Intégration de la structure dans un modèle probabiliste de document

Mathias Géry, Christine Largeron et Franck Thollard

Université Jean Monnet,
Laboratoire Hubert Curien, UMR CNRS 5516, St-Etienne
prenom.nom@univ-st-etienne.fr

Résumé. En fouille de textes comme en recherche d'information, différents modèles, de type probabiliste, vectoriel ou booléen, se sont révélés bien adaptés pour représenter des documents textuels mais, ces modèles présentent l'inconvénient de ne pas tenir compte de la structure du document. Or la plupart des informations disponibles aujourd'hui sur Internet ou dans des bases documentaires sont fortement structurées. Dans cet article[1], nous proposons d'étendre le modèle probabiliste de représentation des documents de façon à tenir compte du poids d'une certaine catégorie d'éléments structurels : les balises représentant la structure logique et la structure de mise en forme. Ce modèle a été évalué à l'aide de la collection de la campagne d'évaluation INEX 2006.

1 Introduction

En fouille de texte comme en recherche d'information (RI), plusieurs modèles sont utilisés pour représenter un document. Ces modèles, de type probabiliste, booléen ou vectoriel, se sont révélés bien adaptés pour représenter des documents textuels. Cependant, ils présentent l'inconvénient de ne pas tenir compte de la structure du document. Or, la plupart des informations disponibles aujourd'hui sur Internet ou dans des bases documentaires sont fortement structurées. C'est la raison pour laquelle des travaux récents, en RI comme en fouille de données se sont intéressés à la structure des documents. Ceci a notamment conduit à l'émergence de la recherche d'information XML orientée contenu dont l'objectif est justement d'exploiter l'information structurelle contenue dans les documents pour concevoir des systèmes de RI plus efficaces. La compétition INEX[2] (INitiative for Evaluation of XML Retrieval) produit d'ailleurs depuis 2002 de larges collections de documents utilisables pour l'évaluation de tels systèmes. L'exploitation de la structure a aussi été étudiée dans des tâches de classement, supervisé ou non, de documents . Dans ce contexte, plusieurs voies ont été envisagées, parmi lesquelles on citera l'extension des modèles usuels de représentation de documents textuels [Doucet et Ahonen-Myka (2002)] ou l'exploitation de la structure arborescente des documents XML [Yi et Sundaresan (2000); Marteau et al. (2005); Vercoustre et al. (2006)]. Enfin, dans le contexte de la détection d'information nouvelle (Novelty Detection), d'autres travaux ont

[1]Ce travail a été partiellement soutenu par l'action collaborative Web Intelligence de la région Rhône-Alpes
[2]http ://inex.is.informatik.uni-duisburg.de/2007/

pris en compte la structure logique des documents en estimant le poids à accorder chacune des parties qui le composent [Jacquenet et Largeron (2006)].

Dans cet article, nous proposons d'étendre le modèle probabiliste de façon à tenir compte du rôle joué par les éléments de structure et de mise en forme pour mettre en évidence des informations importantes. Notre approche nécessite une phase d'apprentissage sur une partie de la collection considérée. Au cours de cet apprentissage, un poids est calculé pour chacune des balises, basé sur la probabilité pour que cette balise distingue les termes pertinents. Dans une seconde phase, le modèle que nous avons développé permet d'estimer, en tenant compte de ce poids, la probabilité qu'un document de la collection soit pertinent pour une requête donnée. Ce modèle est décrit dans la prochaine section tandis que les résultats d'expérimentations obtenus sur la collection INEX 2006 sont présentés dans la troisième section.

2 Un modèle probabiliste de représentation de documents structurés

2.1 Principe d'intégration de la structure dans un modèle probabiliste de documents

En Recherche d'Information, le modèle probabiliste de documents [Robertson et Jones (1976)] aspire à estimer la pertinence d'un document pour une requête à partir de deux probabilités : celle de trouver une information pertinente et celle de trouver une information non pertinente. Ces estimations sont basées sur la probabilité de chacun des termes contenus dans le document d'apparaître dans un document pertinent ou dans un document non pertinent de la collection. Pour ce faire, on utilise une collection de test, composée de documents, de requêtes et de la connaissance des documents pertinents pour chaque requête. Cette collection permet, dans une phase d'apprentissage, d'estimer la probabilité de pertinence de chaque terme en fonction de ses distributions respectivement dans les documents pertinents et les documents non pertinents.

Notre objectif est d'intégrer la structure des documents dans ce modèle afin de parvenir à une recherche d'information structurée. Dans notre modèle, seront considérés des éléments de structure logiques (titre, section, paragraphe, etc.) et de mise en forme (souligné, en gras, centré, etc.). L'intégration de la structure dans le modèle probabiliste s'effectue ensuite à deux niveaux. Dans le premier, la structure logique est utilisée pour identifier les éléments XML qui seront susceptibles d'être indexés par notre système : les sections, paragraphes, tableaux, etc. Dans le second, les balises de structure logique et de mise en forme sont intégrées au modèle probabiliste classique. Cette intégration nécessite une étape préliminaire qui consiste à estimer un poids pour chacune des balises. Ce poids est basé sur la probabilité pour qu'une balise distingue les termes pertinents. Dans la seconde étape d'intégration des balises, le modèle que nous avons développé permet de déterminer la probabilité qu'un document de la collection soit pertinent pour une requête donnée en tenant compte non seulement de la pondération classique des termes du modèle probabiliste, mais aussi de la pondération de chacune des balises qui englobent ces termes. La section suivante présente plus formellement ce modèle probabiliste de représentation de documents structurés.

2.2 Notations

On dispose d'un ensemble $\mathcal{D}$ de documents structurés. En pratique, il s'agira le plus souvent de documents XML. Chaque élément logique (i.e. section, paragraphe,...) e_j d'un document XML représente donc un ensemble de termes, délimité par une balise structurelle logique, qui sera utilisée pour indexer l'élément.

On note :

- $E = \{e_j, j = 1, ..., l\}$, l'ensemble des éléments structurés considérés dans la collection, par exemple des sections, des paragraphes, etc.
- $T = (t_1, ..., t_i, ...t_n)$, un index de termes construit sur E.
- $B = \{b_1, ..., b_k, ..., b_m\}$, l'ensemble des balises logiques et de mise en forme considérées.

Soit E_j, un vecteur de variables aléatoires T_{ij} à valeur dans $\{0, 1\}$:
$E_j = (T_{10}, ..., T_{1k}, ..., T_{1m}, ..., T_{i0}, ..., T_{ik}, .., T_{im}, ..., T_{n0}, ..., T_{nk}, .., T_{nm})$

$$\text{avec} \begin{cases} T_{ik} = 1 & \text{si le terme } t_i \text{ apparaît étiqueté par la balise } b_k \\ T_{ik} = 0 & \text{sinon} \\ T_{i0} = 1 & \text{si le terme } t_i \text{ apparaît sans être étiqueté} \\ & \quad \text{par une des balises de mise en évidence de } B \\ T_{i0} = 0 & \text{si le terme } t_i \text{ n'apparaît pas sans étiquette} \end{cases}$$

On notera $e_j = (t_{10}, ..., t_{1k}, ..., t_{1m}, t_{i0}, ..., t_{ik}, .., t_{im}, t_{n0}, ..., t_{nk}, .., t_{nm})$ une réalisation de la variable aléatoire E_j. À partir de cette représentation, l'objectif est maintenant d'étendre le modèle probabiliste pour prendre en compte la structure de mise en forme des documents.

2.3 Probabilité de pertinence d'un élément XML, basée sur les balises

La fonction de pondération BM25, introduite par [Robertson et Jones (1976)] auquel nous renvoyons le lecteur pour plus de précision, est très largement utilisée dans les systèmes de recherche d'information probabilistes pour estimer le poids d'un terme t_i dans un élément XML e_j. Dans notre modèle, cette première pondération, notée w_{ij}, est enrichie de manière à prendre en compte la structure logique et de mise en forme des documents. Dans un contexte de recherche d'information, on désire en effet estimer la pertinence d'un élément XML e_j relativement à une requête. Ce qui revient à estimer $P(R|e_j)$ (respectivement $P(NR|e_j)$), la probabilité de trouver une information pertinente (respectivement non pertinente) quand on observe l'élément e_j pour une requête donnée.

On introduit une fonction de classement $fc_1(e_j)$ qui permettra, en comparant ces deux probabilités, d'ordonner les documents en fonction de leur pertinence par rapport à la requête :

$$fc_1(e_j) = \frac{P(R|e_j)}{P(NR|e_j)}$$

Plus $fc1(e_j)$ est élevée, plus pertinentes sont les informations contenues dans e_j. Par la formule de Bayes et en éliminant le terme $\frac{P(R)}{P(NR)}$ constant sur la collection pour une requête donnée qui n'interviendra donc pas dans le classement des documents, on obtient fc_2, proportionnelle à fc_1 :

$$fc_2(e_j) = \frac{P(e_j|R)}{P(e_j|NR)}$$

Moyennant l'hypothèse d'indépendance des termes (Binary Independance Model) :

$$P(E_j = e_j|R) = \prod_{t_{ik} \in e_j} P(T_{ik} = 1|R)^{t_{ik}} \times P(T_{ik} = 0|R)^{1-t_{ik}} \quad (1)$$

$$P(E_j = e_j|NR) = \prod_{t_{ik} \in e_j} (P(T_{ik} = 1|NR))^{t_{ik}} \times (P(T_{ik} = 0|NR))^{1-t_{ik}} \quad (2)$$

Pour simplifier les notations, on pose :

$p_0 = P(T_{i0} = 0|R)$ **:** probabilité de ne pas avoir t_i sachant que l'élément est pertinent

$p_{ik} = P(T_{ik} = 1|R)$ **:** probabilité d'avoir t_i étiqueté par b_k sachant que l'élément est pertinent

$q_0 = P(T_{i0} = 0|NR)$ **:** probabilité de ne pas avoir t_i sachant que l'élément est non pertinent

$q_{ik} = P(T_{ik} = 1|NR)$ **:** probabilité d'avoir t_i étiqueté par b_k sachant que l'élément est non pertinent

En reportant dans la fonction de classement $fc_2(e_j)$:

$$fc_2(e_j) = \frac{\prod_{t_{ik} \in e_j} (p_{ik})^{t_{ik}} \times (1 - p_{ik})^{1-t_{ik}}}{\prod_{t_{ik} \in e_j} (q_{ik})^{t_{ik}} \times (1 - q_{ik})^{1-t_{ik}}}$$

La fonction logarithmique étant croissante, en prenant le logarithme de $fc_2(e_j)$ le classement produit par la fonction $fc_3(e_j)$ = $\log(fc_2)$ sera le même que celui produit par fc_2 :

$$fc_3(e_j) = \sum_{t_{ik} \in e_j} t_{ik} \times \left(\log \left(\frac{p_{ik}}{1 - p_{ik}} \right) - \log(\frac{q_{ik}}{1 - q_{ik}}) \right) + \sum_{t_{ik} \in e_j} log(\frac{1 - p_{ik}}{1 - q_{ik}}) \quad (3)$$

Le terme $\sum_{t_{ik} \in e_j} log(\frac{1-p_{ik}}{1-q_{ik}})$ est une constante relativement à la collection (*i.e. indépendant de* t_{ik}), ne pas le considérer ne change pas le classement produit par la fonction. On en déduit la fonction de pertinence, tirée de fc_3 :

$$fc_{balises}(e_j) = \sum_{t_{ik} \in e_j} t_{ik} log \left(\frac{p_{ik}(1 - q_{ik})}{q_{ik}(1 - p_{ik})} \right) \quad (4)$$

Le poids d'un terme t_i étiqueté par la balise b_k est noté w'_{ik} : $w'_{ik} = log(\frac{p_{ik}(1-q_{ik})}{q_{ik}(1-p_{ik})})$

Ainsi, dans ce modèle probabiliste tenant compte de la structure du document, la pertinence d'un élément e_j par rapport aux balises est mesurée par le score $fc_{balises}(e_j)$:
$fc_{balises}(e_j) = \sum_{t_{ik} \in e_j} t_{ik} \times w'_{ik}$

En pratique, pour mesurer cette pertinence, il convient d'estimer les probabilités p_{ik} et q_{ik}, $i \in \{1, .., n\}$ et $k \in \{0, .., m\}$ à partir d'un échantillon d'apprentissage EA constitué d'éléments déjà jugés pour une requête. À partir des ensembles R et NR contenant respectivement les éléments pertinents et non pertinents, on obtient :

– r_{ik} : nombre de termes t_i étiquetés par b_k parmi les éléments pertinents de EA ;

- n_{ik} : nombre de termes t_i étiquetés par b_k parmi les éléments de EA ;
- $r'_{ik} = n_{ik} - r_{ik}$: nombre de termes t_i étiquetés par b_k parmi les éléments non pertinents de EA ;
- $R = \sum_{ik} r_{ik}$: somme des occurrences des termes figurant parmi les éléments pertinents de EA ;
- $N - R = \sum_{ik} r'_{ik}$: somme des occurrences des termes figurant parmi les éléments non pertinents de EA.

On en déduit $p_{ik} = P(t_{ik} = 1|R) = \frac{r_{ik}}{R}$ et $q_{ik} = P(t_{ik} = 1|NR) = \frac{n_{ik} - r_{ik}}{N-R}$

Ayant construit des estimateurs sans biais de pik et de q_{ik}, les probabilités d'avoir le terme t_i étiqueté par b_k sachant que l'élément est respectivement pertinent et non pertinent, on en déduit $p.k$, la probabilité d'avoir la balise b_k sachant que l'élément est pertinent et $q.k$, la probabilité d'avoir la balise b_k sachant que l'élément n'est pas pertinent : $p.k = \sum_i p_{ik}$ et $q.k = \sum_i q_{ik}$.

2.4 Combinaison des pertinences basées sur les termes et sur les balises

Pour obtenir un score global de classement $fc(e_j)$ d'un élément e_j, en fonction des termes et des balises, qui permette d'estimer sa pertinence par rapport à une requête, nous avons proposé une une première approche qui consiste à multiplier le poids w_{ij} de chaque terme de e_j par la moyenne des poids w'_{ik} correspondant à toutes les balises qui englobent le terme. Ainsi, nous calculons $fc(e_j) = \sum_{t_i \in e_j} w_{ij} * \prod_{k/t_{ik}=1} w'_{ik}$

3 Expérimentation sur la collection INEX

3.1 Présentation de la collection

Nous avons évalué notre modèle sur la collection d'INEX 2006 (Initiative for Evaluation of XML Retrieval) composée de 659.388 articles en anglais issus de l'encyclopédie Wikipedia. Le modèle vectoriel basé sur la fonction de pondération BM25, a été utilisé comme modèle de référence et comparé au modèle probabiliste structurel décrit précédemment, qui utilise lui aussi la fonction de pondération BM25, mais en intégrant les balises. Les résultats ont été évalués en utilisant les taux de *précision* et de *rappel* ainsi que la mesure de performance globale *interpolated mean average precision* (iMAP) permettant de les combiner et définie dans [Pehcevski et al. (2007)]

Sur les 114 requêtes de la collection, l'indice iMAP est égal à 2,34% dans le cas du modèle de référence sans utilisation des balises. Il est égal à 1,80% quand toutes les balises sont considérées. Cette tendance est confirmée lorsqu'on considère le rappel et la précision indépendamment l'un de l'autre [Géry et al. (2007)]. Ces résultats préliminaires peu convaincants, ne remettent pas nécessairement en cause l'intérêt d'exploiter l'information structurelle en plus de l'information textuelle mais plutôt les modalités de combinaison des poids des termes avec ceux des éléments structurels.

4 Conclusion

Dans cet article, nous avons proposé d'étendre le modèle probabiliste de représentation des documents structurés de façon à tenir compte du poids des balises représentant la structure logique et la structure de mise en forme. Ces poids sont estimés par apprentissage puis intégrés dans le calcul de la probabilité qu'un document de la collection soit pertinent pour une requête donnée. Bien que les résultats préliminaires obtenus sur la collection test de la campagne d'évaluation INEX 2006 soient peu convaincants, nous pensons qu'ils ne remettent pas en cause l'intérêt d'exploiter l'information structurelle en plus de l'information textuelle, mais que la combinaison des poids des termes avec ceux des balises doit être étudiée de manière plus approfondie.

Références

Doucet, A. et H. Ahonen-Myka (2002). Naive clustering of a large xml document collection. In *Proceedings of the First Workshop of the Initiative for the Evaluation of XML Retrieval (INEX)*, Schloss Dagsuhl, Germany, pp. 81–87.

Géry, M., C. Largeron, et F. Thollard (2007). Probabilistic document model integrating xml structure. In *Workshop of the Initiative for the Evaluation of XML Retrieval (INEX)*.

Jacquenet, F. et C. Largeron (2006). Using the structure of documents to improve the discovery of unexpected information. In *SAC*, pp. 1036–1042.

Marteau, P., G. Ménier, et L. Ekamby (2005). Apport de la prise en compte du contexte structurel dans les modèles bayésiens de classification de documents semi-structurés. In *RNTI, numéro spécial sur la fouille de données complexes*.

Pehcevski, J., J. Kamps, G. Kazai, M. Lalmas, P. Ogilvie, B. Piwowarski, , et S. Robertson (2007). Inex 2007 evaluation measures. In *INEX 2007 Pre-Proceedings*.

Robertson, S. et K. S. Jones (1976). Relevance weighting of search terms. *Journal of the American Society for Information Sciences 27*(3), 129–146.

Vercoustre, A., M. Fegas, Y. Lechevallier, et T. Despeyroux (2006). Classification de documents xml a partir d'une representation lineaire des arbres de ces documents. In *(EGC 2006), RNTI-E-6*, pp. 433–444.

Yi, J. et N. Sundaresan (2000). A classifier for semi-structured documents. In *Proceedings of the sixth ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 340–344.

Summary

In text mining as in information retrieval (IR), differents models, probabilistic, boolean or vectorial, are well suited to manage textuals documents, but they do not use the document structure. Nevertheless, most of the documents available (e.g. on the internet or in textual databases) are strongly strucured. In this article, we propose an extension of the probabilistic model in order to take into account some of the structural elements present in the document. This model has been evaluated using the INEX 2006 evaluation campaign.