait intéressant de la WWlib expérimentale est de considérer une page web comme un élément d'une bibliothèque et de préparer des notices catalographiques décrivant des informations incluant le titre, l'adresse URL, la classe DDC, et la description dans les pages web collectées. En général, les moteurs de recherche sur Internet présentent les résultats par ordre de pertinence par rapport à la question de l'utilisateur, alors que le WWLib fournit les pages pertinentes par rapport à la classification DDC. Les composants du classificateur accomplissent le processus de catégorisation des documents web automatiquement, lequel repose sur une correspondance terme à terme. Le classificateur de la WWLib compare un flux de mots extraits de documents à la description de classes de la DDC (Wallis & Burden, 1995). L'occurrence de mots dans les documents web est pondérée selon leurs identifiants, on applique une technique de lemmatisation. **En outre**, pour utiliser l'avantage de la structure hiérarchique de la DDC, la méthode pren**d** en compte la pertinence d'un document à la fois par rapport à sa classe et par rapport à la classe père. Dans les dernières versions (WWLib), on a pris en considération un ensemble plus riche de descriptions de classes de la DDC **en** incluant des synonymes. Il semble qu'une expérience formelle de mesure de la performance du système n'ait pas été entreprise. Au lieu de cela, **on a signalé** un résultat de test informel avec les 17 URL (WWLib) sélectionnées au hasard où 13 cases sur 17 furent simplement déclarées pertinentes sans donner plus de détails de procédure tels que la méthode d'évaluation ou la sélection des données.

3.5 Résumé

D'un point de vue applicatif, nous avons estimé que la plupart des projets de recherche de catégorisation automatique de données catalographiques et de pages internet ne s'appliquaient pas à des documents en texte intégral. Du point de vue des classifications, soit la LCC ou la DDC, toutes deux parmi les plus populaires des classifications de bibliothèques utilisées en Amérique du Nord, ont été utilisées dans les applications de catégorisation. Néanmoins, la justification **du choix d'**une classification de bibliothèque n'est pas clairement expliquée. Selon les descriptions données dans les articles, la décision de choisir une classification de bibliothèque semble **fondée** sur une préférence personnelle plutôt que sur les tâches proches ou la disponibilité des données.

4. LES DEFIS DE LA CATEGORISATION TEXTUELLE POUR LES CLASSIFICATIONS DE BIBLIOTHEQUE

La discussion des défis de la TC est **interrompue ??? (surmontée)** par les composants du processus de TC

4.1 Classifications

Les classifications de bibliothèque **ont été** développées à l'origine pour organiser des documents imprimés primaires tels que des livres et des périodiques et sont utilisées dans l'organisation traditionnelle des bibliothèques depuis plus d'un siècle. Récemment, l'utilisation de ces classifications s'est **étendue à** l'environnement en ligne pour organiser

[1] http://www.scit.wlv.ac.uk/wwlib

l'information numérique ou le rôle potentiel de la classification de bibliothèque **a été** exploré comme outil pour organiser, parcourir et accéder à l'information. Plusieurs classifications de bibliothèques ont été utilisées dans divers projets et études (Koch et Day 1997), telles que la LCC, la DDC, celle de la bibliothèque Nationale de Médecine (NLM), et la Classification Décimale Universelle (CDU).

4.1.1 Couverture et caractéristiques des catégories

Le premier défi vient de la taille et l'instabilité des catégories. Il y a approximativement 100000 classes différentes dans la LCC et le nombre de classes de la DDC n'en est pas loin. **De ce fait** préparer les données d'entraînement pour chaque catégorie et construire un système de TC correspondant à chaque classe ne semble pas logiquement possible. **En outre**, Les classifications ne sont pas statiques dans le temps, mais font l'objet de mises à jour y compris une mise à jour des classes existantes. Il semblerait **aussi que** toutes les classes précisées dans les classifications ne soient pas utilisées.

Le deuxième défi réside dans la dissemblance des classifications. Les classifications universelles de bibliothèque ont en commun le fait que le sujet est la caractéristique principale pour la classe. **Mais** elles sont très différentes dans leur structure et dans le système de notation retenu.

Pour faire face à ces problèmes, on doit prendre en considération les éléments suivants.

- Déterminer les limites des niveaux de classes selon le sujet général de l'application de TC

  Des applications de TC différentes peuvent être intéressées par différents niveaux ou ensemble de classes. Une application sur l'histoire peut avoir un intérêt pour les classes en histoire alors que l'utilisation de toute la gamme des classes est plus attractive pour une application telle qu'un répertoire sur le web.
- **Mettre en œuvre** les caractéristiques structurelles de la classification

  Dans les classifications, les relations entre classes se reflètent dans leur structure hiérarchique, **ce qui signifie** que la DDC est une classification dans laquelle une classe dans un niveau indique une discipline ou un sujet plus général qu'une classe dans le niveau inférieur. La nature de la hiérarchie de la LCC est similaire à celle de la DDC. Un ensemble de classes principales en tête de hiérarchie représente une liste de disciplines, chacune d'entre elles est divisée en sous classes pour les disciplines plus spécifiques, à l'exception des classes E et F (Histoire de l'Amérique), et Z (bibliographie et bibliothéconomie). Alors, les subdivisions suivantes sont généralement faites par sujet, lieu, temps et forme. De cette manière, les données d'entraînement pour les niveaux inférieurs peuvent aussi être utilisées pour les niveaux supérieurs.

4.2 Les sources de connaissance

Pour faire simple, la connaissance cible acquise par la machine provient directement des données d'entraînement établies en entrée, et la connaissance résultante acquise par le processus de TC peut être affecté**e** de façon inhérente par celle des données d'entraînement. C'est ainsi que des cibles de système de TC aboutissent à des connaissances acquises différentes lorsque des données d'entraînement différentes sont utilisées.

En général, l'acquisition de données d'entraînement est réputée difficile, donnant une importante charge de travail ; **elle peut être coûteuse, voire** infaisable dans certains cas. Les données d'entraînement consistent en *expériences* et connaissance contextuelle. La connaissance contextuelle est un ensemble de données générales applicables à un sujet plus large alors que l'expérience est considérée comme un sujet ou une classe spécifiquement orientée vers la tâche.

4.2.1 Des données d'entraînement plus explicites comme plateforme de test

La recherche en TC fondée sur l'apprentissage machine (ML) dépend des données d'entraînement. Il y a aujourd'hui peu de données d'entraînement standard utilisées comme *benchmark* pour la recherche. Développer plus de données d'entraînement sur des sujets variés devrait permettre des recherches productives et conduire à des améliorations majeures en TC.

4.2.2 Développer la connaissance contextuelle

Souvent confrontée à la rareté des données d'entraînement et à la difficulté d'y accéder, la recherche s'efforce de développer des outils pour lesquels la génération de connaissance contextuelle est significative. Les langages contrôlés et les thésaurii sont des collections de termes autorisés pour des sujets, et représentent aussi certains types de relations entre termes. L'usage potentiel des définitions et relations pour la connaissance contextuelle apparaît proéminente. Aussi, **la mise en œuvre** de l'intégration et des liens entre divers outils d'organisation de la connaissance devrait s'améliorer de manière significative.

4.2.3 Génération automatique de données

Une alternative à la collecte manuelle de données serait très utile car la génération de données est le processus le plus coûteux en TC. L'environnement de l'information est mature pour cela : la disponibilité d'un vaste volume d'information numérique ; un large spectre d'outils et de techniques de traitement de l'information développés dans le domaine de la recherche d'information. Les ressources numériques pré-classifiées par des professionnels peuvent souvent être trouvées dans des systèmes d'information tels que les annuaires sur le web et les bases de données en ligne. L'utilisation de documents fortement pré-contrôlés peut être une option pour une nouvelle collection de données.

Dès qu'un système de TC se me**t** en place, il pren**d** des documents non classifiés en entrée et produit des documents catégorisés en sortie. Une autre possibilité est la réutilisation des documents catégorisés comme ensemble de données d'entraînement.

4.2.4 Intégration de sources multiples

Les modèles et les outils **permettant d'**incorporer de multiples sources telles que la connaissance contextuelle et la connaissance explicite à travers des chemins variés peuvent résider de façon significative dans la synthèse d'indices pour établir des classes pertinentes.

4.3 Techniques / Modèles de catégorisation

Un nombre croissant de chercheurs provenant de divers champs d'études, principalement en informatique et sciences de l'information, ont montré de l'intérêt pour le développement d'outils et de méthodes de catégorisation de documents automatique basée sur le texte. Un

large spectre de techniques et d'algorithmes d'apprentissage inductif, tels que les machines à support vectoriel, les réseaux bayésiens, les arbres de décision et les réseaux neuronaux artificiels, ont été proposés et testés, (Yiming 194 ; Joachims 1998 ; Lewis et Ringuette 1994 ; Mitchell 1997). La recherche en TC s'est concentrée massivement sur le développement de méthodes et d'outils efficaces et d'algorithmes d'apprentissage (Sebastiani 2002).

Des modèles de classification différents soutiennent des représentations différentes de la connaissance et adoptent des méthodes d'apprentissage différentes. Dans les algorithmes de réseaux neuronaux, la connaissance est représentée comme un graphe comportant des nœuds et des côtés, et dans les règles d'induction, on utilise des règles action-condition. Dans d'autres méthodes les fonctions, les programmes logiques et les ensembles de règles, les automates à états et transition, les grammaires et les systèmes de résolution de problèmes ont été adoptés pour représenter la connaissance.

4.3.1 Techniques d'indexation sémantique et modèles de catégorisation

Une des finalités de la catégorisation textuelle peut être la pleine compréhension du sens de documents textuels. Ce défi a longtemps été l'objet des recherches d'abord en linguistique informatique et compréhension du langage naturel, depuis les débuts des traitements automatique des documents dans les années 50.

La TC, en tant que subdivision de la recherche d'information (IR), a adopté diverses techniques et méthodes de l'IR parmi celles existantes. Récemment, les chercheurs en traitement de l'information se sont tournés vers les outils et les méthodes d'apprentissage fondés sur l'intelligence artificielle provenant des réseaux neuronaux, l'apprentissage symbolique, et les algorithmes génétiques. Les techniques et les méthodes probabilistes pour la TC ont été plus attractives dans la décennie précédente.

La plupart des modèles de textes montrent des aspects relativement simples du langage (mots, expressions, noms), et les modèles pour indexer les termes reposent sur un mécanisme simple de comptage tel que fréquence et co-occurrence. De tels modèles ne prennent pas en compte les aspects de structures et de relations sémantiques qui pourraient être plus importants pour caractériser les catégories en matière, sujet, etc. On peut trouver des exemples typiques des problèmes relevés dans les méthodes non sémantiques avec la synonymie (plusieurs mots ayant le même sens) et la polysémie (un mot ayant plusieurs sens).

De nombreux efforts ont été tentés en IR pour résoudre ces problèmes en partie ou en totalité. Une voie prometteuse semble être d'explorer des modèles incorporant des termes d'autorités contrôlés et les relations entre eux.

5. CONCLUSION

Durant le siècle précédent, le rôle des classifications de bibliothèque a été élargi en tant qu'outils pour localiser les documents sur les étagères, pour les retrouver dans les catalogues OPAC (Online Public Access Catalog), et maintenant pour organiser et accéder aux ressources numériques dans des environnements en réseaux. L'adoption de classifications traditionnelles est séduisante, prometteuse et riche de potentiel pour les raisons suivantes : (1) Les grandes classifications de bibliothèque ont été des cadres d'organisation de l'information très populaires ; (2) Un très riche ensemble d'outils d'organisation soutenant et associés avec les classifications de bibliothèque, tels que les vocabulaires contrôlés et les listes d'autorités

ont été développés et sont disponibles ; (3) La description bibliographique des sources d'information telles que les notices catalographiques contiennent l'association entre source d'information et outils bibliographiques utilisés.