Aide à l'interprétation de documents juridiques

Une approche centrée utilisateur

Youssouf SAIDALI* —Julien LECANU* —Eric TRUPIN* Jacques LABICHE*

** Laboratoire d'Informatique, de Traitement de l'Information et des Systèmes, Université de Rouen,*
Avenue de L'Université, 76800 Saint Etienne du Rouvray
Prenon.Nom@univ-rouen.fr

RÉSUMÉ. *Nous présentons un projet de recherche en cours visant à améliorer les interactions d'utilisateurs de différentes catégories professionnelles avec un système d'information dédié au droit du transport et de la logistique. L'objectif vise à concevoir et à mettre au point un environnement numérique de travail (ENT) destiné à un public professionnel (entreprises de la filière logistique, juristes, risk managers, assureurs, avocats, .). et non professionnel (usagers ou salariés des transports). Après avoir posé la question de l'appropriation des contenus dans le cadre des documents numériques, nous décrirons les spécificités de notre corpus de travail. Nous placerons alors notre projet dans un cadre théorique actuellement novateur au sein des sciences cognitives, celui de l'énaction. Ceci nous amènera à proposer une approche résolument centrée utilisateur dans la conception de l'ENT. Nous terminerons par une description des spécifications du futur ENT, qui privilégie une démarche interprétative dans la formulation/reformulation de requêtes, ainsi que la représentation graphique des données.*

ABSTRACT. *In this paper, we present a research project which aims to improve human interactions of various professional categories with an information system dedicated to transport law and logistics. The goal is to conceive and develop a digital environment of work for professionals (companies of the logistic die, lawyers, risk managers, insurers, lawyers,) and non-professionals (users). After having discussed about contents appropriation in digital documents, we will describe specificities of our corpus. Then we will place our project within an innovative theoretical framework in cognitive sciences, that of "enaction". This will lead us to propose user centered approach in the design of the digital environment. We will end with the specifications description of the digital environment, which privileges an interpretative approach in the formulation/reformulation of requests, as well as data vizualisation.*

MOTS-CLÉS : *Environnement numérique, classification, usages, interfaces, visualisation.*

KEYWORDS: *Numerical environment, clustring, uses, interfaces, visualization.*

1. Introduction

1.1. *Interprétation de documents juridiques*

Avec l'apparition des techniques numériques et de l'internet, la distance entre la population et l'information économico-juridique tend apparemment à se réduire. Sur le plan juridique, il est aujourd'hui possible d'accéder à un large pan de la réglementation et de la jurisprudence, qu'elles soient françaises ou étrangères. Mais ce rapprochement technique n'est pas pour autant signe d'une meilleure maîtrise et appropriation de l'information. Les spécialistes en sciences de l'information font le constat des écarts (des fossés) entre prouesses technologiques et appropriation des contenus par des utilisateurs de plus en plus hétérogènes. La question de l'appropriation ne se résout pas en effet par le simple ajout de métadonnées aux documents numériques comme dans le projet du Web Sémantique où l'objectif annoncé par Tim Berners-Lee (1998), initiateur du projet et directeur du W3C, est d'enrichir (notamment au moyen des technologies XML) les documents (à l'aide d'ontologies normalisées, soit automatiquement, soit en assistant leurs auteurs) avec « des informations sur leur propre sémantique qui soient directement interprétables par des agents logiciels sans la supervision d'une interprétation humaine ». Ce positionnement fait l'hypothèse que la valeur sémantique d'un passage de document n'est le fait que de son auteur (alors que c'est tout autant celui de son lecteur confronté à ses pratiques professionnelles). La réponse du Web sémantique est d'indexer les textes avec les concepts d'une ontologie partagée par une large communauté sans que celle-ci ne soit d'ailleurs clairement définie, ni surtout que soient prises en considération la diversité et l'évolution des pratiques langagières au sein de sphères d'activités hétérogènes comme peuvent l'être ici celles des acteurs du transport.

Constatant que les systèmes actuels (IHM, bases de données, ...) conduisent à une interaction Système/Utilisateur forcément appauvrie, parce qu'ancrée dans un environnement prédéfini, (Peschard, 2004) celui des réponses, sous forme de thésaurus qui réorientent la question de l'utilisateur, nous jetterons ici les bases de la conception d'un environnement numérique de travail (E.N.T.) capable de s'enrichir d'apports successifs dus aux interactions de plus en plus denses et complexes au sein de sphères d'activités. Cela nous conduit à sortir de la problématique du mot-clé, ou du figement lexical (représentation de connaissances), pour celle de la thématique des textes et de l'interprétation située. Notre démarche fait l'hypothèse que la valeur sémantique d'un passage est d'abord le fait de son lecteur (entité pouvant être collective) qui *grâce à cette étrange faculté de l'esprit qui est de relier* (Vico, 1986) tracera ses thématiques en fonction de son environnement en même temps qu'il constitue un corpus de textes par sa navigation intertextuelle.

Nous nous intéressons plus particulièrement à la conception d'un *Environnement Numérique de Travail* (E.N.T.), sorte d'extranet dédié aux usages de la filière transport et logistique. Certes, cet ENT ne recèlera guère de fonctionnalités inédites.

En revanche, l'intégration d'un ensemble de ressources et de services interopérables en son tout, dédiés non pas à une collection de cas d'usages particuliers, mais justement à une sphère d'activité (transport et logistique) large et en évolution rapide, constitue une réelle nouveauté, voire une singularité. La mise en œuvre de ce dispositif est susceptible de contribuer à des évolutions notables de l'usage de documents réglementaires et, plus généralement, des activités des acteurs de la filière transport et logistique.

1.2. *Accès au corpus et difficultés de l'appropriation des contenus*

Le corpus réglementaire est encore difficile d'accès malgré une forte demande sociale et économique tout particulièrement en transport et logistique. A ce jour, le corpus et la base documentaire de l'Institut du Droit International du Transport (IDIT[1]) sont accessibles en ligne. Cette base s'adresse à des adhérents spécialistes du droit, mais elle est difficilement utilisable par un novice dans le domaine juridique comme un transporteur, qui chercherait des informations pour la mise en place de conditions de transport de marchandises conformes à la législation en vigueur, par exemple. Cette base documentaire est associée à un thésaurus hiérarchisé « maison » pour améliorer son interrogation. Elle impose aussi la saisie manuelle de comptes-rendus (CR) de jurisprudence et de réglementation sous la forme de fiches. Cette captation de l'information et la veille présentent des difficultés majeures pour renseigner et mettre à jour le système d'information. Nous envisageons, suite à la numérisation (en cours) des collections papiers (des milliers d'articles et de décisions de justice relatifs à des risques et litiges en matière de transports), une aide à l'interprétation de contenus textuels.

Le SI de l'IDIT est conçu pour diffuser des informations aux adhérents afin qu'ils puissent gérer dans les meilleures conditions leurs entreprises et sécuriser leurs activités. Or, la fragmentation de l'information relative au droit des transports et de la logistique qui couvre des domaines très variés rend difficile son accès, d'où la nécessité d'une mise en relief (signalement pour interprétation) de celle-ci.

2. Une approche centrée utilisateur

Notre approche de l'accès aux documents se situe à l'opposé de celles défendues dans le cadre du Web Sémantique. Là où le Web Sémantique cherche à rendre le plus possible partagées de vastes ontologies qui synthétisent une connaissance devant convenir à tous les utilisateurs, nous préférons manipuler des ressources termino-ontologiques (bases de données terminologiques) propres à un utilisateur ou un petit groupe d'utilisateurs et liées à leur tâche, leurs besoins et de leurs centres d'intérêt. Il en découle une certaine *légèreté* sémantique de ces ressources, au sens

[1] L'IDIT, créé en 1969, est une association qui, aux termes de ses statuts, a pour objet l'étude de toutes les questions d'ordre juridique intéressant les transports

de (Perlerin, 2004), dans la mesure où elles ne représentent que ce qui est important du point de vue de l'utilisateur et restent ainsi de taille raisonnable (une centaine de termes) ce qui les rend moins complexes à construire, à maintenir et à enrichir.

Cette approche centrée utilisateur conduit à opérer un certain renversement scientifique relativement aux ressources qu'utilisent les modèles de TAL. Premièrement, d'un point de vue très pratique, force est de constater que des ressources très généralistes, valables pour tout type de traitement envisagé ainsi qu'à destination de tout utilisateur potentiel, ne sont pas facilement disponibles (sous forme électronique pour des traitements automatiques) et encore moins gratuites. Deuxièmement, nous soutenons que l'idée même d'une ressource généraliste est illusoire car elle dépend inévitablement du contexte qui lui préexiste. Le rapport de l'Action Spécifique 32 du CNRS/STIC en 2003 (Charlet *et al.*, 2003) va également dans ce sens en précisant un obstacle au projet du Web Sémantique : la détermination et l'ajout, même de simples méta-données, n'est pas une activité naturelle pour la plupart des personnes.

Les ressources qui sont les plus importantes pour un utilisateur dans une instrumentation informatique pour l'accès aux documents sont celles qui doivent être produites de manière endogène dans une boucle d'interaction entre un outil logiciel, un utilisateur et des corpus. L'accès personnalisé au contenu s'inscrit dans un processus interprétatif en aller-retour entre des outils, des corpus et des ressources personnelles, les uns étant conditionnés par les autres.

Dans notre approche herméneutique et énactive du langage, nous mettons l'accent sur l'interprétation plus que sur les connaissances. Ainsi la priorité est donnée aux spécificités sociolinguistiques des utilisateurs (par exemple leurs centres d'intérêt, leurs habitudes terminologiques, leurs parcours interprétatifs). Ce qui a du sens pour les utilisateurs ne se réduit pas à une représentation et encore moins à une formalisation. Ce n'est pas le résultat d'un calcul, c'est une activité au centre d'une interaction homme-machine. Ainsi, on remet en cause l'idée qu'un mot, une phrase, un texte ou un corpus ait du sens, pour défendre plutôt l'idée qu'ils font sens dans un couplage personne-système.

L'utilisateur et lui seul est capable de dire si un ensemble de documents en retour à sa requête est pertinent ou pas, au vu de sa problématique. Il doit être actif à chaque étape du processus de recherche d'information.

2.1. *Prétraitement et préparation du corpus*

La base de données de l'IDIT comporte plus de 40.000 fiches (jurisprudences, articles, réglementations, acquisitions). Chaque fiche est composée de différentes rubriques (Numéro de la fiche, Thèmes, Date de la décision, Mode de transport, Pays, Objet, Sommaire, Référence), et stockée en format texte.

Pour valider notre démarche, nous effectuerons les premiers tests sur un échantillon restreint du corpus, soient 1369 fiches ayant comme thème « *Conteneur* ».

La première étape de notre développement est de proposer une représentation des documents du corpus. Nous utilisons ici le modèle vectoriel (Salton, 1975), l'une des approches les plus courantes dans le domaine de la recherche d'information.

Nous représentons, les fiches et les requêtes dans un espace vectoriel engendré par l'ensemble des termes contenus dans les fiches. Un terme présent dans un document équivaut alors à une dimension du vecteur [1]. Ce qui permettra par la suite à l'utilisateur de regrouper des items voisins de façon à faire émerger l'information dont il a besoin.

$$T < t_1, t_2, \ldots, t_M > \quad [1]$$

Nous utilisons un algorithme standard pour la pondération des attributs des vecteurs représentatifs des documents, à savoir la fonction TFIDF [2].

$$tfidf_{i,j} = \frac{n_{i,j}}{\sum_k n_{k,j}} . \log \frac{|D|}{\left|\{d_j : t_i \in d_j\}\right|} \quad [2]$$

Avec : - $n_{i,j}$: le nombre d'occurrence du terme t_i dans le document d_j

- $\sum_k n_{k,j}$: nombre d'occurrences de tous les termes dans le document d_j
- $|D|$: le nombre total de fiches dans le corpus
- $\left|\{d_j : t_i \in d_j\}\right|$: le nombre de fiches où le terme t_i apparaît

Nous aurons ainsi la valeur de chaque attribut équivalant à l'importance d'un terme dans un document relativement à l'ensemble des documents. L'avantage de cet algorithme est qu'il va permettre de supprimer la totalité des mots qui seront communs à toutes les fiches (Numéro, jurisprudence, thème, etc.). De plus, cette pondération va permettre à l'utilisateur de mettre en avant les termes les plus représentatifs du document en fonction de sa problématique.

A l'issue de cette étape, nous obtenons un vecteur de 11913 attributs représentatif de chaque document. Le corpus contenant 1369 documents au total, on obtient donc une matrice terme-document de 11913 lignes sur 1369 colonnes. Nous notons ici l'un des inconvénients de ce modèle de représentation. Il crée des vecteurs de très grande taille et de ce fait, difficilement utilisable dans une interaction homme-machine dynamique, notamment pour la classification mais aussi pour la reformulation de requêtes. En effet, un tel vecteur va considérablement augmenter les temps calculs, et ainsi retarder l'obtention des résultats à une requête,

ce qui n'est pas concevable pour un système que l'on veut interactif. Il est donc impératif de réduire sa taille, en limitant au maximum la perte d'information. Nous avons alors cherché à supprimer les termes qui ne portent pas d'information à un instant donné et donc qui ne sont pas discriminants pour la représentation des documents. Nous partons ainsi d'un principe assez répandu en recherche d'information qui dit que les mots les plus fréquents (articles, prépositions, auxiliaires) sont vides de sens. Les premiers utilisateurs (3 spécialistes de l'IDIT), ont défini tout mot comptant moins de quatre lettres comme étant vide de sens. En réalisant cette suppression de termes, nous sommes en mesure de passer la taille d'un vecteur représentatif d'un document à 11178 attributs valable dans le contexte de ce groupe d'utilisateurs. Malgré cette première réduction, le nombre d'attributs de chaque vecteur représentatif reste encore trop élevé. Nous proposons une nouvelle diminution, en nous appuyant sur des méthodes issues de la linguistique, notamment la lemmatisation (ou radicalisation) et sélection de mots pleins, en regroupant les variantes d'un mot.

Exemple : économie, économiquement, économiste économ

Toutefois, un problème non négligeable d'ambiguïté des termes peut se produire. En effet, le fait de réduire les termes à leur radical peut avoir comme conséquence de regrouper des mots ayant le même radical mais n'ayant pas le même sens. Par exemple, on peut faire correspondre « *transporter* » et « *transformer* » à travers le radical « *trans* ». Ces regroupements erronés ont comme conséquence la perte en précision. Cependant, cette radicalisation permet de réduire nos vecteurs représentatifs de manière significative, passant de 11178 à 7243 (Tableau 1).

	à l'origine	Stop words	Lemmatisation
Taille vecteurs	11913	11178	7243

Tableau 1. *Tailles des vecteurs suite aux réductions de dimension*

Le corpus est donc maintenant correctement préparé par et pour l'utilisateur. Il s'agit ensuite de proposer, dans le contexte de l'utilisateur, une classification dans le but de former des ensembles de documents thématiquement proches.

2.2. *Classification contextuelle*

Nous partons simplement du principe que si un document est jugé pertinent à une requête, alors les documents similaires à ce document ont de fortes chances d'être également pertinents. Nous allons donc laisser l'utilisateur regrouper les documents au sein de *clusters*, et relever ses thèmes majeurs ou ceux de sa sphère d'activité. De nombreuses méthodes ont été utilisées en classification de textes (Sebastiani, 2005). Dans la mesure où nous ne disposons pas (et ne souhaitons pas constituer) d'exemples d'items déjà classés et étiquetés, nous nous intéressons ici uniquement aux méthodes de classification non-supervisées pour catégoriser les

documents. Pour faciliter l'usage de l'ENT, et pousser à l'émergence d'information sans connaissances a priori, on intègre plusieurs méthodes exploitables par l'utilisateur.

2.2.1. *Approche par les k-means*

La première méthode de classification que nous avons intégrée dans l'ENT est les k-means (McQueen, 1967). L'intérêt de cette méthode, pour nous, vient du fait que la valeur k, correspondant au nombre de classes à créer n'est pas fixée à l'avance. Il peut donc apparaître comme un élément du contexte utilisateur. Nous avons réalisé les premiers tests de classification du corpus pour différentes valeurs de k. A l'analyse des fiches placées dans un même cluster, les utilisateurs ont ainsi mis en avant que plus k est grand et plus les clusters sont représentatifs de leurs thèmes. Toutefois, si le nombre k devient trop grand, il apparaît une sur-segmentation du corpus. Ceci risque alors de provoquer une perte du contexte pour l'utilisateur dans sa navigation intertextuelle.

L'évaluation de la pertinence dans le tableau ci-dessus correspond à un jugement utilisateur, après navigation dans les fiches contenues dans un cluster.

2.2.2. *Approche par une classification hiérarchique*

Toujours dans le but d'aider à l'émerge nce de connaissances par l'usage de l'ENT, nous proposons également une classification hiérarchique descendante, qui structure le corpus sous forme d'un arbre (Figure 2).

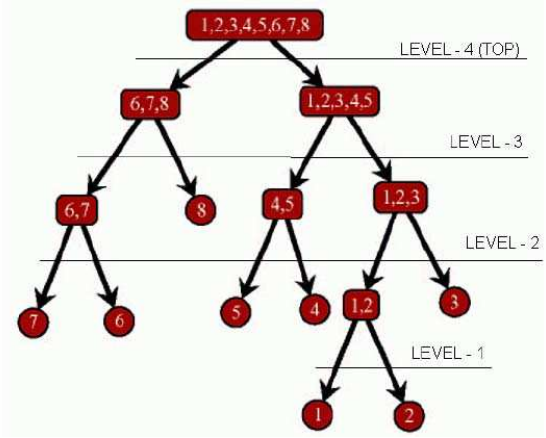

Figure 2. *Exemple de classification hiérarchique ascendante*

Cette classification aboutit à 251 clusters qui contiennent entre 1 et 20 documents chacun. Une analyse rapide de cette hiérarchie valide la répartition des fiches dans les clusters en fonction des thèmes de l'utilisateur. Et montre que dans le contexte courant, deux clusters situés aux extrémités de l'arbre sont jugés totalement

différents. L'avantage de cette seconde approche est qu'elle fournit une topologie facilement représentable et manipulable pour la visualisation et la navigation.

Maintenant que l'utilisateur a classifié ses documents, il peut s'appuyer sur cette structuration pour exprimer sa requête et affiner son besoin d'information.

2.3. *Expression du besoin et reformulation de requête*

Dés lors que le corpus est classifié, et que dans chaque classe les documents sont jugés thématiquement et sémantiquement proches par l'utilisateur, on peut proposer des outils de formulation-reformulation de requêtes dans notre ENT. Cette procédure se décompose en vectorisation de la requête utilisateur, calcul de la distance entre requête et clusters et reformulation par réinjection de pertinence.

2.3.1. *Requête initiale*

Une requête peut être définie comme étant l'expression formalisée d'un besoin d'information exprimé par l'utilisateur. Elle est dans la plupart des cas composée d'une suite de termes permettant de décrire ce besoin d'information. Cependant, pour pouvoir être comparée aux documents du corpus par un calcul de distance, la requête doit être représentée avec le même formalisme que les documents. Elle sera donc représentée par un vecteur (de 7243 attributs) identique à ceux des documents. De plus, à chaque terme de la requête sera affecté un poids, déterminé en fonction du nombre d'occurrences du terme, normalisé par le nombre de termes de la requête. Le poids pour un terme t est ainsi défini par la fonction suivante : $p_t=n/N$ où n est le nombre d'occurrences du terme dans la requête et N le nombre total de termes.

2.3.2. *Calcul de distance*

La requête étant vectorisée, nous pouvons aider l'utilisateur à établir la similarité entre sa requête et les différents documents du corpus [3]. Cette similarité apparaît comme une distance entre les deux vecteurs. Nous aurons donc pour chaque document d_i :

$$sim(q, d_i) = \overrightarrow{v(q)}.\overrightarrow{v(d_i)} \quad [3]$$

Cependant, nous ne travaillons pas directement sur les documents mais sur des clusters de documents, en calculant un vecteur représentatif de chacun d'eux. Un cluster est alors vu comme la moyenne terme à terme des vecteurs représentatifs des documents. Nous aurons donc :

$$\vec{V}_{cluster} = \sum \frac{\overrightarrow{v_d}}{N} \quad [4]$$

Avec $\overrightarrow{v_d}$ les vecteurs représentatifs des documents et N le nombre de documents dans le cluster.

Pour déterminer la similarité entre ses requêtes et les clusters, l'utilisateur peut exploiter un cosinus, une métrique fréquemment utilisée en fouille de texte

2.3.3. *Reformulation par injection de pertinence*

Pour affiner de façon incrémentale le besoin utilisateur, nous utilisons une retro-propagation de la pertinence sur la requête. L'idée est de générer une nouvelle requête par une combinaison linaire des éléments de la requête initiale, et de l'avis de l'utilisateur sur les documents extraits. Dans cette approche, si les *n* premiers clusters trouvés sont jugés pertinents[2], ils sont réutilisés pour reformuler la requête de l'utilisateur (Figure 3).

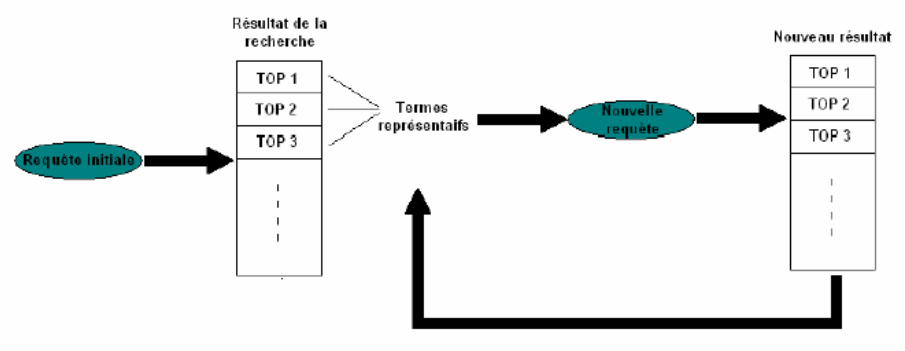

Figure 3. *Schéma de principe de la reformulation*

Les n clusters sélectionnés pour la reformulation sont les premiers clusters du classement réalisés lors de l'étape précédente. Nous ajoutons à la requête initiale de l'utilisateur les termes les plus représentatifs de ces clusters. Par précaution, et pour éviter que les termes ajoutés suite à la reformulation ne deviennent prépondérants par rapport à la requête initiale, ils sont pondérés. Toutefois, on constate les clusters placés loin dans le classement d'origine (*ie* jugés peu pertinents), peuvent remonter.

Nous ne cherchons pas ici, à automatiser le processus, ni à fournir le meilleur résultat. On a fait le choix dès le début de plonger l'utilisateur (ou un petit groupe d'utilisateurs) dans des interactions qui offrent la possibilité de notamment mieux discerner l'homogénéité thématique d'un corpus, de mettre en évidence sa densité, d'en extraire les principales tendances thématiques de chaque document et de permettre un accès rapide à tel ou tel document ou passage de document. Dans le couplage personne-système les interprétations des utilisateurs et les calculs des machines vus précédemment, ne sont pas en concurrence car les uns n'ont en aucun cas le but de supplanter les autres. Au contraire, nous les pensons comme complémentaires dans le sens où l'activité d'une machine a pour objectif de

[2] *n* étant un paramètre que l'utilisateur peut affiner en fonction de son besoin et de la visualisation des résultats

produire dans l'interaction des traces qui vont participer aux interprétations du ou des utilisateurs.

3. *Ver*s un environnement numérique de travail personnalisable

D'un point de vue expérimental, il s'agit de savoir comment un environnement numérique de travail, et le couplage qu'il induit, permettent l'émergence par énaction d'une perception sémantique du corpus et ainsi un meilleur accès aux informations (Varela, 1989).

Dans notre stratégie d'amélioration de la recherche d'information, nous proposons à l'utilisateur plusieurs approches pour naviguer dans l'ensemble des documents, visualiser, manipuler et organiser le résultat de ses recherches. Il pourra notamment s'appuyer sur l'historique de sa navigation, ses propres traces, mais aussi celles qui sont liées à sa sphère d'activité (collectif de travail). Il s'agira donc d'observer l'utilisateur dans son activité, et de lui permettre d'exploiter dynamiquement cette observation. Avec ses traces (volontaires ou involontaires), nous ne cherchons pas à modéliser un comportement pour faire de la prédiction, mais à disposer d'outils de description et d'analyse de la navigation intertextuelle en situation réelle.

La visualisation et l'analyse des résultats de la recherche sont des étapes nécessaires qui s'inscrivent dans le processus global de recherche d'information. La perception de l'information est liée à la prise de décision dans le contexte d'utilisation de l'ENT proposé. L'utilisateur se retrouve au centre d'une boucle itérative « *formulation-analyse-visualisation-reformulation* » dans une représentation globale du processus comme celle de la Figure 4 (inspirée de Kules, 2008).

Le processus est donc initialisé lorsqu'un utilisateur identifie un besoin informationnel et tente de le satisfaire en entreprenant une ou plusieurs tâches de recherche. Il prend des décisions sur la ou les stratégies à adopter, les outils à exploiter et le corpus ou partie du corpus à consulter. Chaque unité d'information découverte peut déclencher de nouvelles idées, suggérer de nouvelles directions et changer la nature même du besoin d'information. On émet alors l'hypothèse que, la gestion sous forme d'historiques de traces (incluant point de blocages et retours arrière) laissées par les différents utilisateurs peut aider à la découverte de nouvelles stratégies et de nouvelles informations.

Pour ce qui est de l'extraction d'information, chaque action implique un engagement cognitif et physique, et peut induire une évolution dynamique de l'interface ou des connaissances en émergence. Nous cherchons à faciliter le couplage et l'engagement de l'utilisateur en lui proposant des outils simples pour la manipulation/sélection/déplacement des documents résultats, ainsi que pour l'expression dynamique des requêtes. Notre démarche consiste donc à utiliser des

représentations spatiales dynamiques pour concevoir et mettre en œuvre une plate-forme générique en personnalisation de la visualisation et en intégration de modalités variées et hétérogènes.

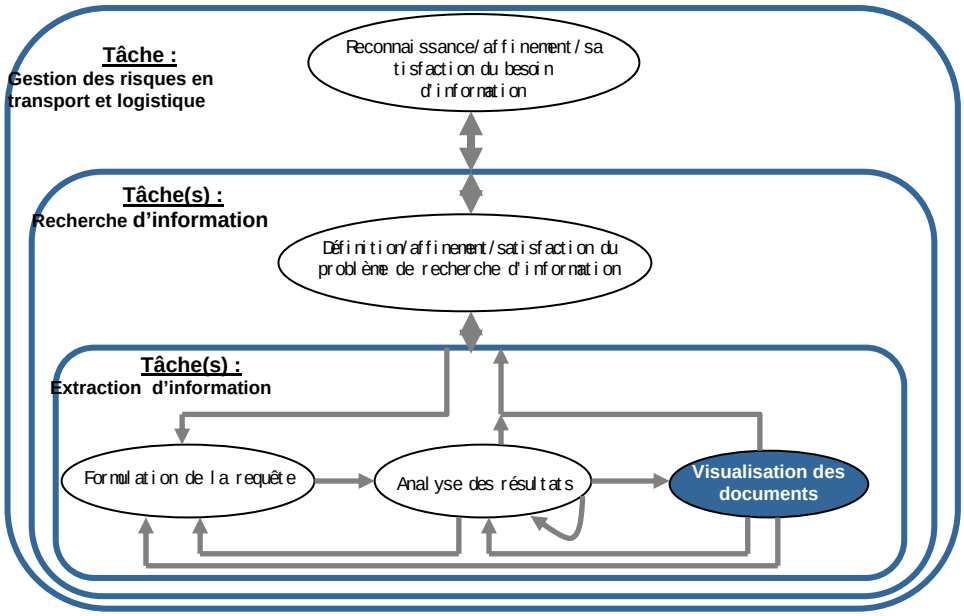

Figure 4. *Visualisation dans un processus d'aide à l'interprétation*

5. Conclusion

A travers cet article et notre projet de mise au point d'un ENT dans le domaine juridique, nous avons voulu poser d'une manière particulière la problématique de la recherche d'information. Nous avons cherché à mettre en avant la complémentarité entre un agent humain énactif et un système classique en RI. Le document électronique est ici considéré de manière indissociable à l'activité du ou des humains qui les produisent, recueillent, indexent et recherchent. Il en découle à notre avis que les approches autour de la recherche d'information ne pourront rester indépendantes des utilisateurs et sans prise en compte des paliers d'intertextualité comme le sont actuellement par exemple les moteurs de recherche ou encore certains projets dans le contexte du web sémantique. Nous proposons d'intégrer plusieurs modalités de manipulation et de navigation dans les données. Cette navigation dynamique et visualisation à différents niveaux de granularité de

l'ensemble des documents permet alors à l'utilisateur de se créer son propre parcours interprétatif.

Plus que jamais la problématique de la recherche d'information requiert des collaborations pluridisciplinaires pour mettre au point, expérimenter et évaluer les conditions d'une relation entre documents et interprétants d'où puisse émerger du sens et de nouveaux usages.

6. Bibliographie

Salton G., Wong A., and Yang C.S., « A Vector Space Model for Automatic Indexing », *Commun. ACM*, vol. 18 (11), 1975, p. 613-620.

Bates M J : «*The design of browsing and berrypicking techniques for the online search information* ». Online review, 13, 407,-431, 1989.

Berners-Lee, T. « *What the semantic web can represent?* » W3C, Disponible à : www.w3.org/designissues/rdfnot.html., 1998.

Peschard, I. « *La réalité sans représentation, la théorie de l'enaction et sa légitimité épistémologique* ». Thèse de Philosophie, 2004, Ecole Polytechnique.

Ricoeur, P. « *Du texte à l'action : essais d'herméneutique* ». Point Seuil, 1986, Paris.

Vico, G. « *Principes d'une science nouvelle* ». (trad JL. Lemoigne) Nagel 1986.

Perlerin, V. *Sémantique légère pour le document* ». Thèse d'informatique. Université de Caen. 2004 .

Charlet, J., Laublet, P., Reynaud, G. « *Web sémantique* ». Rapport de l'Action Spécifique 32 CNRS/STIC, 2003.

Fabrizio Sebastiani. « *Text Categorization* ». In Alessandro Zanasi, editor, Text Mining and its Applications to Intelligence, CRM and Knowledge Management, pages 109--129. WIT Press, Southampton, UK, 2005.

McQueen J. B. « *Some Methods for classification and Analysis of Multivariate Observations* », Proceedings of 5-th Berkeley Symposium on Mathematical Statistics and Probability", 1967, Berkeley, University of California Press, 1:281-297.

Varela, F. « *Invitation aux sciences cognitives* ». Seuil, 1989, Paris.

Kules, B., Shneiderman, B. « *Users can change their web search tactics : Designs guidelines for categorized overviews* ». Information Processing and Management, 2008.

Jacko, J. A., Sears, A. «*The Human-Computer Interaction Handbook : Fundamentals, Evolving Technologies and Emerging Applications* ». 2nd Edition, Lawrence Erlbaum Associates, 2006.

Sebastiani, F. «*Text Categorization* ». In Alessandro Zanasi, editor,Text Mining and its Applications to Intelligence, CRM and Knowledge Management, pages 109--129. WIT Press, Southampton, UK, 2005.