Un modèle pour la recherche d'information sur des documents structurés

Benjamin Piwowarski, Ludovic Denoyer, Patrick Gallinari

LIP6 – 8 rue du capitaine Scott – 75015 PARIS – FRANCE –
{Benjamin.Piwowarski,Ludovic.Denoyer,Patrick.Gallinari}@lip6.fr

Abstract

As new standards, like XHTML or DocBook where documents structure is clearly present, are promised to a great development, the information retrieval (IR) community has become more and more concerned to use this new source of evidence. This task is hard to achieve as it aims at working with two different sources of information, the structure and the textual information. Since then, a few models appeared but they lack maturity and were not completly thought to use the full information contained in the structure. The model we present, which is based on bayesian networks, has the pretention to overcome these limits. Although it is a very experimental model, the theoric framework we provide deals with two information retrieval tasks (categorization and ad-hoc retrieval), can adapt itself to new databases with machine learning algorithms. It also can be extended to cope with new tasks as interactive navigation.

Résumé

Avec l'émergence de nouveaux standards comme le XHTML ou le DocBook où la structure des documents est apparente, la communauté de recherche d'information a commencé à s'intéresser à l'utilisation de cette nouvelle source d'information. La tâche est ardue, car il s'agit de concilier de sources d'informations de natures différentes, à savoir le texte et la structure. Quelques modèles ont fait leur apparition ; mais ces travaux manquent encore de maturité et n'utilisent la structure que d'une manière simple. Le cadre théorique que nous présentons dans ce papier a pour vocation de permettre une prise en compte de la structure dans les tâches de recherche documentaire et de catégorisation. Ce modèle basé sur l'utilisation de réseaux bayésiens est capable de s'adapter à de nouvelles bases de données grâce à des techniques d'apprentissage numérique. Il offre également des perspectives de développement intéressantes comme par exemple la navigation interactive dans une base de données.

Mots-clés : information retrieval, categorization, ad-hoc retrieval, structured document, bayesian network

1. Introduction

La récente explosion du Web donne accès à une quantité d'informations énorme et sans cesse croissante qui se présente majoritairement sous forme textuelle (courriers électroniques, pages HTML...). Le traitement automatique des documents textuels est le domaine de la Recherche d'Information (RI) ou Recherche Documentaire (RD). Les systèmes de RI commercialisés de nos jours ont conceptuellement peu évolué depuis la fin des années 1970. Ils utilisent des représentations généralement très simples des textes et proposent des outils de base pour accomplir des tâches comme la caractérisation de la pertinence de textes par rapport à des requêtes, leur classement en différentes catégories ou le filtrage.

Ces dernières années ont vu d'importants changements à la fois dans la nature des documents textuels et des corpus, et dans les besoins utilisateurs. Les corpus sont devenus beaucoup plus gros et sont souvent composés de documents hétérogènes aussi bien dans leur forme que dans

leur contenu. De nouveaux standards de représentation des documents ont commencé à se développer en liaison avec le Web et avec les bases de données. En particulier, les représentations structurées se développent avec la proposition de standards issus du langage XML, comme par exemple le XHTML (World Wide Web Consortium, 1998) pour les ressources du Web et le DOCBOOK (OASIS, 2001) issu d'OASIS (Organization for the Advancement of Structured Information Standards) pour tous les documents. L'utilisation de descriptions structurées pour les documents s'est rapidement développée et est en train de s'imposer.

Paradoxalement, la RI apporte peu d'outils pour le traitement de ces documents structurés : les représentations des document en RI ont été conçus pour des documents "plats" et homogènes, ils ne sont pas adaptés au traitement simultané de la structure et du contenu – pas plus qu'à la prise en compte d'informations de natures différentes ou d'importances inégales telles qu'elles peuvent apparaître dans une description structurée. C'est le problème du traitement simultané de la structure et du contenu que nous étudions ici. Nous nous plaçons dans le cadre de la RI ; nous proposons un modèle probabiliste basé sur les réseaux bayésiens, qui nous permettent d'envisager des tâches d'inférence [1] et d'apprentissage sur des corpus structurés.

Cet outil a été développé pour deux tâches classiques de RI, à savoir les réponses aux requêtes ouvertes et la classification de documents (requêtes fermées). En (2), nous présenterons la problématique et ferons un état de l'art sur le traitement de documents structurés en RI en (3). Nous introduirons notre modèle en (4) et des expériences préliminaires en (5).

2. Problématique

Les systèmes de recherche d'information actuels se basent sur une représentation du texte dite "sac de mots" où les documents sont représentés simplement par la présence, l'absence ou la fréquence d'apparition des termes. Quelques systèmes ont commencé à prendre en compte l'information séquentielle [2] présente dans les textes pour des tâches spécifiques de la RI (Denoyer et al., 2001). Pourtant, avec l'apparition de nouveaux standards comme le XML et son dérivé le XHTML pour les pages Web, la tendance actuelle est d'une part de séparer la structure logique d'un document de son apparence et d'autre part de lier structure logique et contenu. Un texte sera ainsi représenté par une information plus complète que son simple contenu textuel, il sera accessible à différents niveaux de description et pourra contenir différents types d'information (e.g. texte et meta-données). L'accès simultané à la structure des documents et à leur contenu permet d'envisager de nouveaux modes d'exploitation de l'information textuelle. Celle-ci nécessite la création de nouveaux outils et de nouveaux modèles capables d'exploiter de telles données.

3. État de l'art

Différents travaux concernant l'utilisation de l'information de structure pour la RI ont déjà été développés. Il est nécessaire de différencier les travaux concernant la tâche de recherche documentaire de ceux concernant la tâche de classification. En effet, ils reposent sur deux notions très différentes de la structure. Pour mémoire, nous rappelons brièvement en quoi consistent ces deux tâches. La recherche documentaire (ad-hoc retrieval) a pour but de trouver, parmi un

1. ce qui correspond à s'intéresser à la pertinence d'un document, ou d'une partie d'un document, pour une requête ou une catégorie donnée

2. Un mot dépend de son contexte local

ensemble de documents, celui ou ceux qui répondent le mieux à une requête exprimée en langage naturel. La classification est utilisée dans plusieurs tâches de RI. Elle consiste à attribuer à chaque document une catégorie, parmi un ensemble de classes connues à l'avance (e.g. économie, science ou littérature). Il s'agit d'une tâche de discrimination entre deux ou plusieurs classes.

3.1. Recherche documentaire et documents structurés

Un premier travail où la structure est directement employée dans une tâche de recherche documentaire est celui de Wilkinson (Wilkinson, 1994) qui a étudié sur une sous-partie de la base de données TREC (Text REtrieval Conference) l'influence de la prise en compte du découpage d'un document en sections, et de la nature des sections. Les expériences montrent que la structure apporte une information supplémentaire, même s'il faut rester prudent sur les conclusions vu la petite taille de la base de données utilisée.

Une autre approche est celle qui a été suivie à l'université de Glasgow par Lalmas et al., et qui a pour base la théorie de l'évidence de Dempster-Schafer (Lalmas et al., 1997; Lalmas, 1997; Lalmas and Moutogianni, 2000). Pour tout document et toute requête, il est possible de définir deux mesures : la croyance et l'incertitude. Un opérateur d'aggrégation permet de combiner la croyance de ces sous-structures pour calculer la pertinence d'un document.

Depuis Inquery (Callan et al., 1992; Turtle and Croft, 1991), les réseaux bayésiens (RB) ont fait leur apparition dans le domaine de la recherche documentaire. Le modèle de RB très simple utilisé dans Inquery permet de calculer la probabilité qu'une requête soit satisfaite par un document [3]. Ce modèle est utilisé dans un cadre standard de RD, *i.e.* pour des documents "plats". Les modèles de (Ribeiro and Muntz, 1996) et de (Indrawan et al., 1994) proposent des approches légèrement différentes. Une extension du modèle développé pour INQUERY proposée par (Myaeng et al., 1998) prend en compte la structure des documents. Un document est représenté par un arbre. A chaque nœud de l'arbre correspond une partie de la structure de ce document (section, chapitre, etc.) et le texte qui lui est associé. A la différence D'INQUERY, ce réseau est donc directement calqué sur la structure du document et l'information "descend" du document pour arriver jusqu'aux termes. Lorsqu'une requête arrive, on calcule pour tout document D la probabilité que les termes de la requête représentent bien le document. Pour cela, il est nécessaire de calculer la probabilité qu'une section représente bien le document D, puis qu'un terme représente bien une section, et enfin qu'une question représente bien ce terme. Les auteurs procèdent à de nombreuses approximations pour rendre les calculs réalisables ce qui limite la portée du modèle.

D'autres approches considèrent également le cadre de requêtes structurées dans le cas où l'information textuelle est réduite à la présence ou l'absence d'un terme. La plus connue est celle des Proximal Nodes (Baeza-Yates and Ribeiro-Neto, 1999). L'accent est mis, dans ces modèles, sur les problèmes de complexité dans de grands corpus.

3.2. Classification

En ce qui concerne la tâche de classification, (McCallum et al., 1998; Chakrabarti et al., 1997; Koller and Sahami, 1997) considèrent comme structure une hiérarchie sur les classes des documents, et utilisent cette hiérarchie pour réaliser la classification. La structure est abordée de

3. Pour reprendre leur terminologie, on dira que le document *représente bien* la requête

deux manières différentes

- *Combinaisons de classifieurs*. (Chakrabarti et al., 1997),(Koller and Sahami, 1997) présentent des modèles assez proches. Ils associent à chaque nœud de l'arbre, un classifieur "spécialisé" dans la discrimination des classes fils de ce nœud. Ainsi, un document " descend" dans l'arborescence jusqu'à maximiser sa probabilité de pertinence en passant successivement dans plusieurs classifieurs.
- *Shrinkage*. (McCallum et al., 1998) utilise également une hiérarchie de classes, ils associent une distribution de probabilité à chaque nœud de cette hiérarchie qui modélise la distribution des mots dans la classe correspondante. Ils estiment les distributions locales à chacun des nœuds, puis définissent la distribution associée à un nœud comme une combinaison linéaire de cette distribution locale et des distributions associées aux parents du nœud dans la hiérarchie. Les poids de la combinaison sont appris. Cela permet d'obtenir des estimateurs plus robustes en particulier pour les classes faiblement représentées.

4. Présentation du modèle

Le modèle que nous présentons s'applique aussi bien à la recherche documentaire qu'à la catégorisation. Nous nous sommes basés sur une modélisation probabiliste qui est celle des réseaux bayésiens. Ce formalisme permet de façon assez naturelle de traiter des structures, et permet de faire des inférences assez complexes sur les différents éléments de la structure. L'originalité de notre modèle tient en plusieurs points. C'est un modèle générique, qui peut être appliqué à différentes tâches de la RI comme la recherche documentaire et la catégorisation; les paramètres du modèle sont estimés directement à partir du corpus contrairement à tous les autres modèles qui traitent des documents structurés où les paramètres sont fixés à la main; le modèle peut être étendu à d'autres tâches de la recherche d'information.

4.1. Les réseaux bayésiens

Les réseaux bayésiens (Pearl, 1988; Jensen, 1996; Krause, 1998; Murphy, 2000) sont *un formalisme probabiliste* qui exploite des relations d'indépendance conditionnelle entre différentes variables aléatoires caractérisant un même phénomène. Ces relations d'indépendance, sont en général basées sur les connaissances que l'on a *a priori* sur le problème[4]. Elles permettent de simplifier l'expression des probabilités jointes ou conditionnelles. Nous allons illustrer cela sur un exemple. Considérons un document (figure 1) comportant deux sections et trois paragraphes. Associons à chaque entité structurelle du document une variable aléatoire binaire qui indique la pertinence/non-pertinence de cette entité pour une requête ou une catégorie. Pour effectuer des inférences avec ce modèle (e.g. calculer la pertinence du document ou d'une section), il nous faudra connaître la distribution de probabilité $P(d, s_1, s_2, p_1, p_2, p_3)$[5] où d, s et p représentent respectivement les variables document, section et paragraphe. Pour peu que l'on ait a traiter de grands documents, ce type de modèle est clairement impossible à gérer.

Pour contourner cette difficulté, nous ajoutons des hypothèses d'indépendance qui permettent de simplifier le problème tout en modélisant au mieux la tâche abordée. Dans la figure 1(b), les sections 2 et 3 sont les *parents* de la variable "document" : la probabilité que le paragraphe 1 soit pertinent ne dépend pas du reste des variables. La probabilité que la section 2 soit pertinente ne

4. nous ne considérerons pas l'apprentissage de la structure des RB qui est en soit un problème complexe
5. Ce qui représente $2^6 - 1$ valeurs de probabilités

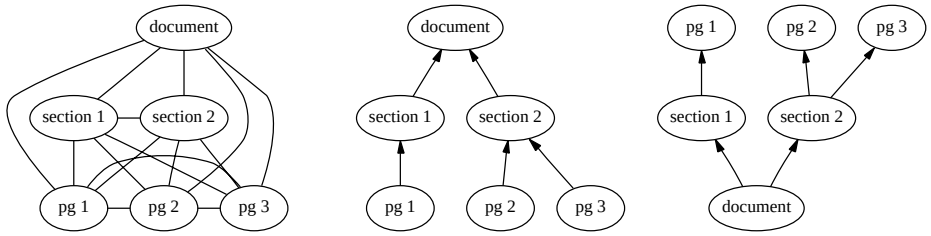


FIG. 1: *Trois modélisations d'un même document avec un réseau bayésien - (a) toutes les parties sont dépendantes, (b) et (c) représentent deux modélisations qui encodent différentes dépendances conditionnelles entre parties.*

dépend que de la pertinence de ses paragraphes (2 et 3). De même, la probabilité que le document soit pertinent ne dépend que de la pertinence de ses sections. Avec une telle modélisation, l'information sur la pertinence (d'un document pour une catégorie donnée) remonte des entités structurelles les plus fines (le paragraphe) pour se propager jusqu'à l'entité structurelle la plus grosse (le document).

Dans ce cas, $P(d, s_1, s_2, p_1, p_2, p_3) = P(d|s_1, s_2)P(s_1|p_1)P(s_2|p_2, p_3)P(p_1)P(p_2)P(p_3)$.

Avec une telle modélisation, la complexité passe de $O(2^N)$ où N représente le nombre de variables aléatoires à $O(2^{N_{\max}})$ où $N_{\max}$ est le nombre maximum de parents pour un nœud du réseau. Dans un réseau bayésien, on associe à chaque nœud une distribution conditionnelle (e.g. $P(s_1|p_1)$). Cette distribution est soit fixée à la main, soit apprise à partir d'exemples. Là encore, les hypothèses d'indépendance simplifient considérablement les problèmes d'estimation. Bien sûr, d'autres modélisations d'un texte sont possibles avec les RB. La modélisation que nous proposons est assez générale.

4.2. Réseaux bayésiens pour la structure : présentation des idées du modèle

Nous allons présenter les principales idées de notre modèle. Pour une tâche donnée, et pour un type de structure (e.g. la structure logique définie dans le modèle du document ou encore la structure hiérarchique en sections et paragraphes) nous allons construire un réseau bayésien par document permettant de modéliser la pertinence de ce document pour la tâche considérée. Ce réseau sera construit directement à partir de la structure considérée. Par apprentissage, nous estimerons les paramètres de nos réseaux (lois conditionnelles), ces paramètres seront partagés par tous les réseaux bayésiens construits pour la tâche. Voici les hypothèses utilisées (nous expliquerons plus loin les différences entre la RD et la classification).

Les variables aléatoires sont des variables binaires à valeur dans l'ensemble {Relevant, Irrelevant}. Chaque variable est représentative de la pertinence d'une partie du document pour la tâche considérée. Par exemple, pour la figure 1, la variable associée à section 1 représentera la pertinence de cette section, celle associée à paragraphe 2 la pertinence du paragraphe 2 et celle associé à d celle du document tout entier. L'évidence dans ce modèle sera obtenue en calculant la pertinence d'une entité. Sans que cela soit restrictif, nous calculerons cette pertinence uniquement pour les feuilles de l'arbre. Pour cela, nous utiliserons des mesures classiques dans le domaine comme par exemple le TF-IDF dans le cas des requêtes ouvertes et le modèle Naive-Bayes dans le cas de la classification.

Nous considérerons uniquement des modèles de type arbre. Dans le cas de l'exemple de la figure 1(b), la pertinence d'une section dépend alors uniquement de la pertinence des paragraphes qui composent la section et la pertinence du document dépend de la pertinence des sections. Nous ne considérons pas les dépendances entre nœuds frères. Cette hypothèse permet de simplifier suffisament le modèle pour qu'il soit calculable, au prix d'une modélisation moins complexe de la réalité.

Les paramètres de nos réseaux qui correspondent aux probabilités conditionnelles associées à chaque nœud de celui-ci (par exemple , $P(d|s_1, s_2)$ dans le cas de la figure 1(b)) étant partagés par tous les documents, la probabilité $P(d|s_1, s_2)$ est la même quel que soit le document considéré. On suppose donc que la structure des documents du corpus impose la nature des dépendances fonctionnelles entre les entités du document. Là aussi cette hypothèse forte est nécessaire pour permettre une estimation robuste des paramètres. Comme pour d'autres modèles stochastiques (e.g. modèles de Markov cachés), il a été observé avec les RB en général ce ne sont pas les modèles les plus réalistes, qui sont souvent trop complexes, qui donnent les meilleurs résultats en pratique. Des hypothèses simplificatrices même drastiques sont nécessaires pour obtenir des performances acceptables sur des données.

Les principales difficultés rencontrées dans l'établissement du modèle sont liées d'une part au fait que les documents ayant en général des structures différentes, les arbres qui les représentent sont d'arité et de profondeur variables, ce qui rend complexe le calcul des probabilités conditionnelles en chaque noeud et le partage de paramètres. D'autre part l'apprentissage à partir de données est assez lourd dans les RB, en particulier dans le cas d'arbres de grande dimension comme ceux que nous manipulons. Pour les les problèmes d'arité et de profondeur variables, nous proposons différentes représentations qui permettent d'affecter des paramètres communs à des arbres possédant des structures différentes. Pour l'apprentissage, nous avons employé une version de l'algorithme EM adaptée aux RB.

4.2.1. Recherche Documentaire

Nous allons construire un réseau bayésien par document et par requête. Pour les documents, la structure du RB est définie comme expliqué en 4.2. Nous allons utiliser comme exemple, l'application que nous avons traitée, où les documents sont les pages d'un site web, et la structure d'arbre est construite à partir de la structure du site (figure 2). Pour chaque page HTML, nous avons deux nœuds parents (la pertinence du texte et celle de la page "mère"). La pertinence du texte dépend de la représentation textuelle proprement dite. Cette représentation textuelle est l'évidence introduite dans le réseau bayésien. La probabilité conditionnelle $P(\text{pertinence du texte}|\text{texte})$ est calculée pour chaque nouvelle requête et est approximée par un modèle classique TF-IDF. Les paramètres du modèle correspondent à la distribution de probabilité

$$P(\text{pertinence d'une page}|\text{pertinence du texte} \wedge \text{pertinence de la page mère}) \qquad (1)$$

Lors de l'apprentissage, l'évidence est introduite au niveau des nœuds "texte" et des nœuds correspondant aux différentes pages jugées pertinentes dans la base d'apprentissage. Les paramètres de la distribution (1) sont alors ré-estimés par le biais de l'algorithme EM (Dempster et al., 1977). Lors de la recherche documentaire, on se sert du réseau bayésien pour faire de l'inférence. La seule évidence est introduite au niveau des nœuds "texte", et la recherche correspond à calculer les différentes probabilités $P(P_i)$ pour les différentes pages P_i.

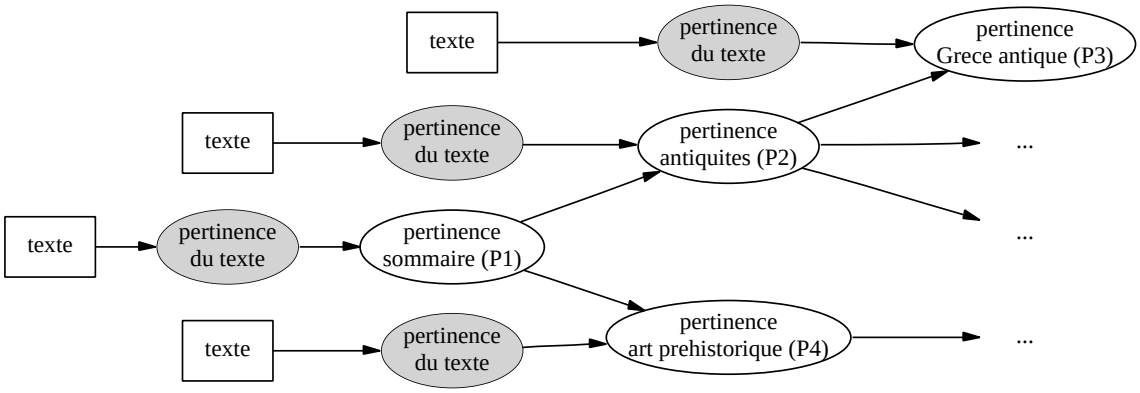


FIG. 2: *Réseau correspondant aux premières pages du site Web du musée de l'hermitage. L'exemple montre quatre pages sur trois niveaux (accueil; antiquités, art préhistorique; Grèce antique). Le site utilisé dans les expériences contient environ 450 pages.*

4.2.2. Classification

Nous allons construire un réseau bayésien par document et par classe. Les paramètres de ces réseaux seront les mêmes pour une même classe. Nous considérerons qu'un problème de classification à n classes est équivalent à n problèmes de classification à deux classes $Relevant, Irrelevant$. Dans la suite, nous considérerons donc que nous sommes dans le cas d'un problème à deux classes. Pour simplifier, nous allons décrire nos modèles pour la tâche de classification que nous avons abordée dans les expériences. Comme pour la RD, il existe aujourd'hui peu de corpus de documents structurés pour la tâche de classification. Nous avons réalisé notre propre corpus. Nous avons fait le choix de classer des sites Web en considérant ceux-ci comme des documents structurés arborescents. Nous allons détailler deux modèles que nous avons développés pour le traitement de sites Web. Ces deux modèles considèrent que la pertinence d'une page d'un site dépend de la pertinence du texte contenu dans cette page (calculée ici par le modèle Naive-Bayes) et de la pertinence des pages auxquelles on peut accéder à partir de celle-ci (voir la figure 3). La pertinence du site sera celle de sa page d'accueil. Les deux modèles sont issus d'hypothèses différentes quant à la pertinence des pages "filles".

P_p est la probabilité de pertinence de la page.
P_t est la probabilité de pertinence du texte de la page (calculée avec le modèle Naive Bayes).
P_{pf} est la probabilité de pertinence de l'ensemble des pages filles.

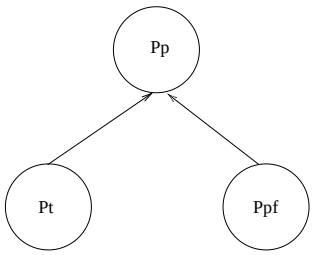


FIG. 3: *Ce schéma présente la dépendance conditionnelle entre la pertinence d'une page, la pertinence du texte de cette page et la pertinence de ses pages filles.*

Différences entre les modèles Les figures 4 et 5 illustrent les différences de ces modèles. Nos documents structurés (sites Web) ne possèdent pas tous la même structure et ne sont donc pas représentés par le même arbre, ce qui ne correpond pas à l'utilisation usuelle des RB où l'on

a un RB pour une tâche de discrimination donnée. Comment alors apprendre les paramètres des différents RB ?, Pour gérer ce problème d'arité et de profondeur variable des arbres, nous devons émettre des hypothèses permettant de généraliser les réseaux bayésiens au cas de réseaux de taille différente.

Le premier modèle est très simple. Il est inspiré du modèle Naive-Bayes. Nous considérons que, pour que l'ensemble des pages filles d'une même page soit non pertinent, il est nécessaire que chacune de ces pages soit non pertinente. D'un point de vue probabiliste, cela se traduit par le fait que la non-pertinence $\bar{P}_p$ des pages filles est le produit de la non-pertinence de chacune de ces pages. Considérons le réseau bayésien de la figure 4.

Nous supposons pour ce premier modèle : $\bar{P_{pf}} = \prod_{i=1}^{N} \bar{P}_i$. Cette fonction correspond à la fonction "OU" probabiliste. Les paramètres calculés par apprentissage sont alors uniquement les probabilités conditionnelles $P_i(p_p|p_t, p_{pf})$ où i indique la profondeur du noeud p_p.

Le second modèle est un peu plus complexe. Il considère que la pertinence des pages filles dépend de la pertinence de la page dont le texte est le plus pertinent et de la pertinence des pages restantes etc. récursivement. La figure 4 illustre ces dépendances. Ce modèle introduit un ordre dans les pages filles de la page considérée. En effet, l'ordre naturel dans le cas de documents plus classiques (1er paragraphe, 2nd paragraphe, ...), n'existe pas dans le cas des sites Web. Nous avons donc introduit un ordre artificiel provenant de la pertinence directe du texte de chacune des pages filles. Cet ordre est cohérent avec le partage des paramètres entre les différents réseaux associés à une classe. Il règle le problème d'arité puisqu'en un noeud, pour tout document la pertinence du noeud dépend de celle de ses 2 fils. Les paramètres calculés par apprentissage sont, d'une part, et comme dans le modèle précédent, les $P_i(p_p|p_t, p_{pf})$, d'autre part, les $P(p_{pm..n}|p_{pm}, p_{pm+1;n})$ (voir figure 5).

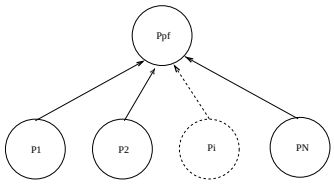
Dans le cadre de nos expériences, et pour avoir un temps d'apprentissage raisonnable, nous avons appris les probabilités $P(p_{pm..n}|p_{pm}, p_{pm+1;n})$ pour m de 1 à 10 et considéré que, pour $m \geq 11$, $P(p_{pm..n}|p_{pm}, p_{pm+1;n}) = P(p_{p10..n}|p_{pm}, p_{pm+1;n})$.

5. Expériences et Résultats

Nous allons présenter les expériences et les résultats obtenus pour les tâches de recherche documentaire et de classification. De façon générale, les corpus structurés sont encore très rares aujourd'hui, car cette problématique est très récente. Cette absence de corpus de référence est bien sûr problématique pour les évaluations. Dans le cas de la recherche documentaire, nous avons pu utiliser un corpus de petite taille déjà constitué, pour la classification, nous avons réalisé notre propre corpus.

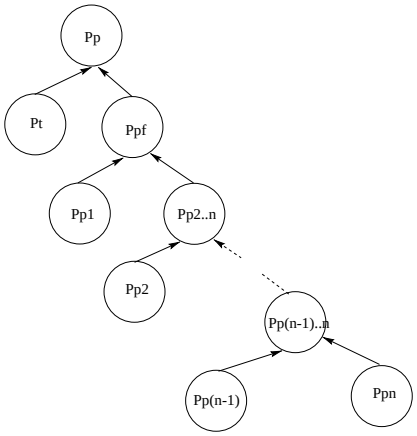
5.1. Recherche documentaire

Le corpus utilisé est constitué de l'ensemble des pages du site du musée de l'hermitage. Ce corpus a fait l'objet d'expériences (Lalmas and Moutogianni, 2000) et nous a été fourni par les auteurs de cet article. Ce corpus contient 445 documents (pages du site) présentant la collection du musée de l'hermitage à Saint-Petersbourg – allant d'un sommaire d'une galerie à la présentation détaillée d'une œuvre. La nature strictement hiérarchique du site permet de le transformer en un réseau bayésien de manière simple. Quinze questions et leurs jugements de pertinence associés sont présents dans ce corpus. La mesure de performance utilisée est la courbe rappel-précision (figure 6). Les courbes présentées sont celles obtenues avec le modèle



P_{pf} est la probabilité de pertinence de l'ensemble des pages filles.
P_i pour i de 1 à N est la probabilité de la page fille i.

FIG. 4: Modèle 1*: Cette figure illustre la dépendance conditionnelle entre la pertinence de l'ensemble des pages filles d'une page donnée et la pertinence de chacune de ces pages*



$P_{pi..j}$ est la probabilité de pertinence de l'ensemble des pages filles de i à j.
P_{pi} pour i de 1 à n est la probabilité de la page fille i.

FIG. 5: Modèle 2*: Cette figure illustre la dépendance conditionnelle entre la pertinence globale de l'ensemble des pages filles (de 1 à n) d'une page donnée p et la pertinence de chacune de ces pages P_{pi}*

classique TF-IDF/cosine, et avec le modèle proposé après 5 itérations EM. La convergence du réseau est très rapide étant donné le faible nombre de questions et de jugements de pertinence disponible, et le nombre de paramètres a été limité à quatre pour éviter le sur-apprentissage.

5.2. Classification

Nous avons utilisé comme mesure de performance le *breakeven point*. Ce point correspond au moment où la précision et le rappel sont égaux. C'est une mesure classique de performance des systèmes de classification. Nous avons comparé notre modèle à un modèle de référence qui est le modèle Naive-Bayes classique. Le corpus présenté ici à été construit par nous-mêmes à partir de l'annuaire de sites Web Yahoo.

Le corpus est constitué de 7006 sites Web répartis en 19 classes différentes. Ces 19 classes sont les sous-classes du thème *Computers and Internet* de l'annuaire international de Yahoo. Chaque site Web est transformé en un arbre de documents HTML correspondant à chacune des pages du site. Seuls les sites de plus de 2 pages et de moins de 2000 pages ont été conservés afin de réduire la taille de la base de données de manière acceptable. Nous n'avons pas effectué de sélection de variables et le vocabulaire contient 358995 mots. Chaque site est ensuite transformé en un document XML contenant l'information de structure (l'organisation des pages) ainsi que l'information textuelle (le texte des pages HTML débarrassé des différents tags et pré-traité par le stemmer de Porter). Le corpus a été découpé en deux corpus de même taille, un pour l'entraînement, l'autre pour le test. Nous présentons ici les résultats concernant les 5 plus grosses classes. Des expériences complétementaires sont en cours.

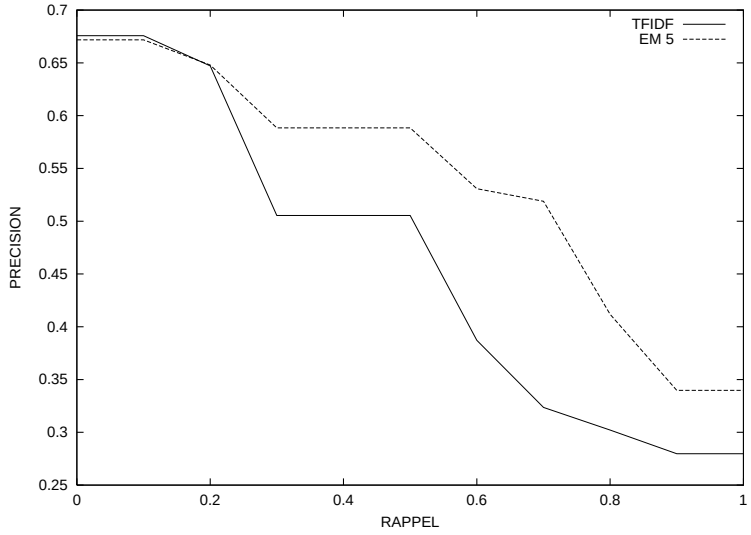


FIG. 6: *Courbe rappel-précision (corpus Hermitage) EM i représente les performances pour la ième itération de l'algorithme EM*

Classes	Nombre de documents	Nombre de pages moyen	Naive Bayes	Modèle 1	Modèle 2
Communications	183	20.45	33.8	35.6	31.6
Internet	849	27.8	64.1	56.4	57.2
Multimedia	114	16.85	19.3	12.5	16.73
Product Reviews	493	36.7	68.3	66.2	75.1
Software	896	30.28	39	37.5	37.6

FIG. 7: *Résultats de classification*

5.3. Commentaires des résultats

Dans les deux cas (RD et classification) on observe des performances similaires à celles obtenues avec des modèles bien plus simples (tf-idf pour la RD et naïve Bayes pour la classification). Pour une preemière étape, nous considérons cela comme encourageant, car le développement des modèles et l'apprentissage pour ces tâches ne sont pas triviaux et les solutions proposées ici sont surement largement perfectible. Pour la RD, nous avons fait figurer sur la figure 2 différentes itérations de l'algorithme EM, au dela d'une dizaine d'itérations, le RB a tendance à surapprendre. Pour la classification, on peut noter que le second modèle, qui est un peu plus complexe donne des résultats légèrement meilleurs que le premier. Les meilleurs résultats obtenus dans la classe "Product Reviews" nous conduisent à examiner l'influence du nombre de page moyen par site sur le comportement de l'algorithme.

6. Observations et limites

Au vu des résultats qui sont très proches de ceux de modèles bien plus simples, et au-delà des perspectives présentées dans la suite, il est évidemment important d'améliorer les performances brutes de nos modèles. D'autres limites nous paraissent importantes à dépasser. La première est liée au choix de ne faire apparaitre l'information textuelle dans le modèle que par le biais

de la pertinence de différents noeuds. Ce choix nous a permis de mettre en place les premières expériences car il simplifie de manière conséquente les modèles présentés. Mais il les limite également car les *a priori* introduits pour représenter la pertinence d'un texte pour une requête ou une catégorie ne sont pas toujours cohérent avec le modèle probabiliste utilisé. D'autre part, bien que les tâches de recherche documentaire et de catégorisation soient proches, il reste une différence notable : en catégorisation, les requêtes (catégories) sont finies et fixées une fois pour toutes alors que l'ensemble des documents est en perpétuelle évolution. En recherche documentaire, c'est l'inverse. La façon de modéliser la tâche, la façon dont l'information de pertinence est transmise aux différents nœuds de la hiérarchie peut donc gagner à être plus spécifique à la tâche envisagée.

7. Perspectives et Conclusions

Notre modèle ne possède pas que des applications pour les tâches classiques de recherche documentaire et de classification de documents structurés. La flexibilité des réseaux bayésiens et la possibilité d'effectuer de nombreuses inférences laisse entrevoir d'autres applications.

À l'heure actuelle, nous introduisons de l'évidence afin de caractériser la pertinence de l'information textuelle du document pour la requête ou la classe donnée en utilisant des modèles classiques de RI. Nous pouvons envisager d'introduire de l'évidence à n'importe quel endroit dans le réseau bayésien. Considérons qu'un utilisateur nous donne son avis quant à la pertinence d'une certaine partie du document; cet avis peut-être interprété en terme d'évidence qui sera introduite "interactivement" dans le réseau. Ce dernier pourra alors recalculer, en fonction de l'avis de l'utilisateur, la pertinence du document et de toutes ses parties. Notre modèle deviendrait ainsi un *modèle interactif de recherche d'information*. Appliqué au cas particulier des sites Web, il pourrait devenir un assistant de navigation permettant à tout instant d'orienter un utilisateur vers les pages les plus susceptible de l'intéresser en tenant compte de son parcours dans les sites Web. Une autre extension intéressante, qui se rapproche de la problématique présente dans la communauté base de données (BD), serait que l'utilisateur puisse poser des *questions structurées* en recherche documentaire (e.g. des requêtes qui établissent des contraintes sur la structure, comme par exemple *Trouver les livres dont le premier chapitre contient les mots x, y et z*). La structure, qu'elle soit interne ou externe, permet aussi d'envisager de nouvelles techniques de présentation et de visualisation des résultats : qu'il s'agisse de regrouper les documents par catégories dans le cas de bases de données hiérarchiques, de synthétiser les informations à la manière des "coupes" dans les hypercubes ou tout simplement de choisir entre présenter un sommaire ou un article particulier directement, la structure ouvre de nouvelles perspectives dans un domaine qui a finalement été très peu exploré par la communauté IR - peut-être une évaluation plus complexe des méthodes est-elle à l'origine de cet état des choses.

Références

Baeza-Yates R. and Ribeiro-Neto B. (1999). *Modern Information Retrieval*. Addison Wesley, New York, USA.

Callan J. P., Croft W. B., and Harding S. M. (1992). The INQUERY Retrieval System. In Tjoa A. M. and Ramos I. editors, *Database and Expert Systems Applications, Proceedings of the International Conference*, pages 78–83, Valencia, Spain. Springer-Verlag.

Chakrabarti S., Dom B., Agrawal R., and Raghavan P. (1997). Using taxonomy, discriminants, and

signatures for navigating in text databases. In *23rd International Conference on Very Large Data Bases*, Athens, Greece.

Dempster A. P., Laird N. M., and Rubin D. B. (1977). Maximum Likelihood from incomplete data via de EM algorithm. *J. Royal Statistical Society Seriers B*, 39:1–37.

Denoyer L., Zaragoza H., and Gallinari P. (2001). Hmm-based passage models for document classification and ranking. In *23rd European Colloquium on Information Retrieval Research*.

Indrawan M., Ghazfan D., and Srinivasan B. (1994). Using Bayesian Networks as Retrieval Engines. In *ACIS 5th Australasian Conference on Information Systems*, pages 259–271, Melbourne, Australia.

Jensen F. V. (1996). *An introduction to Bayesian Networks*. UCL Press, London, England.

Koller D. and Sahami M. (1997). Hierarchically Classifying Documents Using Very Few Words. In *ICML-97: Proceedings of the Fourteenth International Conference on Machine Learning*, pages 435–443, San Francisco, CA, USA. Morgan Kaufmann.

Krause P. (1998). Learning probabilistic networks.

Lalmas M. (1997). Dempster-Shafer's Theory of Evidence Applied to Structured Documents: Modelling Uncertainty. In *Proceedings of the 20th Annual International ACM SIGIR*, pages 110–118, Philadelphia, PA, USA. ACM.

Lalmas M. and Moutogianni E. (2000). A Dempster-Shafer indexing for the focussed retrieval of a hierarchically structured document space: Implementation and experiments on a web museum co. In *6th RIAO Conference, Content-Based Multimedia Information Access*, Paris, France.

Lalmas M., Ruthven I., and Theophylactou M. (1997). Structured document retrieval using Dempster-Shafer's Theory of Evidence: Implementation and evaluation. Technical report, University of Glasgow, UK.

McCallum A., Rosenfeld R., Mitchell T., and Ng A. Y. (1998). Improving Text Classification by Shrinkage in a Hierarchy of Classes. In Brasko I. and Dzeroski S. editors, *International Conference on Machine Learning (ICML'98)*, pages 359–367. Morgan Kaufmann.

Murphy K. P. (2000). A Brief Introduction to Graphical Models and Bayesian Networks. web: http://www.cs.berkeley.edu/˜murphyk/Bayes/bayes.html.

Myaeng S. H., Jang D.-H., Kim M.-S., and Zhoo Z.-C. (1998). A Flexible Model for Retrieval of SGML documents. In Croft W. B., Moffat A., van Rijsbergen C., Wilkinson R., and Zobel J. editors, *Proc 21st ACM SIGIR*, pages 138–140, Melbourne, Australia. ACM Press, New York.

OASIS (2001). Docbook standard. `http://www.oasis-open.org/specs/docbook.shtml`.

Pearl J. (1988). *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufmann.

Ribeiro B. A. N. and Muntz R. (1996). A Belief Network Model for IR. In *Proc 19th ACM–SIGIR*, pages 253–260.

Turtle H. R. and Croft W. B. (1991). Evaluation of an Inference Network-Based Retrieval Model. *ACM Transactions On Information Systems*, 9(3):187–222.

Wilkinson R. (1994). Effective retrieval of structured documents. In Croft W. and van Rijsbergen C. editors, *Proc 17th ACM SIGIR*, pages 311–317, Dublin, Ireland. Springer-Verlag.

World Wide Web Consortium (1998). XHTML standard. `http://w3c.org`.