

La recherche documentaire : une activité langagière

Vincent Perlerin

Laboratoire GREYC – Université de Caen
bd Maréchal Juin – F14032 Caen Cedex
perlerin@info.unicaen.fr

Résumé

Un nombre important de requêtes soumises aux moteurs de recherche du W3 ne satisfont pas pleinement les attentes des utilisateurs. La liste de documents proposée en retour est souvent trop longue : son exploration représente un travail exagérément laborieux pour l'auteur de la requête. Nous proposons d'apporter une valeur ajoutée aux systèmes de recherche documentaire (RD) existants en y ajoutant un filtrage n'utilisant que des données fournies par l'utilisateur. L'objectif de notre étude est de confronter un modèle dynamique de la mémoire sémantique des individus (ou des agents) développé par notre équipe à une tâche nécessitant une compétence interprétative de la part des machines. Nous souhaitons dépasser la *sémantique lexicale* couramment utilisée dans ce champ d'application pour aboutir à l'utilisation d'une *sémantique des textes* et accroître par ce biais, à la fois la qualité des résultats et la qualité de leur présentation aux usagers.

Mots-clés : recherche documentaire, système anthropocentré, filtrage documentaire, sémantique différentielle et componentielle, isotopie.

Abstract

Lot of the searches submitted to the main search engines on the www do not bring about satisfactory results for users. The list of documents proposed in return is often too long: its handling constitutes an exceedingly strenuous task for the user. We propose to improve existing document retrieval systems by adding a filter made from criteria determined by the user. The objective of our study is to contrast a dynamic model of individual's (or agent's) semantic memory created by our team with a task requiring an interpretative competence from the machines. By this means, we hope to go beyond the *lexical semantics* commonly used in this field of application and achieve the use of *textual semantics*.

Keywords : document retrieval, anthropocentred system, document filtering, differential and decomponential semantics, isotopy.

1 Une recherche documentaire personnalisée

1.1 Degrés de systématique de la sémantique unifiée

Il est important de noter ici que nous n'avons pas la prétention de supplanter les techniques habituelles de RD qui présentent souvent des résultats probants (Strzalkowski, 1999) comme l'ont montré par exemple les conférences TREC¹. L'accroissement incessant des documents potentiellement accessibles aux utilisateurs rend cependant ces techniques toujours plus perfectibles et souvent toujours plus exigeantes en ressources et en traitement. Il est couramment admis que les outils de TAL (Traitement Automatique des Langues) pour la RD requièrent des connaissances linguistiques massives (Jacquemin, 2000). Pour notre part, nous tentons de replacer l'utilisateur au centre du modèle en adoptant une approche résolument anthropocentrée : « (dans un système anthropocentré) ...ce n'est pas l'homme qui essaie d'entrer dans un monde informatique quasi-autosuffisant, mais la machine qui se construit autour des besoins précis de l'homme pour mieux l'assister. » (Thlivitis, 1998). Nous imaginons ainsi les machines comme des mémoires externes individuelles (Lévy, 1990) pouvant, grâce à des techniques de TAL, filtrer les résultats de systèmes plus généraux en fonction des seules ressources fournies par l'utilisateur dans le cadre d'une tâche donnée.

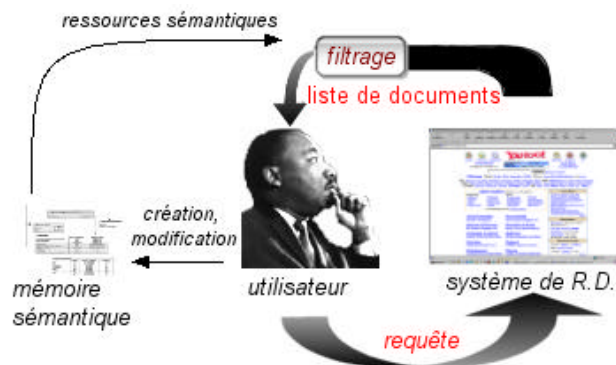


Figure 1 : Utilisation d'une mémoire sémantique pour la RD

Certains systèmes préconisent l'utilisation d'ontologies (Pretschner et al., 1999) ou de thésaurus (Yang et al., 1998) pour l'amélioration de l'indexation ou de la recherche des documents. Nous nous distinguons de ces techniques essentiellement sur la place donnée à l'utilisateur pour exprimer ses besoins et s'impliquer dans le processus de recherche. Dans (Rastier et al., 1994), on décrit les degrés de systématique de la sémantique unifiée comme étant : le système fonctionnel de la langue (les normes de la langue), les normes sociales (considérations essentiellement culturelles) et les normes idiolectales (l'ensemble des régularités personnelles dans les actes linguistiques). Notre approche anthropocentrée trouve une part de sa justification dans cette remarque. Nous pensons que la RD a de nombreux points communs avec une activité linguistique et que les critères énoncés par Rastier doivent donc être pris en compte dans le modèle que nous proposons.

¹ Text Retrieval Conference – <http://trec.nist.gov>

Dans notre système, l'ensemble des descriptions sémantiques nécessaires à la tâche (Figure 1) est créé par l'utilisateur grâce à un logiciel². Ces descriptions en terme d'opposition (1.2) proviennent de l'expérience de l'utilisateur dans son domaine d'intérêt ou de son utilisation d'une aide basée sur une étude de corpus (3.1). Des travaux communs entre les universités de Caen et du Mans sont en cours pour permettre la construction semi-automatique de ces relations à partir d'un système de dialogue Homme/Machine.

Dans notre modèle de la mémoire sémantique, l'utilisateur imprime donc ses normes dialectales : il choisit la langue qu'il désire voir utilisée dans les documents qu'il recherche, il restreint le sens de certains mots polysémiques... Ce qui relève des normes idiolectales est implicite au modèle, car les relations sémantiques sont construites par un utilisateur pour un sujet de recherche donné. Les normes sociolectales font appel à une activité particulière de la part de l'utilisateur. Dans le cadre de la RD, celui-ci se trouve en effet dans une situation dialogique dont il est le seul acteur : il doit prendre en compte pour son objectif, un auteur qu'il ne connaît pas *a priori* (à travers des documents dont il ne connaît pas *a priori* l'existence) mais avec qui il partage un centre d'intérêt : le sujet de sa recherche. L'individu construisant des relations sémantiques doit tenir compte d'une utilisation possible de la langue par un auteur. Cette remarque relève de la définition de la langue en tant que phénomène social (et donc partagé). Lorsqu'on utilise des ontologies ou des thésaurus pré-définis, les normes idiolectales et sociolectales sont délaissées : l'utilisateur doit s'adapter aux définitions préexistantes quelle que soit sa recherche et quel que soit son profil. Certaines méthodes de construction des ontologies tendent à rendre plus importante la place de l'utilisateur à travers l'analyse de son comportement ou la construction de plusieurs ontologies comparées lors d'une recherche (une ontologie du site construite par un agent, une ontologie personnelle de l'utilisateur) (Obiwan, 2000) (Chaffee, 2000). Nous verrons dans la suite de l'article en quoi nos propositions sont différentes.

1.2 Un modèle dynamique de la mémoire individuelle

Notre modèle de la mémoire sémantique des individus (appelé ANADIA - (Coursil, 2000)) trouve ses fondements dans l'observation des interactions sociales. Ce modèle est un *analogie* de la mémoire humaine car il ne repose pas sur des considérations biologiques mais s'inspire d'observations sur le fonctionnement humain et est adapté aux compétences des machines (Auroux, 1998). Nous pensons que la mémoire est une organisation des différences perçues dans l'interaction avec le monde et les autres (Tyvaert, 2000). Cette hypothèse a été renforcée par l'étude d'un corpus de dialogues – Projet PIC³ - (Beust, 1998) qui a établi que pour « *montrer les choses dont on veut parler, pour établir leur valeur sémiotique, on les décrit juste assez pour les différencier des choses avec lesquelles elles pourraient être confondues dans la situation* ». Notre modèle se base ainsi sur l'observation de la construction d'un terrain commun dans un dialogue. C'est en fait, une organisation de choses différenciées entre elles, trouvant leur place à l'intérieur d'un dispositif de catégorisation. Il est fondé sur un principe de discrimination et non de description ensembliste en définissant des types sous lesquels tombent ou non des objets du monde connu sans partir des choses

² Anadia, logiciel libre disponible à l'adresse http://www.info.unicaen.fr/~perlerin/recherche/anadia_prog.htm

³ Le projet PIC propose une étude expérimentale et une modélisation des processus cognitifs et sociaux de conception distribuée à travers les traces qu'ils laissent dans les dialogues. Pour plus d'informations c.f. <http://www.info.unicaen.fr/~fgerard/pic/pic.html>

existantes pour les classer. Ainsi, ce modèle de représentation de la mémoire n'est pas un système de représentations de connaissances au même titre que les ontologies car les choses représentées ne sont pas présumées connues et la représentation qui en est faite n'est ni complète ni intrinsèque. ANADIA a fait l'objet en 2001 d'un article d'une quarantaine de pages (Nicolle et al., 2001). Nous y renvoyons le lecteur désireux d'en connaître les fondements informatiques, linguistiques et cognitifs.

Pour catégoriser les choses, on définit un *dispositif* de discrimination (Figure 2). Au premier stade de sa construction, nous trouvons les *attributs* i.e. des propriétés que l'on peut ou non attribuer aux choses (ex : type *matériel* ou *logiciel* pour différencier les //matériaux de l'informatique//). Le regroupement d'*attributs* pertinents pour un genre à partitionner dans un certain contexte s'appelle *registre* (ex : *matériaux de l'informatique* regroupe les attributs *type* et *rôle* dans la table éponyme). Les *tables* résultent de la combinatoire des places obtenues par croisement des valeurs des *attributs* d'un *registre*, elles sont un moyen de distinguer des catégories dans un même genre. La *sélection* d'une table donnée est l'ensemble des places de cette table reconnues comme valides par l'auteur du dispositif qui a pu les nommer ou en donner des exemples. A certains niveaux de catégorisation, des phénomènes de synonymie locale peuvent intervenir comme dans la table //périphériques// où la ligne *mode d'utilisation* : *sortie* et type de *médium* : *image et texte* correspond à la fois à *écran* et *imprimante*. Finalement, les *topiques* (Figure 3) sont les graphes indiquant la proximité sémantique entre les places sélectionnées d'une table. Elles permettent dans le logiciel Anadia (Nicolle et al., 2001) de présenter des points de vue dont l'utilisateur n'a pas nécessairement conscience en montrant les différences retenues pour les discriminations.

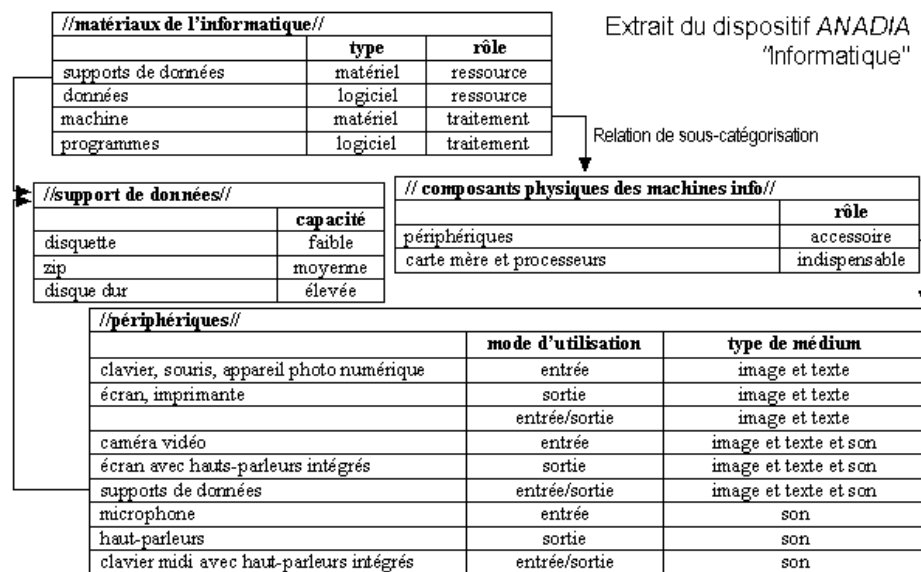


Figure 2 : Extrait d'un dispositif ANADIA en rapport avec l'Informatique

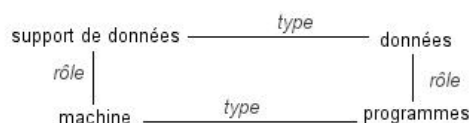


Figure 3 : Topique à 1 différence près de la table //matériaux de l'informatique//

Dans le cadre de la RD, nous ne souhaitons pas utiliser des ressources censées représenter le monde (comme les ontologies) mais des données représentant la façon dont on en parle. L'utilisation d'ANADIA est appropriée à cet objectif : utilisé dans le cadre d'une sémantique componentielle dans une approche différentielle (2.1) et couplé avec le principe d'isotopie (2.3), il permet d'amorcer une compétence interprétative pour les machines, utile pour une tâche de RD.

2 Vers une sémantique des textes.

Les effets de sens sont la seule manifestation des significations : la signification d'un mot dépend de sa contextualisation. Le calcul du sens dans cette contextualisation s'articule autour de deux axes : l'axe paradigmatique qui consiste à représenter les significations du point de vue de leur implication dans l'initiation ou dans le renforcement des effets de sens et l'axe syntagmatique qui doit mettre en évidence l'émergence d'effets de sens au sein des énoncés (ce fait a été exprimé depuis de nombreuses années par Saussure, Hjelmslev ou encore Jakobson (Mel'cuk, 1998)).

2.1 Dimension structurelle : axe paradigmatique

ANADIA peut être utilisé comme modèle sémantique de l'axe paradigmatique i.e. comme un modèle de sémantique lexicale. La représentation du sens relève alors d'une sémantique componentielle différentielle. En sémantique componentielle (Greimas, 1966) (Pottier, 1964), les significations correspondent à des combinaisons de *sèmes* ou traits sémantiques (ex : BLEU 1. *qualificatif, couleur...* 2. *qualificatif, être humain, novice...*). Ici, la notion de sème est redéfinie comme étant un couple *<domaine d'interprétation, jeu d'oppositions>*. Les sèmes sont alors les « *clefs des significations dans les relations qu'elles entretiennent entre elles plutôt que dans les relations que chacune d'entre elles, considérée isolément, entretient avec le monde* » (Nyckees, 1998).

Par exemple (Figure 2) :

<u>Domaine d'interprétation</u>	:	matériaux de l'informatique
<u>Oppositions</u>	:	<i>matériel vs. ressources</i>

Dans cette utilisation d'ANADIA, le domaine d'interprétation d'un sème est représenté par la propriété d'un *attribut* et les traits qui constituent les oppositions forment les *valeurs de l'attribut*. Les sèmes ne représentent ainsi plus un outillage descriptif de la langue. Ils sont rendus opérationnels au regard d'un système de représentation pour une machine par le renforcement du statut de différence qui leur confère une valeur calculable au niveau syntagmatique grâce au principe d'isotopie.

2.2 Dimension fonctionnelle : axe syntagmatique

Depuis les travaux de Greimas, l'*isotopie* est définie comme la récurrence d'un même sème au sein d'une entité textuelle (énoncé, phrase, paragraphe...). Le principe d'isotopie permet de décrire une articulation entre les ordres paradigmatique et syntagmatique et ce, à tous les niveaux de l'analyse sémantique. Avec le modèle du sème défini ci-dessus, le principe d'isotopie s'applique aux oppositions, quelle que soit la valeur marquée dans la lexie, au lieu de s'appliquer aux valeurs elles-mêmes. Au niveau des lexies, les isotopies s'inscrivent dans le choix des contenus retenus comme sens en contexte par afférence. Au niveau de la

phrase et au-delà, elles rendent compte des effets de sens co-textuels. La Figure 4 montre les possibilités du modèle sur deux niveaux touchant à l'interprétation : la levée des ambiguïtés sémantiques et la recherche d'impressions référentielles (l'effet de sens global). Ce sont ces deux aspects de la compétence interprétative qui permettent l'utilisation de ce modèle dans le cadre de la RD.

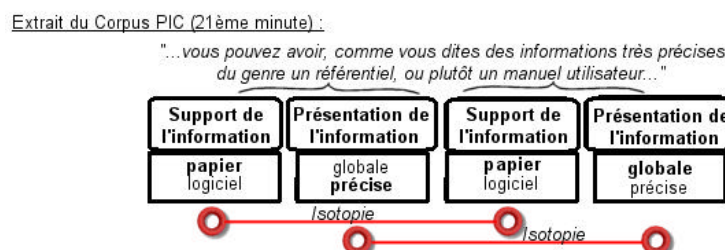


Figure 4 : Principe d'Isotopie – Les sèmes utilisés correspondent à ceux de la Figure 5.

Documentation	Support	Présentation
Manuel	Papier	Globale
Référentiel	Papier	Précise
Aide en ligne	Logiciel	Globale
	Logiciel	Précise

Aide en ligne	Accès	Structure
Assistant	Interactif	hiérarchique
Full text	Interactif	linéaire
Sommaire	Figé	hiérarchique
Index	Figé	linéaire

Figure 5 : Extrait d'un dispositif ANADIA en rapport avec l'Informatique

2.2.1 Désambiguï sation sémantique

La Figure 4 présentent deux isotopies parallèles (différentes formes d'isotopies peuvent être repérées (Beust p127-129, 1998)). La désambiguï sation sémantique porte ici sur *manuel*. L'analyse en terme de sèmes permet, du fait de la présence de *référentiel*, de ne pas retenir d'autres éventuelles descriptions sémantiques de *manuel* pouvant correspondre par exemple à l'adjectif substantivé : *une personne qui travaille avec ses mains*. En effet, la présence d'une isotopie faisant intervenir l'un des sens de *manuel* explicité par l'utilisateur permet à notre système de ne pas retenir d'autres sens (donc d'autres sèmes).

Envisagée indépendamment d'une tâche précise (comme pour son évaluation dans le projet SENSEVAL (Kilgarriff et al., 1998)), la désambiguï sation sémantique (DS) nécessite de larges bases de données comme des dictionnaires ou des thésaurus. Pour la RD, l'objectif de la recherche circonscrit le domaine. Dans notre modèle, la DS consiste alors, dans le cas où une même lexie serait décrite de façon à exprimer plusieurs sens différents, à ne retenir qu'un seul de ces sens en fonction des isotopies découvertes. Il ne s'agit pas d'aboutir à une levée exhaustive de tous les cas d'ambiguï té sémantique, mais de ne retenir que les sens décrits par l'utilisateur et retenus par lui dans le cadre d'une recherche donnée pour décider de la pertinence d'un document. Les ressources utiles à cette tâche sont donc réduites à celles fournies par l'utilisateur.

2.2.2 Recherche d'impressions référentielles

Les isotopies relevées dans la Figure 4 permettent également la découverte des effets de sens au sein de l'énoncé analysé. Au vu des sèmes mis en jeu dans les isotopies repérées,

un module d'analyse peut conclure que l'on parle ici de //Documentation// et sur l'ensemble de la conversation, de //Documentation//, d'//Informatique// et de //Logiciels//. Pour les textes, l'analyse à l'échelle de la phrase, du paragraphe, du document ou d'un ensemble de documents permet la découverte de *faisceaux* d'isotopies. C'est en fonction de la présence et de la couverture de ces faisceaux que la pertinence d'un document par rapport aux attentes de l'utilisateur est décidée. L'impression référentielle calculée à partir des isotopies découvertes au sein des documents est donc au cœur du filtrage proposé. La machine met ainsi en place une interprétation sélective et supervisée car l'utilisateur a toujours la possibilité de modifier son dispositif pour affiner l'analyse sémantique et paramétrer au mieux l'outil logiciel par rapport à ses exigences.

3 Réalisations

3.1 Etude de corpus – Aide à la création des tables

Pour réaliser des logiciels rendant effectives nos propositions, nous avons débuté nos travaux par l'étude d'un corpus homogène de 1783 dépêches journalistiques de l'agence *Reuter*TM traitant du monde de la finance (Perlerin, 2000). L'étude a tout d'abord porté sur la recherche de mots révélateurs du domaine grâce à la loi de Zipf dont l'application au matériau linguistique a montré que la liste des mots d'un corpus de textes classés par ordre décroissants d'occurrence présente tout d'abord les mots grammaticaux, les mots du domaine puis les mots non significatifs du corpus. Les mots retenus ont ensuite fait l'objet d'un calcul de co-occurrence au sein des textes (Figure 6).

	société	banque	groupe	filiale	établissement	entreprise	compagnie	agence	pôle	bancassurance	holding	réseau	assurance
société		77%	66%	18%	40%	43%	18%	11%	9%	5%	23%	29%	27%
banque	80%		49%	18%	44%	40%	21%	12%	9%	6%	18%	27%	30%
groupe	75%	54%		40%	21%	33%	24%	8%	14%	6%	22%	26%	39%
filiale	46%	44%	90%		11%	21%	17%	6%	14%	6%	25%	30%	45%
établissement	95%	100%	44%	10%		64%	16%	14%	4%	1%	12%	29%	19%
entreprises	88%	78%	58%	17%	55%		17%	17%	12%	4%	9%	35%	15%
compagnie	69%	75%	79%	25%	25%	31%		10%	21%	13%	31%	35%	60%
agence	91%	91%	57%	17%	48%	65%	22%		13%	13%	22%	52%	30%
pôle	63%	59%	85%	37%	11%	41%	37%	11%		22%	33%	22%	48%
bancassurance	82%	100%	82%	36%	9%	36%	55%	27%	55%		64%	36%	91%
holding	93%	69%	78%	40%	20%	18%	33%	11%	20%	16%		18%	67%
réseau	90%	81%	71%	36%	38%	53%	29%	21%	10%	7%	14%		31%
assurance	60%	63%	76%	39%	18%	16%	35%	9%	16%	12%	37%	22%	

Figure 6 : Table n°1 des co-occurrences des lexies retenues après étude du corpus. Le tableau se lit : « Dans 77% des fichiers où société apparaît, banque est présent également. »

Ce type de table a permis la mise en place d'une aide à la création des tables de catégorisation. Par exemple, si un utilisateur a partiellement pu nommer les lignes d'une table avec les lexies *société*, *entreprise* et *établissement*. L'aide lui fournit la liste des lexies présentant les plus fortes co-occurrences avec celles déjà présentes dans la table. En l'occurrence : *agence*, *réseau*, *compagnie*, *banque*, *groupe*... Nos expériences itérées sur l'ensemble des tables construites par nos soins à partir des mots du domaine ont montré

(Perlerin, 2000) qu'en supprimant une ligne déjà nommée, le mot initialement présent se trouvait à chaque fois au moins dans les 10 premières occurrences de la liste proposée par l'aide. Nous poursuivons actuellement des travaux dans ce sens.

3.1.1 *Selana* : un logiciel de filtrage à base de dispositif de catégorisation sémantique

Pour tester nos techniques de filtrage et les évaluer par rapport aux techniques les plus basiques faisant appel au pattern matching, nous avons créé deux logiciels. Le premier, appelé *ReDoClef*, permet l'interrogation d'une base de documents (en local) à partir d'une requête de type booléen. Le critère de base du positionnement d'un document retenu dans la liste des résultats est le nombre d'occurrences des mots ou expressions de la requête dans ce dernier.

Le second logiciel, appelé *Selana*, permet l'exploration d'une base documentaire en fonction d'un dispositif de catégorisation sémantique préalablement construit (du type de la (Figure 2)). Les dispositifs utilisés pour nos expériences ont été créés suite à l'étude de corpus décrite ci-dessus. Bien que dans sa version actuelle, le logiciel ne propose pas toutes les fonctionnalités que nous souhaitons lui adjoindre (prise en compte du degré de recouvrement des isotopies à l'intérieur des textes...), nous avons souhaité effectuer une première évaluation permettant la mise au jour des atouts probants de l'utilisation de notre modèle dynamique de la mémoire sémantique de l'utilisateur dans le cadre de la RD. Pour cela, nous avons sélectionné une dizaine de documents dans notre corpus et des corpus traitants de sujets différents. Ne souhaitant pas biaiser nos résultats au vu de l'état actuel de nos programmes, nous n'avons pas mis en oeuvre un protocole expérimental poussé du type calcul de *rappel* et *précision* sur un corpus validé par des experts de type TREC. Les textes sélectionnés ont été notés de 0 à 2 pour décider de leur validité par rapport à une requête traitant du *développement de l'actionnariat*. *ReDoClef* a été ainsi soumis à la requête « *développement de l'actionnariat* » tandis que *Selana* était dirigé dans ses recherches par la sélection de 3 tables en rapport avec le sujet : la table //actions financières// qui regroupe les lexies 'actionnariat', 'présidence' et 'salarial', //acteurs financiers// contenant les lexies 'actionnaire', 'président', 'salarier' et 'petit porteur' ainsi que //établissements financiers// contenant les lexies 'société', 'entreprise', 'compagnie' et 'établissement'.

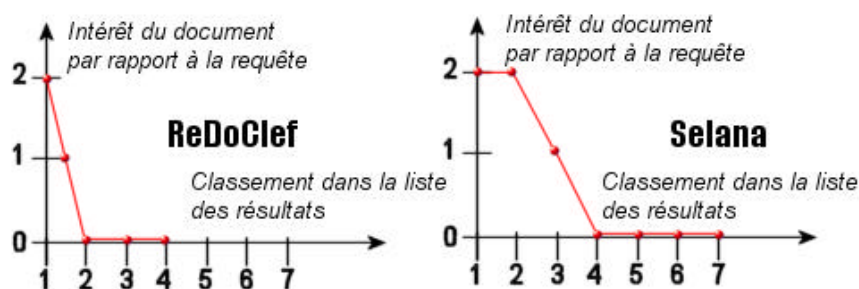


Figure 7 : Résultats de la première expérimentation de nos propositions.

La Figure 7 présente les résultats obtenus par cette expérimentation (détaillée dans (Perlerin, 2000)). Les courbes exposées prêtent un intérêt légèrement supérieur à *Selana* essentiellement parce que le nombre de lexies prises en compte lors de la recherche est plus important dans ce logiciel du fait de l'utilisation du dispositif de catégorisation. Une version plus complexe de *Selana* est en cours de réalisation.

L'utilisation d'une sémantique textuelle permet d'apporter une valeur ajoutée conséquente à la RD ou à la veille technologique, en particulier quant à l'exposition des

résultats d'une recherche. Il en effet est courant dans les systèmes classiques de donner une place meilleure dans la liste des résultats à un document présentant un nombre important d'occurrences des mots de la requête. L'analyse des isotopies couplée à la prise en compte de la morphologie des textes permet de classer de façon originale des textes ne présentant aucune différence en terme de nombre d'occurrences d'un mot de la requête :

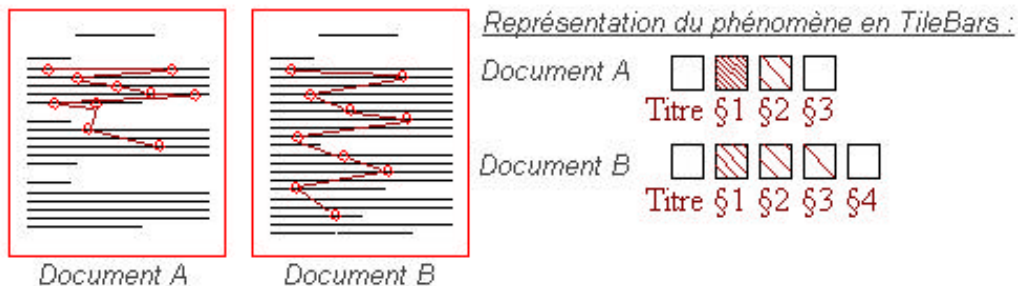


Figure 8 : Faisceaux d'isotopie et représentation en *TileBars*.

Dans la Figure 8, le faisceau d'isotopies du *Doc. A* ne couvre essentiellement que le premier paragraphe. Dans le *Doc. B*, le recouvrement par le faisceau d'isotopies est plus important mais chaque paragraphe pris indépendamment présente moins d'isotopies que le premier paragraphe du *Doc. A*. Deux solutions sont alors envisageables pour tenir compte de ce phénomène. Soit l'utilisateur privilégie l'un des critères suivant : le taux de recouvrement des isotopies, le nombre d'isotopies (...), soit on lui présente les résultats du filtrage personnalisé à l'aide d'objets graphiques de type *Tilebars* (Hearst, 1995) pour lui permettre de visualiser la distribution des paragraphes pertinents par rapport à ses attentes. Dans notre exemple, la densité des traits obliques dans chaque *TileBar* est proportionnelle au nombre d'isotopies repérées dans le paragraphe correspondant.

4 Conclusion

Nos travaux se placent dans un nouveau courant de l'informatique qui préconise une meilleure prise en compte des utilisateurs à tous les niveaux de la conception des modèles calculatoires. L'approche anthropocentrée permet de limiter les ressources pour des tâches jusque là considérées comme nécessairement avides des données toujours plus exhaustives : ici, les données nécessaires sont limitées à celles jugées utiles par l'utilisateur en fonction des résultats obtenus du logiciel. Les rapports entre TAL et RD sont complexes (Jacquemin, 2000) : les systèmes tombent souvent dans l'écueil de performances faibles pour un coût calculatoire élevé. Par nos propositions, nous tentons de fournir à la machine une capacité interprétative n'allant pas au-delà des attentes de l'utilisateur et déterminée par la tâche à réaliser.

Remerciements

L'auteur remercie Anne Nicolle et Pierre Beust pour leur soutien et leur aide précieuse pour les travaux exposés, et Thomas Lebarbé pour l'encouragement à la rédaction de cet article.

Références

Auroux S. (1998), *La raison, le langage, et les normes*, Paris, PUF, 1998.

- Beust P. (1998), *Contribution à un modèle interactionniste du sens*, Thèse de doctorat d'informatique de l'université de Caen.
- Chaffee J.(2000), *Personal Ontologies for Web Navigation*, Thèse d'informatique de l'Université du Kansas.
- Coursil J. (2000), *La fonction muette du langage*, Petit-Bourg, Ibis Rouge Éditions, 2000.
- Greimas A.J. (1966), *Sémantique Structurale*, Paris, Larousse.
- Kilgarriff A., Palmer M. (2000), Special issue on SENSEVAL : Evaluating word sense disambiguation programs, *Computer and the Humanities*.
- Lévy P. (1990), *Les Technologies de l'Intelligence*, Paris, Editions La Découverte.
- Mel'cuk I. (1998), *Vers une linguistique Sens-Texte*, Chaire du Collège de France.
- Nicolle A., Beust P., Perlerin V. (2001), Un analogue de la mémoire pour un agent interactif (en cours de publication - <http://www.info.unicaen.fr/~perlerin>).
- Nyckees V. (1998), *La Sémantique*, Paris, Belin, 1998.
- Obiwan (2000), Ontology Based Informing Web Agent Navigation, projet du National Science Foundation CARREER, *Cooperative Agents for Conceptual Search and Browsing of World Wide Web resources*, www.ittc.ukans.edu/obiwan/index.html.
- Perlerin V. (2000), *Sémantique lexicale pour la RD*, Stage de DEA de l'université de Caen (<http://www.info.unicaen.fr/~perlerin>).
- Pottier B. (1973), sous la direction de Pottier B., *Le langage*, Les encyclopédies du savoir moderne, Paris, Retz.
- Pretschner A., Gauch S. (1999), Ontology Based Personalized Search, Actes de Proc. 11th IEEE Intl. Conf. on Tools with Artificial Intelligence, pp. 391-398.
- Rastier F., Cavazza M., Abeillé A. (1994), *Sémantique pour l'Analyse*, Paris, Masson.
- Strzalkowski T. (1999), *Natural Language Information Retrieval*, Boston, MA, Kulwer.
- Jacquemin C. (2000), Présentation, *T.A.L.*, Vol. 41, pp.327-332.
- Thlivitit T. (1998), *Sémantique interprétative Intertextuelle : assistance anthropocentrée à la compréhension des textes*, Thèse d'informatique de l'Université de Rennes I.
- Tyvaert J.E. (2000), Réflexions épistémologiques et programmatiques sur la place des langues dans la cognition humaine, *InCognito*, Vol. 18, pp.1-10.
- Yang M.C., Wood W.H., Cutkosky M.R. (1998), *Data Mining for Thesaurus Generation in Informal Design Informal Design Information Retrieval*, Standford, Standford University.