PROTOTYPE DE SYSTÈME EXPERT POUR L'AIDE À L'ANALYSE (TRI, CLASSIFICATION, INDEXATION) DES DOCUMENTS DE JURISPRUDENCE

Suzanne BERTRAND-GASTALDY et Gracia PAGOLA
École de bibliothéconomie et des sciences de l'information
Université de Montréal
Case Postale 6128, Station A
Montréal, Québec, CANADA H3C 3J7

Jean-Guy MEUNIER, François Daoust
et Louis-Claude PAQUIN
Centre ATO•CI
Université du Québec à Montréal
Case Postale 8888, Station A
Montreal, Québec, CANADA H3C 3P8

RÉSUMÉ

Le projet vise à concevoir un système expert pour l'aide à l'analyse des jugements qui vont parvenir d'ici peu en grande quantité sur support magnétique à la SOQUIJ (Société québécoise d'information juridique). Il ne cherche pas à remplacer la pratique courante, mais à l'optimiser et à l'enrichir par des informations complémentaires. Il consiste d'abord à modéliser les décisions prises par les conseillers juridiques pour sélectionner ces jugements, pour les trier, les classifier et les indexer, puis à construire et à opérationnaliser des algorithmes (d'ordre linguistique et statistique) d'analyse en plein texte pour assister les opérations humaines, et finalement à mettre au point une maquette de système expert.

Le projet s'accompagne d'activités d'information et de formation pour transférer l'expertise dans l'organisation. La productivité et la performance de l'entreprise seront accrues, lui permettant de mieux répondre aux besoins d'une clientèle nombreuse qui consacre beaucoup de temps à la recherche documentaire.

La méthodologie et les analyseurs mis au point pourront être adaptés dans les organisations qui doivent faciliter l'accès à des textes juridiques et réglementaires informatisés.

CONTEXTE

La Société québécoise d'information juridique (SOQUIJ) a pour mandat de "promouvoir la recherche, le traitement et le développement de l'information juridique en vue d'en améliorer l'accessibilité au profit de la collectivité" et, plus particulièrement, de publier les jugements rendus par les tribunaux judiciaires du Québec dans des recueils imprimés et des banques de données interrogeables en direct. Elle reçoit actuellement plus de 10 000 jugements par année, chiffre auquel il faut ajouter les jugements de la Cour suprême du Canada et les décisions des tribunaux administratifs.

La saisie électronique des jugements à la source mise en place progressivement par le ministère de la Justice du Québec portera à près de 50 000 le nombre de jugements acheminés sans aucune sélection préalable.

Afin de maintenir le même service sans accroître indûment le personnel, SOQUIJ a confié à une équipe de recherche du CEFRIO (Centre francophone de recherche en informatisation des organisations) le mandat de concevoir un prototype de système expert qui allègera certaines des tâches des conseillers juridiques chargés de traiter les jugements à plusieurs niveaux (tri, sélection, classification, indexation, résumé, documentation, édition).

OBJECTIFS:

Le système a pour objectif d'assister les conseillers juridiques dans l'analyse de ces documents.

Plus spécifiquement, le système doit:

1) aider à éliminer à la source les jugements non motivés et ne présentant pas d'intérêt;
2) aider à trier et à classifier les jugements selon les différents domaines du droit répertoriés dans un plan de classification subdivisé en 57 domaines;
3) suggérer des descripteurs extraits d'un thésaurus (qui contient plus de 700 descripteurs) et des mots-clés libres pour l'indexation.

HYPOTHÈSES DE TRAVAIL

Le projet s'appuie sur la triple hypothèse qu'il est possible:

1) de modéliser les décisions prises par les conseillers juridiques pour analyser les arrêts;
2) de construire des algorithmes d'analyse des jugements en plein texte pour assister ces décisions;
3) d'opérationnaliser les algorithmes dans un milieu réel pour un corpus fortement normalisé comme celui de la jurisprudence.

FONCTIONNALITÉS GÉNÉRALES DU SYSTÈME EXPERT

Plusieurs chaînes de traitement transforment les textes intégraux et attribuent aux unités lexicales et textuelles une série de propriétés à partir desquelles diverses opérations sont réalisées. Au terme de ces opérations, les données sont filtrées et analysées sur le plan statistico-linguistique et des coefficients de confiance sont attribués. Grâce à la prise en compte de l'incertitude, le système peut aboutir à plusieurs réponses, chacune étant qualifiée d'un coefficient cumulé indiquant le degré de confiance qu'on peut avoir.

Les opérations suivantes sont effectuées sur les textes intégraux:

- Conversion des codes d'édition.
- Prétraitement de certains caractères: désambiguïsation du point d'abréviation et de fin de phrase; identification des majuscules de noms propres.
- Segmentation des textes en paragraphes et identification des passages exposant le litige, le contexte et la décision du juge.
- Catégorisation des deux premières et des deux dernières phrases de chaque paragraphe.
- Repérage et catégorisation de différentes informations accompagnant le jugement: intitulé, provenance, lois et articles de lois cités, parties au litige, etc.
- Identification des termes simples et complexes du domaine, des rubriques de classification et des descripteurs du thésaurus.
- Divers calculs statistiques.
- Etc.

À partir de ces données, les règles proposent un diagnostic pour l'élimination éventuelle du jugement, répartissent les jugements en domaines du droit, suggèrent l'attribution d'une ou plusieurs rubriques de classification et l'assignation de descripteurs et de mots-clés libres.

MÉTHODOLOGIE DE CONCEPTION

Les sources de données

La modélisation, la mise au point et l'opérationnalisation des algorithmes s'appuient sur plusieurs sources de données. Certaines sont enregistrées sur support informatique. Il s'agit des notices bibliographiques produites à la suite des différentes opérations d'analyse, des textes intégraux et des outils documentaires (plan de classification et thésaurus). D'autres données sont constituées des critères, généralement implicites, auxquels recourent les conseillers juridiques pour prendre leurs décisions et qu'il s'agit de mettre au jour.

Extraction de l'expertise

L'extraction de l'expertise s'effectue par deux approches complémentaires qui s'enrichissent mutuellement. D'une part, à l'aide de traitements statistico-linguistiques, on compare les données des différents textes en interrelation (textes primaires et textes secondaires). Ainsi, à travers les traces laissées par les conseillers juridiques dans les résultats des différents types d'analyse, on essaie de reconstituer les opérations cognitives. On tente de trouver quelles unités lexicales et textuelles et lesquelles de leurs propriétés contribuent le plus à différencier les éléments retenus des éléments non retenus, aux différents niveaux de condensation. D'autre part, on interviewe les conseillers juridiques, on les observe dans l'exercice de leurs tâches et on recueille leurs commentaires sur les résultats des analyses textuelles.

L'ajout de propriétés dans le logiciel SATO (Système d'Analyse de Textes par Ordinateur)

Diverses propriétés sont ajoutées, de façon automatique ou semi-automatique, aux chaînes de caractères et aux segments textuels; elles sont d'ordre éditique, grammatical, lexico-sémantique, "documentaire", textuel, pragmatique, statistique, etc.

Dans l' exemple suivant extrait d'une notice, on peut voir, en contexte, le marquage de certaines propriétés et de certaines de leurs valeurs. On notera:

1) l'identification de segments comme la provenance, la manchette (à mi-chemin entre le titre enrichi et l'indexation), les subdivisions correspondant au litige, au contexte et à la problématique, etc.
2) la numérotation des phrases à l'intérieur des paragraphes et leur catégorisation en première (pr), deuxième (deux), dernière (de), avant-dernière (ad), autres (au);
3) la mention de l'italique identifiant, entre autres, des lois citées;
4) la distinction entre mots-clés libres et unités lexicales issues du plan de classification ou du thésaurus, distinction qui est rendue possible par la consultation automatique de ces outils documentaires.

NOTICE 91-3.STR

***par=ident*typo=nil**<ND>91-3 ***par=provenance** <HD>COUR D'APPEL

par=manchette** ASSURANCEpc=oui** -- assurance de responsabilité***mot-clé=oui** -- recours contre le tiers responsable***mot-clé=oui** -- option***th=oui** -- article 2603 C.C. -- interdiction de cumul***mot-clé=oui** -- amendement***th=oui**.

***par=litige *phr=1 *ord=(ad,pr)** Appel d'un jugement de la \Cour supérieure ayant accueilli une requête en irrecevabilité. ***phr=2 *ord=de** Rejeté, avec dissidence.

***par=contexte *phr=1 *ord=pr** Le 18 février 1988, l'appelante a intenté une action contre la mise en cause \Fontaine, lui réclamant 23 688$ à titre de dommages à la suite d'un incendie provoqué par sa négligence. ***phr=2 *ord=deux** Quelques mois plus tard, l'appelante a fait signifier une déclaration amendée qui ajoutait la compagnie d'assurances intimée à titre de défenderesse et qui concluait à la condamnation conjointe et solidaire des codéfenderesses. ***phr=3 *ord=au** L'intimée a alors présenté une requête en irrecevabilité fondée sur le fait que l'appelante n'avait aucun recours contre elle puisque, en poursuivant \Fontaine, elle avait exercé l'option prévue à l'article 2603 \C..\C.. . ***phr=4 *ord=au** La requête en irrecevabilité a été accueillie malgré la demande verbale d'amendement présentée par l'appelante visant à modifier la désignation des parties et à ne maintenir que l'intimée à titre de défenderesse, reléguant \Fontaine au rang de mise en cause. [...]

***par=décision *phr=1*ord=pr *typo=italique** \Mme la juge \Tourigny et \M.. le juge \Proulx: ***typo=nil** Les dispositions du ***typo=italique** Code de procédure civile ***typo=nil** relatives à l'amendement doivent recevoir une interprétation aussi large que possible. ***phr=2 *ord=deux** Cependant, une interprétation, aussi large soit--

elle, ne peut écarter une disposition de droit substantif incluse dans le ***typo=italique*** \Code civil. ***typo=nil *phr=3 *ord=au*** Le législateur a voulu que, en intentant un recours, la partie demanderesse fasse un choix, ainsi que l'a confirmé \M.. le juge Mayrand dans l'arrêt \L'\Union québécoise, mutuelle d'assurance contre l'incendie c.. \Mutuelle des \Bois-Francs: [...]

par=référence \Compagnie d'assurances \Traders générale c. \Laurentienne générale, \Compagnie d'assurances inc.. Juges \Tourigny, \Proulx et \Chouinard (diss..). C.A.

Ces autres exemples montrent des propriétés statistiques, grammaticales, pragmatiques (appartenance au domaine du droit) et lexico-sémantiques, hors contexte, dans le lexique:

Propriétés statistiques

Mo.	Éc.	Rép.	Dis.	Chi2	lexique
9.73	4.02	100.0%	0.00	28.79	a
0.04	0.19	3.8%	0.45	21.98	abri
0.12	0.42	7.7%	0.46	25.03	absence
0.04	0.19	3.8%	0.45	21.98	acceptant
0.08	0.27	7.7%	0.36	16.65	acceptation
0.04	0.19	3.8%	0.49	24.26	acceptation du risque
0.04	0.19	3.8%	0.31	15.07	acceptent
0.04	0.19	3.8%	0.49	24.26	accepter
0.04	0.19	3.8%	0.27	12.88	accès aux documents
0.35	0.87	15.4%	0.92	49.72	accident
0.04	0.19	3.8%	0.31	15.07	accident d' automobile
0.04	0.19	3.8%	0.56	28.05	accident du travail
[...]					
0.12	0.58	3.8%	1.61	80.00	acte criminel
0.04	0.19	3.8%	0.44	21.76	actes fautifs
0.58	0.93	30.8%	0.65	35.07	action
0.04	0.19	3.8%	0.31	15.07	action en dommages
0.08	0.27	7.7%	0.35	19.05	action en dommages-intérêts
0.23	0.42	23.1%	0.32	19.68	action en réclamation

Propriétés grammaticales

fréq	gramr	(lexique)
13828	préposition-ff	à
10	nomcommun_fs	abandon
1	conjugué_fv	abandonna
1	conjugué_fv	abandonnaient
1	partprésent_fv	abandonnant
2	conjugué_fv	abandonne
20	partpassé_fv	abandonné
1	partpassé_fv	abandonnée
5	infinitif_fv	abandonner
1	partpassé_fv	abandonnés
[...]		
1	partpassé_fv	abattu
10	nompropre	Abitibi
1	adjectif_fs	abitibienne
1	nompropre	Abitibi-Témiscamingue
1	nomcommun_fs	ablation
1	nomcommun_fs	abolition
4	adjectif_fs	abondante
1	adjectif_fs	abondantes
32	nomcommun_fs	abordages

Appartenance au domaine de droit

domaine	(lexique)
non	Abitibi
oui	acte_d'_accusation
oui	action_en_dommages-intérêts
oui	agents_de_la_paix
oui	agression_sexuelle
peut-être	arme_à_feu
oui	arrestation_sans_mandat
oui	arrêt_des_procédures
oui	centre_de_détention
oui	chef_d'_accusation
oui	conduite_avec_facultés_affaiblies
oui	conseil_de_famille
oui	contrat_de_mariage
oui	contrat_de_vente
oui	divorce
oui	dommages_exemplaires
oui	donation_entre_vifs
oui	droits_de_la_personne
oui	jeunes_contrevenants

Synonymie

synonyme	(lexique)
agression_sexuelle	abus_sexuel
activité_pyramidale	vente_pyramidale
alcool	état_d'_ébriété
alcool	Régie_d'_alcool_du_Québec

On peut ainsi attribuer autant de propriétés que l'on juge utile pour la reconstitution de l'expertise, d'après l'intuition, les résultats des analyses et le savoir-faire des conseillers juridiques. À tout moment, on peut se positionner sur une unité lexicale et obtenir, par une simple commande, la liste de toutes les propriétés et valeurs de propriétés attribuées:

communauté_de_biens

*alphabet	=	fr
*fréqtot	=	11
*longueur	=	19
*gramr	=	tcomposé
*poids	=	51
*typo	=	nil
*par	=	contexte
*phr	=	4
*ord	=	ad
*term	=	non-descript

Filtrage des données

Afin de connaître les meilleurs prédicteurs pour les opérations d'analyse, on procède au filtrage des données selon différentes propriétés et combinaisons de propriétés. On peut ainsi isoler des:

- formes simples ou complexes
- termes du domaine
- rubriques de classification

- mots-clés libres ou contrôlés
- titres de lois
- noms des parties en présence

et opérer une sélection supplémentaire selon:

- leur position dans la macro-structure et la micro-structure du texte
- leur fréquence absolue ou relative
- leur valeur discriminante
- le chi2
- le domaine du droit
- le tribunal d'où provient le jugement
- le type de parties en présence
- etc.

Analyse des données

Les données filtrées sont soumises à des analyses statistiques, dont l'analyse de discrimination, et font l'objet de comparaison et d'évaluation par les experts.

Résultats de l'enquête cognitive

Dans le cas de l'indexation, opération très difficile à formaliser, les résultats obtenus par l'analyse statistico-linguistique sont confrontés aux résultats de l'analyse humaine et soumis aux conseillers juridiques qui , petit à petit, réfléchissent à leur démarche, explicitent leur savoir-faire, ce qui permet d'affiner, par itération successive, le choix des propriétés discriminantes et la pertinence des règles.

Pour le tri et la classification, les conseillers juridiques ont d'emblée identifié les indices pertinents (l'intitulé du jugement, le nom des parties, les lois ou articles de lois cités, la présence de certains termes dans le texte du jugement).

Ainsi, pour le domaine Assurance:

- l'intitulé n'est pas un bon indice;
- si le nom d'une des parties désigne une compagnie d'assurances, on peut hésiter entre le domaine Assurance et le domaine Responsabilités (une compagnie d'assurances peut, en effet, poursuivre la personne qui lui a causé des dommages);
- si les articles 2468 à 2676 du Code civil ou une loi sur les assurances sont cités, cela renforce l'indice précédent;
- si le jugement comporte des termes comme: assurance-automobile, assurance collective, assurance-vie, assurance-invalidité, etc., alors la décision de classer le jugement en Assurances est fiable.

Adaptation des outils documentaires

Le plan de classification et le thésaurus sont harmonisés; les quelques incohérences sont corrigées; le thésaurus subit un enrichissement important: variations flexionnelles, morphologiques et syntaxiques, synonymes documentaires.

Mise au point du système expert

Pour le tri et la classification, les indices qui permettent de prendre une décision sont de valeur différente. Certains peuvent s'avérer prépondérants et simplifier la tâche comme la cour, l'intitulé du jugement ou les lois citées. Par contre, en l'absence de tels indices, un examen plus approfondi est nécessaire pour extraire des indices terminologiques. Il est impossible de constituer a priori un univers de référence qui permettrait de distribuer un certain poids sur les indices. On traite donc indépendamment chacun des indices et on opère un cumul rapporté sur la ou les rubriques vers lesquelles pointe l'indice. On s'appuie sur la théorie des fonctions de confiance de Shafer (1976). Elle fournit une indication quant à la contradiction générée par la combinaison d'indices qui pointent dans des directions différentes. Cette mesure indique la fiabilité à accorder aux résultats.

Le même fonctionnement est adopté pour l'indexation: les indices sont constitués de la nature des unités lexicales, de leur fréquence, de leur valeur discriminante, de leur appartenance au domaine et de leur position dans certains passages-clé du jugement. Des règles faisant appel à la fois à la structuration du thésaurus et au contexte d'emploi (cooccurrences) sont appliquées.

Le système informatique consiste en une chaîne de traitements séquentiels. SATO est utilisé d'une part pour appliquer les bases de connaissances permettant de reconnaître des indices dans les textes (certains indices peuvent être reconnus hors contexte, alors que la reconnaissance d'autres indices dépend du contexte d'occurrence), d'autre part pour constituer et gérer ces bases de connaissances. Nous envisageons de formaliser les modèles de traitement après validation dans un système expert, ce qui nous permettrait un contrôle plus fin et une plus grande sensibilité au contexte. Cette modélisation en système expert sera faite sur ACTE (Atelier Cognitif et TExtuel) qui est en cours de développement au centre ATO•CI. Quant à la théorie des fonctions de confiance, elle est actuellement programmée en ICON. Le design est ainsi fait qu'il est possible de calibrer par essai-erreur l'attribution de poids aux différentes classes d'indices: un multiterme qui appartient au thésaurus vaut plus qu'un multiterme du domaine qui lui-même vaut plus qu'un terme simple, etc.

BÉNÉFICES ESCOMPTÉS POUR SOQUIJ

La démarche de conception du système expert aura eu des retombées positives pour SOQUIJ avant même que celui-ci soit implanté. En effet, la nécessité d'expliciter le processus d'analyse et les règles suivies par chacun des conseillers juridiques aboutit à une prise de conscience de certaines divergences selon les individus; en outre, les résultats des analyses statistico-linguistiques poussent à un examen critique des outils documentaires et de leurs interrelations ainsi que des pratiques "manuelles" et de leurs conséquences. Les changements sont effectués par les conseillers juridiques eux-mêmes, à leur initiative. Au terme de l'opération, un recueil de politiques et procédures

est disponible et le thésaurus se trouve enrichi pour les nécessités d'une analyse assistée par ordinateur.

L'implication constante des futurs utilisateurs du système garantit la pertinence des solutions proposées et crée une disponibilité à l'apprentissage de la nouvelle technologie.

Les suggestions du système expert respectent le plus possible les habitudes actuelles, avec tout le savoir qu'elles supposent concernant le domaine, les textes, les besoins des utilisateurs finals des banques de données et recueils imprimés, les contraintes éditoriales, etc.

Les analyses mixtes qui précèdent la mise au point du système constituent un compromis tenant compte de caractéristiques exigeant parfois des solutions contradictoires: matériau textuel très complexe à analyser, volume important des données prohibant des traitements très fins et rendant possibles des effets de nombre, nécessité de prendre en compte à la fois les caractéristiques linguistiques et les opérations cognitives d'experts au savoir-faire très riche.

BIBLIOGRAPHIE

Shafer, G., (1976). *A mathematical theory of evidence*. Princeton University Press, Princeton, New Jersey. 296 p.