TALN 2004, Fès, 19–21 avril 2004

Une mesure de pertinence pour le tri de l'information dans un index de "fin de livre"

Touria Ait El Mekki et Adeline Nazarenko
Laboratoire d'Informatique de Paris-Nord
99 Avenue J.-B. Clément
93430 Villetaneuse FRANCE
taem,nazarenko@lipn.univ-paris13.fr

Résumé - Abstract

Nous nous intéressons à la construction des index de fin de livres. Nous avons développé le système IndDoc qui aide la construction de tels index. L'un des enjeux de la construction d'index est la sélection des informations : sélection des entrées les plus pertinentes et des renvois au texte les plus intéressants. Cette sélection est évidemment utile pour le lecteur qui doit trouver suffisamment d'information mais sans en être submergé. Elle est également précieuse pour l'auteur de l'index qui doit valider et corriger une ébauche d'index produite automatiquement par IndDoc. Nous montrons comment cette sélection de l'information est réalisée par IndDoc. Nous proposons une mesure qui permet de trier les entrées par ordre de pertinence décroissante et une méthode pour calculer les renvois au texte à associer à chaque entrée de l'index.

This paper deals with the construction of end-of-book indexes. We have developed the IndDoc system which assists the construction of such indexes. One of the stakes of the construction of an index is the information selection: selection of the most relevant entries and the most interesting textual fragments. This selection is obviously useful for the reader who is looking for information. It is also invaluable for the index author who has to validate and correct an outline of index produced automatically by IndDoc. We show how this information selection is carried out by IndDoc. We put forward a measure which sorts the entries in decreasing relevance order and a method to calculate the references to text for each entry.

Mots-clefs – Keywords

Segmentation thématique de texte, extraction d'information, indexation automatique
Text segmentation, information extraction, automatic indexing

Touria Ait El Mekki, Adeline Nazarenko

1 Introduction

Les index de documents sont des outils de navigation traditionnels. Considérant que la présence d'index est souhaitable dans beaucoup de documents mais que le coût de construction des index est parfois rédhibitoire, nous proposons une méthode qui permet de dégrossir le travail d'indexation tout en assurant sa cohérence et en tirant parti de la dimension électronique du document. IndDoc est un système d'aide à la création d'un index de document qui s'appuie sur le texte du document pour proposer une ébauche d'index construite automatiquement à l'indexeur qui peut alors la valider et/ou la compléter. Dans cette ébauche d'index l'information est triée. Cela améliore l'accès à l'information pour le lecteur mais cela facilite aussi le travail de l'indexeur : il peut focaliser son travail de validation sur les éléments les plus pertinents et construire des index de différentes tailles en faisant varier le seuil de pertinence.

Nous mettons ici l'accent sur ces problèmes de sélection et de tri de l'information pertinente et nous montrons comment ils sont pris en compte dans IndDoc. La section 2 présente les index de fin de livre. La section 3 donne un aperçu de IndDoc en situant notre travail par rapport aux travaux existants. Les sections 3 et 4 montrent concrètement comment le calcul des renvois et le tri sont réalisés dans IndDoc.

2 L'index comme outil de navigation dans le document

Face au flux toujours croissant de la documentation imprimée ou numérique, il est important de se doter d'outils d'accès au contenu des documents : une information qui ne peut être localisée dans un flux documentaire "n'existe pas". L'index est l'un des dispositifs les plus anciens permettent d'accéder au contenu des documents.

Nous désignons par *index*, la table alphabétique placée à la fin d'un document ou d'un livre, qui présente les sujets traités dans le document sous forme d'une liste de descripteurs accompagnés de renvois, c'est-à-dire de références aux segments de texte où se trouvent mentionnés ces descripteurs dans le document (sous la forme vedette ou sous une forme variante). L'index vise à guider et éclairer le lecteur. Les lecteurs consultent l'index dans un but précis. S'ils localisent rapidement l'information qu'il recherchent dans le document, ils sont satisfaits. Sinon ils abandonnent l'index.

Entre l'indexation des documents et celle d'un index de fin de livre, le niveau de granularité de l'indexation du texte varie mais l'objectif est identique (Maniez, 2002) : aider un demandeur d'information à trouver rapidement un texte ou un extrait pertinent. L'index est une représentation synthétique de document, il offre un accès rapide et direct à l'information. Dans les deux cas la méthode consiste à repérer, au-delà des mots, les sujets susceptibles de retenir l'attention d'un utilisateur. Dans le premier cas, l'indexeur effectue une analyse réductrice pour ne dégager que les thèmes essentiels du document. Le deuxième cas exige une analyse détaillée pour retenir les concepts portés par chaque page et l'index final peut contenir quelques centaines d'entrées.

Un index se compose de deux parties : la nomenclature et l'ensemble des renvois.

La *nomenclature* de l'index, qui représente les entrées de l'index, est une liste structurée de descripteurs. Les descripteurs sont des termes simples ou complexes, des noms propres, des symboles, des commandes, des sigles, etc. La présentation et l'organisation de cette liste varient

d'un index à l'autre et même au sein d'un domaine ou d'un genre de documents particulier. Elle dépend essentiellement du choix de l'auteur. Cette liste est structurée dans la mesure où les entrées sont reliées les unes aux autres par des relations syntaxiques, sémantiques, ou conceptuelles. Cette structure est d'abord hiérarchique (un terme générique qui constitue une entrée peut être associé à des termes plus spécifiques : les sous-entrées). À côté de ces relations hiérarchiques, on trouve aussi des liens de synonymie et des relations d'association (liens de type *Voir aussi*). La nomenclature offre au lecteur une vue sur les sujets traités (terminologie utilisée, entités nommées...) : une vue globale sur l'ontologie du domaine et une vue locale sur des descripteurs particuliers.

La *liste des renvois* mélange souvent renvois au texte et références croisées (renvois à une autre entrée ou sous-entrée). Comme les références croisées établissent des liens entre descripteurs (synonymie ou lien d'association), nous considérons qu'elles relèvent de la nomenclature qu'elles contribuent à structurer. Dans la suite, nous réservons le terme de "renvoi" pour désigner le renvoi au texte qui établit un lien entre une entrée de la nomenclature et un segment de document qui peut être une page, un groupe de pages, une section ou une sous-section. Pour une entrée donnée, l'index ne renvoie pas à toutes les occurrences du descripteur dans le document. Seules les occurrences les plus pertinentes sont sélectionnées mais les critères de choix varient d'un index à l'autre. L'ordre dans lequel sont présentées les différentes occurrences n'est pas un ordre de pertinence : elles sont classées par ordre d'apparition dans le texte. Les renvois au texte donnent une idée de l'importance des sujets dans le document : la liste des renvois d'un descripteur reflète à la fois la fréquence du descripteur et sa répartition dans le document.

3 Création automatique d'une ébauche d'index

Hormis quelques expériences relativement ponctuelles (Gros & Assadi, 1997; Aussenac & Condamines, 1998; Bourigault & Charlet, 1999) la construction automatique des index de fin de livre n'a guère été étudiée. On peut cependant s'appuyer sur des travaux existants dans différents domaines. Nous présentons la construction de l'index comme un processus en quatre étapes et nous montrons que l'on peut, à chaque étape, tirer profit de travaux antérieurs :

- *Elaborer la nomenclature* d'un index suppose de repérer les descripteurs ou termes pertinents du document et de structurer la liste obtenue en ajoutant des liens sémantiques entre descripteurs. Il s'agit de deux problèmes connus en terminologie computationnelle : l'extraction de termes et la structuration d'une liste de termes.

- *Calculer les renvois textuels* pour chaque descripteur suppose d'identifier les occurrences du descripteur et de déterminer à chaque fois la taille de l'empan de texte (le segment de texte) auquel il est pertinent de renvoyer. Il s'agit d'une tâche de segmentation.

- Une fois établies la nomenclature de l'index et la liste des renvois, il reste à les *trier par ordre de pertinence*. Ce tri sert pour la présentation des résultats. Dans certains cas, en effet, l'ordre alphabétique de présentation des entrées de l'index n'est pas le plus judicieux et le tri des renvois par numéros de page n'est pas très informatif. Ce tri sert aussi à sélectionner l'information la plus utile, lorsque l'indexeur choisit de ne retenir qu'un sous-ensemble des informations contenues dans l'index pour produire un index plus petit (pour le passage d'un index électronique à un index papier, par exemple). Cette

question de la pertinence a été étudiée à la fois pour la recherche d'information et pour le résumé automatique.

- La construction d'index ne pouvant être considérée comme un processus entièrement automatique, l'expertise humaine reste nécessaire pour *valider les résultats* obtenus. InDDoc comporte donc également une interface qui permet d'organiser le travail et d'en assurer la cohérence.

L'extraction des termes a fait l'objet de nombreuses recherches : aujourd'hui, plusieurs extracteurs (voir (Jacquemin & Bourigault, 2003) pour une synthèse) permettent de repérer dans un corpus les termes du domaine. Ces extracteurs fournissent une liste volumineuse de candidats termes qui nécessite un filtrage et un travail de validation. Nous utilisons pour notre part Lexter ou Syntex (Bourigault & Fabre, 2000). Nous structurons ensuite cette liste pour établir des liens entre les entrées et les sous-entrées ainsi que pour établir des références croisées entre entrées. La structuration consiste donc à transformer une liste de termes en un réseau où les termes sont reliés entre eux par des liens syntaxiques et sémantiques (Nazarenko & Hamon, 2002). Les outils de structuration existants reposent sur l'analyse de la structure interne des termes ou/et sur les contextes d'emploi pour rapprocher deux termes. Le module de structuration de INdDoc s'inspire des outils existants. Nous ne le présentons pas en détail ici[1].

Les méthodes de segmentation s'appuient sur la structure physique des documents (phrases, paragraphes, titres, marques typographiques...) et/ou sur leur contenu (passages thématiquement homogènes). Deux approches sont utilisées : la première repose sur la cohésion lexicale (Ferret *et al.*, 1998) (Kozima, 1993) et opère des regroupements de paragraphes (ou au contraire des découpages) sur la base du vocabulaire employé. La deuxième cherche à repérer des connecteurs marquant des relations d'enchaînement entre paragraphes ou au contraire des solutions de continuité. Nous n'exploitons ces méthodes que ponctuellement dans IndDoc car elles s'avèrent difficiles à utiliser pour le calcul des renvois. L'approche par cohésion lexicale est généralement utilisée pour découper des documents qui, contrairement aux nôtres, sont hétérogènes et lexicalement disparates. Concernant l'approche à base de marqueurs, nous avons identifié un certain nombre de marqueurs assez génériques (Porhiel, 1998) : des marqueurs de liaison (*par exemple, Pour ce faire...* en début de paragraphe), des reprises anaphoriques, des marqueurs d'intégration linéaire (*De plus, D'autre part, Ensuite...*). L'analyse des documents montre cependant qu'on trouve à la fois des unités documentaires adjacentes et liées entre elles sans que ce lien soit marqué formellement et des marqueurs de continuité qui ne jouent pas leur rôle dans certains emplois. Le principal défaut de ces approches pour le calcul des renvois est qu'elles proposent une segmentation indépendante du point de vue, alors que dans un index, la segmentation est en partie relative à l'entrée de l'index que l'on cherche à décrire. Un ensemble de paragraphes peut être retenu comme Unité Documentaire (UD) cohérente pour une entrée donnée alors qu'un autre découpage peut s'avérer plus pertinent pour une autre entrée. Les travaux qui proposent une segmentation des documents à la volée en fonction de la requête de l'utilisateur se rapprochent donc davantage de notre objectif (Bellot, 2000).

De nombreuses mesures de pertinence de l'information ont été proposées. En Recherche d'Information (RI), il s'agit d'associer un ensemble de documents et une requête contenant un terme ou plus. Dans notre application, on cherche à apparier un terme à un ensemble de segments du document. Les techniques classiquement utilisées en RI sont de trois types : *les modèles booléens* (Apte *et al.*, 1994) caractérisent simplement la présence ou l'absence des termes dans

[1](AitElMekki & Nazarenko, 2003) présente ce module dans son ensemble.

le document. *Les modèles probabilistes* capturent la distribution des mots dans les documents et s'en servent pour calculer la probabilité de pertinence par rapport à une requête (Bookstein & Swanson, 1974; Crestani *et al.*, 1998). Cette approche est difficile à appliquer dans notre cas parce qu'elle estime des probabilités ou des densités dans des espaces de grande dimension et parce que la plupart des modèles supposent une indépendance complète des termes du document. *Les modèles vectoriels* sont très classiques. L'indexation décrit les documents et les requêtes comme des vecteurs dans un même espace vectoriel. On extrait les termes de la requête et du document, puis on attribue à chaque terme un poids qui reflète son importance. Les représentations les plus courantes reposent sur le codage tf/idf de (Salton, 1989). On mesure la similarité entre chaque document d de la base de documents et la requête q puis les documents sont triés par ordre décroissant de similarité avec la requête.

Des mesures de pertinence ont également été proposées pour sélectionner les phrases pertinentes dans les travaux portant sur le résumé automatique. Dans ce contexte, pour identifier les phrases importantes du texte source qui vont être assemblée en résumé, on exploite des indices statistiques et/ou linguistiques (Paice, 1981; Saggion & Lapalme, 2002) : cooccurrence de mots et lexique des titres, par exemple.

4 Identification des segments de renvoi dans IndDoc

Pour identifier les segments de renvoi, nous partons d'une segmentation brute du document en UD minimales (UDM)(Hearst, 1994; Ouerfelli & Lallich-Boidin, 2000). Une UDM correspond à un paragraphe : cette segmentation fine se justifie dans la mesure où IndDoc, au delà des index papier, vise à construire des index sur support électronique. Il faut ensuite repérer, pour un descripteur donné d_i, quelles UDM contiennent une occurrence de d_i (ou d'une de ses variantes) et identifier la taille du segment de renvoi, c'est-à-dire la taille de l'empan de texte auquel il est pertinent de renvoyer.

La taille des unités documentaires (le niveau de granularité de la segmentation[2]) dépend de la tâche à réaliser (indexation, résumé automatique etc.). Comme nous l'avons mentionné ci-dessus, dans un index, la segmentation est de surcroît relative à l'entrée de l'index que l'on cherche à décrire.

Nous distinguons dans la suite les *unités documentaires* qui résultent d'un découpage absolu du document des *segments de renvoi* qui sont relatifs à un descripteur donné.

Notre méthode de segmentation repose sur : la structure physique des textes (phrases, paragraphes, titres, marques typographiques, etc.), ce qui donne de bons résultats pour les textes scientifiques et techniques qui sont généralement structurés par l'auteur et organisés en sections et sous-sections, et les marques d'intégration linguistique et typographique.

Le segment de texte auquel l'index renvoie est donc vu ici comme une unité discursive qui présente des critères de cohésion sémantique, syntaxique et typographique (Ferret *et al.*, 1998) (Kozima, 1993), parmi lesquels nous retenons : la présence de marqueurs d'intégration linéaire (*si, alors, ensuite, ainsi, de plus, d'autre part,...*) ou des mots de liaison (*par exemple, pour cela, pour ce faire,...*) au début du paragraphe ; la reprise anaphorique au début du paragraphe (reprise par un démonstratif (*ce, cette, ces*), par un pronom personnel (*il, elle*) ; la forte cohésion lexicale

[2]Celle-ci peut s'exprimer en nombre de paragraphes, en nombre de phrases ou même en nombre de mots

des paragraphes consécutifs ; la continuité typographique entre deux paragraphes consécutifs[3].

L'algorithme de segmentation dans IndDoc est donné dans la figure 1. Notre méthode de segmentation permet d'identifier les segments de renvoi (de taille variable) pour chaque descripteur. Elle comporte deux phases. *Une segmentation absolue* qui ne dépend que du document. à ce stade, on segmente le document en UDM, puis on élargit ces UDM en UD en fonction des marqueurs linguistiques et typographiques tout en respectant la structure logique du document (une UD ne peut pas être à cheval entre deux sections de document). Cette segmentation permet donc de découper le document en UD linguistiquement ou typographiquement homogènes. À l'issue de cette phase, le document est représenté comme une liste d'UD. Vient ensuite une phase de *segmentation relative* qui dépend d'un descripteur donné. Cette phase est nécessaire pour établir la liste des segments de renvoi (liste des renvois) d'un descripteur. Elle comporte trois étapes : (1) on identifie les segments de renvoi (les UD qui contiennent le descripteur ou une de ses variantes) ; (2) on regroupe les segments adjacents dans le texte du document, ce qui permet d'obtenir une liste simplifiée de segments de renvoi ; (3) on généralise la séquence des segments d'une même section en un unique renvoi à la section, lorsqu'une partie suffisamment grande de la section figure dans la liste des segments établie en 2. Le fonctionnement de cet

Soit P l'ensemble des paragraphes (UDM) du document.
// élargissement des UDM en UD
Soit UD la liste des unités documentaires du document, initialisée à P
Pour chaque ud_i de UD
élargir ud_i aux UDM qui la suivent dans le texte
s'il n'y a pas de frontière de section entre elles
et s'il y a une continuité linguistique ou typographique entre elles
// Calcul des segments de renvoi
Soit $D = d_1,...,d_m$ la liste des descripteurs extraits
Pour chaque descripteur d_i de D:
Calculer d_i^+, la classe formée par d_i et ses variantes.
Construire S_i^+, la liste ordonnée des segments de renvoi de d_i^+ (les UD où les d_i^+ apparaissent).
// Regroupement des segments occurrences adjacents
Regrouper les segments de S_i^+ qui correspondent à des unités documentaires adjacentes dans le texte
// Généralisation à la section
Soit $\sum$ la liste de toutes les sections et sous-sections du document
pour chaque σ_j de $\sum$
Identifier l'ensemble e_{ij} des segments de S_i^+ qui appartiennent à σ_j.
remplacer l'ensemble e_{ij} par un unique segment-section
si la proportion de paragraphes de l'ensemble e_{ij} dans la section σ_j est supérieure à un seuil donné
Continuité linguistique il y a continuité linguistique entre deux paragraphes, s'il existe un marqueur de liaison linguistique dans l'un d'entre eux (*au delà, d'autre part, ensuite...*)
Continuité typographique il y a continuité typographique entre deux paragraphes, si les deux paragraphes sont typographiquement marqués de la même manière : ils sont tous les deux mis en relief (italique) ou ils constituent deux items de la même liste.

Figure 1: *Algorithme de segmentation.*

algorithme est illustré par la figure 2. Considérons par exemple le texte de la figure 3. Dans un premier temps, ce texte est découpé en 4 paragraphes (UDM). Du fait de la présence de marques de liaison (*effet, De plus*), l'UDM correspondant au paragraphe $\S_i$ est élargie aux paragraphes $\S_{i+1}$ et $\S_{i+2}$. La segmentation absolue donne donc 2 UD : $\S_i$-$\S_{i+2}$ et $\S_{i+3}$. Considérons par ailleurs l'entrée "contexte d'insertion". La seule occurrence de "contexte d'insertion" dans ce document est celle qui figure dans le paragraphe $\S_i$, c'est-à-dire dans l'UD $\S_i$-$\S_{i+2}$. On a donc un unique segment de renvoi mais celui-ci est finalement élargi à la section k elle-même parce que le segment de renvoi contient trois des quatre paragraphes de la section. Nos premières expériences portent sur quatre corpus. L'ouvrage *Ingénierie des Connaissance* (IC99),

[3]Par ex. si les deux paragraphes sont mis en relief de la même manière ou constituent deux items de la même liste)

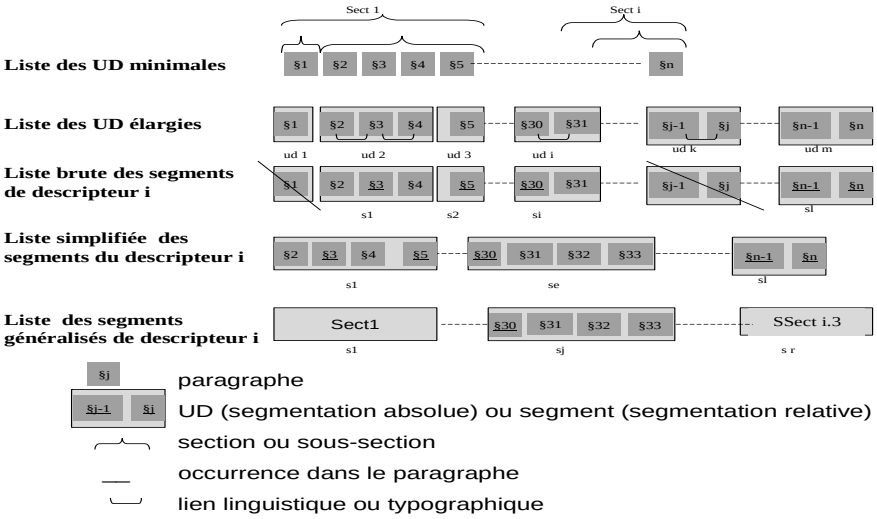

Figure 2: *Illustration de processus de segmentation sur un exemple*

Debut de section k
[§$_i$] Le *contexte d'insertion* d'une ACCA a nécessairement des incidences sur son processus de conception. Il est important...
[§$_{i+1}$] ***En effet,*** pour atteindre l'objectif de création de groupe, l'activité doit prendre place au tout début de l'année, et donc avant la plupart des enseignements. Il faut donc opter pour un domaine qui ne soit ni complètement inconnu des étudiants ni maîtrisé.
[§$_{i+2}$] ***De plus,*** même si dans notre cas le domaine est une variable libre, il faut qu'il se prête à des échanges de point de vue, afin de stimuler les discussions entre les étudiants, les encourager à communiquer et à défendre leurs positions. Le but n'est pas qu'ils se mettent immédiatement d'accord mais, au contraire, qu'ils aient matière à discuter, argumenter et négocier (Berger [1999]).
[§$_{i+3}$] Ces différentes considérations nous ont conduit à proposer une activité liée aux critères...
Fin de section k

Figure 3: *Exemple de segmentation (Les balises entre []sont ajoutées par nous, ainsi que le gras et l'italique).*

qui est un recueil d'articles scientifiques[4], a été choisi pour fournir un point de comparaison avec (Bourigault & Charlet, 1999)[5] . L'ouvrage *Ingénierie des Connaissance* (IC2001) fait suite au précédent. Nous avons également considéré deux manuels d'intelligence artificielle (RC)[6] et de linguistique (EC)[7] pour tester notre approche sur des documents variés. Nous avons appliqué cet algorithme sur ces quatre corpus. On constate que la segmentation réduit effectivement le nombre de renvois mais que le facteur de réduction dépend de la nature du document (monographie *vs* collection) et du style de la rédaction. Le tableau de la figure 4 montre que c'est pour le corpus EC, qui est de style plus littéraire et emploie beaucoup de marqueurs linguistiques, que le facteur de réduction dans le passage des UDMs aux UDs élargies est le plus important. On observe aussi que les étapes de simplification et de généralisation sont moins marquées dans IC, où les articles sont généralement fortement structurés.

[4]L'ouvrage rassemble une sélection de 35 articles publiés dans les actes des Journées Acquisition des Connaissances (JAC), en 1995 et 1996, et dans les actes des journées Ingénierie des Connaissances (IC), en 1997 et 1998.

[5]Les auteurs ont construit un index pour cet ouvrage mais la structuration et le calcul des renvois ont été établis manuellement dans cette première expérience.

[6]*Représentation des connaissance*, D. Kayser, Hermes 1997.

[7]*La cause et son expression en français*, A. Nazarenko, Ophrys 2000.

	IC01	IC99	RC	EC
Nb de UDMs	1085	4929	7386	793
Nb de UD élargies	907	4698	7245	634
Nb de Segments de renvoi	3097	9863	8823	2569
Nb de Segments après simplification	3008	9786	5157	1893
Nb de Segments après généralisation	2997	9728	4469	950

Figure 4: *Résultats de la segmentation sur différents corpus.*

5 Tri de l'information

Nous nous inspirons de l'approche tf/idf pour l'évaluation de la pertinence des différentes clefs d'indexation dans une base documentaire. Cette mesure est adaptée pour prendre en compte, outre le poids d'un mot dans l'ensemble du document et sa fréquence dans le segment, le poids d'une occurrence particulière (qui peut être mise en valeur typographiquement ou dans le discours, figurer dans un titre, etc.).

IndDoc prend en compte deux poids différents : un poids de descripteur ($d_{score}(i)$ pour le descripteur d_i) et un poids de segment ($s_{score}(i,j)$ pour le $j$ième segment occurrence de d_i). Ces poids sont dépendants l'un de l'autre : le poids du segment augmente s'il contient des descripteurs importants et le poids du descripteur augmente s'il apparaît dans des segments importants du document (Maynard & Ananiadou, 2001). En pratique, nous ne tenons compte dans le poids de segment que du nombre d'occurrences des descripteurs que contient le segment ainsi que de leur répartition dans le document.

Le score $s_{score}(i,j)$ est donné par: $s_{score}(i,j) = pts_j . \sum_{k=1}^{D}(\alpha.pds_{kj})$ où D est le nombre de descripteurs dans le document et $\alpha = 1$ pour d_i ou une de ses variantes et 0.5 pour les autres d_k. Le paramètre α permet donc de majorer le poids du descripteur d_i dont on calcule le segment tout en tenant compte des autres descripteurs. Ce score dépend de deux poids élémentaires.

Le poids typographique du segment (pts_j) est intrinsèque au segment s_j. Il est élevé si s_j contient des marques typographiques (gras, italique) ou s'il introduit de nouveaux descripteurs. Ce poids dépend aussi du statut du segment dans le document. Il est majoré, si s_j contient des références ou si s_j est un titre. En revanche ce poids diminue si s_j est un résumé, une introduction ou une conclusion. Nous associons à chaque critère un poids élémentaire et le poids pts_j est une combinaison linéaire de ces poids élémentaires.

Le poids discriminant du segment s_j (pds_{ij}) est relatif au descripteur d_i . pds_{ij} dépend du nombre de descripteurs qui appartiennent à s_j, de leur poids et de leur répartition dans s_j ainsi que dans le document. Ce poids est une mesure révisée de tf/idf : $pds_{ij} = occ_{ij}.log(\frac{P}{P_i})$ où occ_{ij} est le nombre d'occurrences de d_i dans s_{ij}, P est le nombre total de paragraphes dans le document et P_i est le nombre de paragraphes dans lesquels d_i apparaît.

Le score $d_{score}(i)$ est défini par : $d_{score}(i) = psd_i.\sqrt{pdd_i.ptd_i.\sum_{j=1}^{p_i}\frac{s_{score}(i,j)}{p_i}}$. Ce score est calculé à partir de trois poids élémentaires.

Le poids typographique du descripteur ptd_i dépend des caractéristiques typographiques des différentes occurrences de d_i et du poids des segments dans lesquels ce descripteur apparaît. ptd_i augmente si le descripteur apparaît dans des segments importants du document (tels que les titres, le résumé, l'introduction). Nous associons à chaque critère un poids élémentaire, le poids ptd_i est une combinaison linéaire de ces poids élémentaires.

Le poids discriminant de descripteur pdd_i dépend du nombre normalisé des occurrences de d_i

et de sa distribution dans le document. pdd_i est élevé si d_i est irrégulièrement réparti dans le document. Ce poids est une mesure révisée de tf/idf. On pose : $pdd_i = \frac{occ_i}{\overline{occ}}.log(\frac{P}{P_i})$ où $\overline{occ}$ est le nombre moyen d'occurrences par descripteur.

Le poids sémantique du descripteur psd_i dépend du nombre de descripteurs auxquels le descripteur d_i est lié dans le réseau sémantique de la nomenclature d'index. psd_i est élevé si la densité de liens sémantiques de d_i est forte dans le réseau.

La pertinence dépend ainsi d'un ensemble varié d'indices. La fréquence n'est qu'un critère de pertinence parmi d'autres et la typographie, la structure du document, la distribution et la densité sémantique de réseau sont également pris en compte. Nous avons appliqué ces mesures sur nos quatre corpus. Il est difficile d'évaluer ce tri dans l'abstrait mais l'analyse des exemples permet de vérifier le comportement de notre mesure de pertinence. Indépendamment de la segmentation, le tri donne des résultats plus marqués dans IC et EC du fait de leur forte structuration de l'utilisation du gras et d'italique. Au titre d'exemple dans le corpus EC, considérons le descripteur "Contrainte temporelle" qui a une fréquence 12 et possède 3 renvois : le premier dans l'ordre du texte, $S1$, donne une définition du descripteur dont c'est la première mention, et il est écrit en gras. Le segment de renvoi est petit, en revanche. Le deuxième segment $S2$ regroupe une sous-section qui traite de la "contrainte temporelle"(le descripteur figure dans le titre de la sous-section) et deux autres sous-sections où le descripteur est mentionné et dont les titres contiennent respectivement les termes "concordance des temps" et "relation temporelle" qui sont en relation sémantique avec le descripteur " contrainte temporelle". Dans le troisième segment $S3$, le descripteur apparaît en début de segment mais le segment lui-même est inclus dans la conclusion d'un chapitre. Le système a ordonné les renvois en privilégiant $S2$ pour la quantité de l'information et l'apparition dans le titre devant $S1$ dont l'apport d'information est faible même si c'est la première apparition du descripteur et s'il figure en gras. Le renvoi $S3$ est placé en dernière position parce qu'il s'agit d'une partie de conclusion. Signalons pour finir, même si la comparaison ne saurait avoir valeur d'évaluation, que l'index publié de ce livre donne exactement les même segments, dans leur ordre d'apparition, auxquels s'ajoute un renvoi vide (le descripteur n'apparaît pas dans la page)

6 Conclusion

Nous avons montré comment la sélection de l'information est prise en compte dans IndDoc. La méthode proposée s'inspire des pratiques des indexeurs et tire parti des résultats en matière de mesure de pertinence en recherche d'information et en résumé automatique. Nous mettons ici l'accent sur les documents techniques mais ce travail s'intègre dans une réflexion plus globale sur l'intérêt et l'avenir des index. L'originalité de cette approche repose sur la diversité des indices pris en compte : la sélection de l'information dans IndDoc repose à la fois sur des critères de fréquence, de répartition, de typographie et de structure discursive.

Références

AITELMEKKI T. & NAZARENKO A. (2003). Le réseau terminologique, un élément central pour la construction d'index de documents. In *Actes des cinquièmes rencontres Terminologie et Intelligence Artificielle*, p. 1–10.

APTE C., DAMERAU F. & WEISS S. (1994). Automated learning of decision rules for text categorisation. In *Transactions of Office Information Systems*, volume 12.

AUSSENAC N. & CONDAMINES A. (1998). Bases de connaissances terminologiques : enjeux pour la consultation documentaire. *Conférence du chapitre Français de L'ISKO*.

BELLOT P. (2000). *Méthodes de classification et de segmentation locales non supervisées pour la recherche documentaire*. Doctorat informatique, Université d'Avignon et des Pays de Vaucluse.

BOOKSTEIN A. & SWANSON D. (1974). Probabilistic models for automatic indexing. *Journal of the American Society for Information Science*, **25**(5), 312–318.

BOURIGAULT D. & CHARLET J. (1999). Construction d'un index thématique de l'ingénierie des connaissances. *Actes de IC99*.

BOURIGAULT D. & FABRE C. (2000). Approche linguistique pour l'analyse syntaxique de corpus. In *Cahier de Grammaires*, number 25, p. 131–151. Univ. Toulouse-Le Mirial.

CRESTANI F., LALMAS M., VANRIJSBERCEN C. & CAMPBELL I. (1998). "is this document relevant ? ... probably" : a survey of probabilistic models in information retrieval. In *ACM Computing Surveys*, volume 30, p. 528–552.

FERRET O., GRAU B. & MASSON N. (1998). Thematic segmentation of texts : two methods for two kinds of texts. In *actes de Coling-ACL*, p. 392–396, Montréal, Canada.

GROS C. & ASSADI H. (1997). Intégration de connaissances dans un système de consultation de documentation technique. In *Actes des Premières Rencontres du Chapitre Français de l'ISKO (ISKO'97)*: Presses Universitaire du Septentrion.

HEARST M. (1994). Multi-paragraph segmentation of expository text. In *Proceedings of the 32nd Annual Meeting of the Association for Computational Linguistics*,, Las Cruces.

JACQUEMIN C. & BOURIGAULT D. (2003). *Handbook of Computational Linguistics*, chapter Term extraction and automatic indexing, p. 599–615.

KOZIMA H. (1993). Text segmentation based on similarity between words. In *actes du 31 congrès de association for Computational Linguistics ACL (student session)*, p. 286–288.

MANIEZ J. (2002). *Actualité des langages documentaires, fondements théoriques de la recherche d'information.* Adbs edition.

MAYNARD D. & ANANIADOU S. (2001). Term extraction using a similarity-based approach. In D.BOURIGAULT, C. JACQUEMIN & M. L'HOMME, Eds., *Recent Advances in Computational Terminology*, p. 261–278.

A. NAZARENKO & T. HAMON, Eds. (2002). *Structuration de terminologie*.

OUERFELLI T. & LALLICH-BOIDIN G. (2000). Pratiques d'indexation dans les bases textuelles structurées : Application aux textes techniques sous format html. In *ACSI 2000 : les dimension d'une science de l'information globale*, 28e congrès annuel.

PAICE C. (1981). The automatic generation of literary abtracts : An approach based on identification of self-indicating phrases. In O. NORMAN, S. ROBERTSON, C. V. RIJISBERGEN & P. WILLIAMS, Eds., *Information retrieval research*.

PORHIEL S. (1998). *Les introducteurs d'intérêt.* PhD thesis, ParisXIII.

SAGGION H. & LAPALME G. (2002). Generating indicative-informative summaries with sumum. *Computational Linguistics*, **28**.

SALTON G. (1989). Automatic text processing, the transformation, analysis, and retrieval of information by computer. In *Addison-Wesley,Reading*.