Université de Montréal

L'utilisation de vecteurs de liens bibliographiques comme descripteurs de documents juridiques dans un système de recherche d'information

Par

Ernst PERPIGNAND

Département d'informatique et de recherche opérationnelle

Faculté des arts et des sciences

Mémoire présenté à la Faculté des études supérieures
en vue de l'obtention du grade de
Maître ès sciences (M.Sc.)
en informatique

Août 1998

Université de Montréal
Faculté des études supérieures

Ce mémoire intitulé :

« L'utilisation de vecteurs de liens bibliographiques
comme descripteurs de documents juridiques
dans un système de recherche d'information »

Présenté par

Ernst PERPIGNAND

a été évalué par un jury composé des personnes suivantes :

Jean Vaucher, président du jury

Gena Hann, membre du jury

Paul Bratley, directeur

Daniel Poulin, codirecteur

Mémoire accepté le : 22.12.1998

Sommaire

Ce mémoire présente un travail de recherche relatif à la représentation des documents pour des fins de recherche d'information. Au cours de ce projet, nous nous sommes intéressés à l'utilisation des liens bibliographiques tels que les références et les citations comme descripteurs de documents juridiques.

Le but de ce travail est de vérifier l'hypothèse que des décisions présentant une similarité au niveau des différents types de liens bibliographiques induits par les références jurisprudentielles sont également similaires au niveau de leur contenu. Dans cet optique, nous avons élaboré un modèle vectoriel où les composantes sont rattachées à ces liens et la similarité des documents est évaluée suivant la similarité de ces vecteurs.

Dans un premier temps, ce mémoire présente les inconvénients intrinsèques aux représentations par mots clés utilisées dans la plupart des systèmes de recherche d'information traditionnels. Dans la même foulée, les avantages des liens bibliographiques sont mis en évidence et différents travaux antérieurs sont passés en revue.

Dans un deuxième temps, ce rapport décrit le modèle de représentation de document retenu ainsi qu'un prototype développé selon ce modèle. Par la suite, les résultats obtenus avec le prototype sont présentés et analysés.

Le mémoire se termine par une explorations des différentes améliorations et avenues intéressantes envisageables dans des travaux futurs.

Table des matières

Tables des figures et des tableaux

Remerciements

À

Évelyne, ma mère,

qui n'a jamais cessé de m'encourager

Je n'aurais pas pu terminer ce mémoire sans le soutien de plusieurs personnes. Aussi, je profite de ces quelques lignes pour les remercier vivement. Mes codirecteurs, Daniel Poulin et Paul Bratley, me viennent en premier à l'esprit. Tout au long de mon parcours à l'université, Daniel fut pour moi un patron, un modèle et un ami hors pair. Son attention et sa confiance quasi paternelles m'ont procuré un support inestimable tandis que ses conseils judicieux m'ont toujours guidé dans les décisions importantes que je devais prendre pour mener à bien ce projet. Paul, de son côté, fut toujours disponible et n'a jamais hésité à me prodiguer ses conseils judicieux tout au long de l'élaboration de mon projet de maîtrise.

Je tiens également à remercier le Centre de recherche en droit public et son personnel pour leur appui. Une pensée spéciale va au professeur Pierre-André Côté qui m'a suggéré des manuels d'introduction au droit me permettant ainsi de mieux cerner les aspects juridiques du sujet traité dans ce mémoire.

Ce mémoire n'aurait pas pu être achevé sans le soutien financier de l'équipe LexUM du Centre de recherche en droit public dirigée par Daniel Poulin. Je le remercie de ce support.

Finalement, je tiens à remercier ma famille et mes amis. Les encouragements et le soutien inconditionnels qu'ils m'ont donnés m'ont permis de mener à bien ces études. Je leur suis très reconnaissant.

Introduction

Tôt ou tard, la plupart des gens éprouvent un besoin d'information ; que ce soit un scientifique à la recherche d'un article sur un sujet particulier, un médecin voulant connaître les effets d'un nouveau médicament sur des patients atteints d'une certaine maladie, un juriste fouillant la jurisprudence à la recherche de précédents pour un cas particulier ou tout simplement une personne s'intéressant aux dernières nouvelles concernant son activité favorite. Fort heureusement, de nos jours, les ressources d'information abondent. L'émergence d'Internet et d'autres technologies informatiques nous a, un peu sans crier gare, propulsés dans ce qu'il est convenu d'appeler «l'ère de l'information». Le «Word Wide Web», communément appelé le Web, illustre mieux que n'importe quel outil l'avènement de ce phénomène de prolifération de l'information. En effet, de nos jours, nombreux sont ceux qui parcourent le Web régulièrement en quête d'articles scientifiques, de documents juridiques ou de publicité pour des produits et services. Le nombre élevé de sites publiant et mettant régulièrement à jour leur information sur le Web ne laisse pas douter de la quantité et de la diversité de l'information qui y est accessible de part le monde.

Cependant, si la grande quantité d'information disponible sur le Web est l'un de ses avantages, cette richesse peut aussi constituer un de ses inconvénients, car l'utilisateur risque d'être incapable de l'exploiter utilement. En effet, même si la navigation dans l'environnement hypertexte qu'offre le Web peut être considérée comme un procédé de recherche d'information adéquat, elle ne peut à elle seule suffire à protéger l'utilisateur d'interminables et, trop souvent, vaines heures de consultation. En effet, Savoy

[Savoy 96] remarquait que dès que le nombre de documents (ou liens hypertextes) devenait grand (plus de 500 par exemple), la méthode perdait de son efficacité. Dans un tel contexte, la nécessité d'un outil permettant de repérer l'information pertinente se fait sentir.

Ceci dit, le besoin d'outils de recherche d'information existait bien avant l'émergence du Web. La plupart des moteurs de recherche aujourd'hui offerts pour le Web, basés sur différents modèles tels que le modèle booléen, le modèle probabiliste ou le modèle vectoriel, pour ne citer que ceux-là, ont été développés bien avant le développement fulgurant d'Internet. Voyons rapidement les principales caractéristiques de ces diverses approches.

À part les systèmes de recherche utilisant la reconnaissance de patterns sur lesquels nous ne nous attarderons pas ici, le système de recherche d'information le plus simple se fonde sur le modèle booléen. Dans la mise en œuvre de ce modèle, l'utilisateur spécifie les termes pertinents, les relations logiques et l'ordre qu'ils doivent respecter au moyen d'une requête booléenne. Le système analyse cette requête et y répond en fournissant un ensemble de documents satisfaisants aux contraintes données. Cependant, le modèle booléen, malgré sa simplicité, exige que l'utilisateur maîtrise le domaine sur lequel il veut avoir des informations. En effet, comme les systèmes basés sur le modèle booléen ont une tendance à introduire beaucoup de bruit (documents non pertinents), ils obligent l'utilisateur à procéder à des modifications et à des raffinements itératifs de ses requêtes ; tâche souvent ardue et ingrate. De plus, ces systèmes ne disposent pas d'informations supplémentaires leur permettant de présenter les documents retrouvés dans un ordre décroissant de pertinence par rapport à la requête. Certains systèmes de

types booléens contournent ce problème en triant les documents selon leur date de parution par exemple, avant de les présenter à l'utilisateur. Il est clair, cependant, que ce genre de procédé ne donne pas toujours les résultats escomptés.

Le modèle booléen hybride constitue une alternative plus efficace. Dans ce second modèle, un ensemble de termes est associé à chaque document et un poids exprimant l'importance du terme pour le document est associé à chaque paire terme document. Dans le contexte d'une requête, le degré de pertinence des documents satisfaisant aux contraintes logiques est évalué en tenant compte du poids des termes présents à la fois dans le document et la requête.

Une autre approche permettant de mesurer de façon plus précise le degré de pertinence d'un document a été proposée par Gérard Salton dans le système SMART [Salton 83]. Dans ce système, chaque document est considéré dans un espace vectoriel où les dimensions sont associées aux termes d'indexation. Les termes d'indexation sont choisis parmi les termes des documents de façon automatique par un algorithme qui extrait les mots représentant les concepts les plus importants du corpus. À l'issue de l'étape d'indexation, un document est représenté par un vecteur où les composantes représentent le poids d'un terme d'indexation (associé à la dimension correspondante) dans le document. Les requêtes, exprimées dans un langage naturel, sont également traduites dans un format similaire. La pertinence d'un document par rapport à la requête est alors évaluée en calculant le cosinus de l'angle formé par les deux vecteurs, celui de la requête et celui du document. Plus le document est pertinent à la requête, plus les vecteurs se rapprochent et, par conséquent, plus le cosinus de leur angle est élevé. Par la suite, un

tri sur ces valeurs permet d'ordonner les documents suivant leur degré de pertinence par rapport à la requête.

Finalement, une autre approche, celle des réseaux bayésiens, utilise la théorie des probabilités pour modéliser le processus de recherche d'information. Selon cet autre modèle, un document est retourné dès que la probabilité qu'il soit pertinent à la requête est supérieure à la probabilité qu'il ne le soit pas. En utilisant cette stratégie de recherche, le théorème de Bayes et en supposant que les termes apparaissent de façon indépendante dans les documents, il est possible de dériver une formule de pondération des termes de la requête de telle façon que la réponse du système soit optimale. Remarquons que la supposition d'indépendance des termes dans un texte n'est pas tout à fait juste. Cependant, cette approche donne quand même de bons résultats.

Ces modèles possèdent leurs avantages et leurs inconvénients. Cependant, ils sont tous basés sur les termes contenus dans les documents et cela pose des problèmes tant à leurs concepteurs qu'à leurs utilisateurs. À ce chapitre, trois problèmes peuvent être identifiés.

La difficulté de préparer une requête à la fois précise et descriptive à l'aide des termes susceptibles d'apparaître dans les documents cherchés constitue le premier problème de cette famille de systèmes de repérage. Le système booléen est à cet égard le pire. La difficulté s'explique aisément si on se rappelle qu'un concept n'est pas déterminé de façon unique par un terme. En effet, dépendant du contexte dans lequel ils sont utilisés, certains mots tels que : enfant, mineur, juvénile, sont synonymes ou non. Réciproquement, des termes tels que : droit, banque, nouvelle, ont plus d'une signification. Dans le premier cas, une solution consiste à introduire tous ces termes dans

la requête alors que dans l'autre, il faut penser à formuler la requête de sorte à exclure les interprétations indésirables. Il en résulte que la formulation d'un besoin d'information est plus compliquée pour un utilisateur qui ne maîtrise pas les concepts du sujet sur lequel porte sa recherche.

Les systèmes basés sur le modèle vectoriel ou probabiliste, en permettant à l'utilisateur de formuler plus facilement sa requête dans un langage naturel, ne résolvent qu'en partie ce problème. L'intégration de thesaurus et de dictionnaires de synonymes, construits manuellement ou de façon automatique en se basant sur la cooccurrence des termes, ainsi que la rétroaction sont des procédés susceptibles d'augmenter l'efficacité, c'est-à-dire, le nombre de documents pertinents retournés. Malheureusement, la plupart des utilisateurs ne comprennent pas le fonctionnement de ces outils et ils n'en tirent pas profit.

Un deuxième problème des systèmes de repérage traditionnels est en rapport à la flexibilité de la syntaxe de la langue naturelle. En effet, si le modèle booléen propose un mécanisme permettant de spécifier l'ordre dans lequel devraient apparaître les termes cherchés, le fait qu'il existe plusieurs façons d'ordonner les mots pour exprimer un concept réduit l'efficacité d'un tel procédé. Ainsi, exiger que le terme « mari » précède « femme » ne permet pas de distinguer le concept «le mari trompe sa femme» de «la femme trompe son mari», car ce dernier peut être exprimé de la façon suivante : « le mari est trompé par sa femme ». D'un autre côté, les variantes syntaxiques de certains mots nuisent à leur identification. Aussi, plusieurs algorithmes de lemmatisation qui consistent à extraire les lemmes des mots doivent-ils être utilisés lors de l'indexation automatique

afin d'identifier la racine des termes. Malheureusement, ce procédé introduit de nouveaux homographes inattendus et, encore une fois, c'est la précision des systèmes qui en souffre.

Le troisième problème des systèmes de repérage classiques vient du fait qu'en attribuant un ensemble de termes aux documents, ces systèmes laissent de côté la sémantique des phrases. En effet, dans le modèle vectoriel, par exemple, «le mari trompe sa femme» et «la femme trompe son mari» sont considérés comme identiques parce que représentés par le même vecteur alors que ces deux phrases ne véhiculent pas tout à fait la même idée. Le traitement du langage naturel est une branche assez récente de la recherche d'information qui essaie de remédier à cette lacune des systèmes traditionnels et qui tente de représenter le contenu sémantique des documents ou des requêtes. En plus de faciliter l'interaction homme machine, de telles approches devraient permettre d'indexer des structures de phrases plus complexes que les termes et ainsi d'améliorer la précision des systèmes. Malheureusement, les traitements du langage naturel ne sont pas suffisamment maîtrisés et efficaces pour être inclus dans les systèmes de recherche d'information actuels.

Ainsi, l'utilisation des mots comme descripteurs du contenu des documents n'est sans créer de problèmes. Deux approches peuvent être explorées pour les solutionner. La première, que nous avons effleurée dans notre description précédente, consiste à améliorer les systèmes basés sur les termes existants. Ceci peut se faire de différentes façons. Il est possible, par exemple, d'utiliser des thesaurus ou de concevoir d'autres modèles permettant de mieux cibler le besoin d'information des utilisateurs. La seconde approche, que nous privilégions, favorise l'exploration d'autres moyens de représentation

du contenu des documents, et plus précisément dans le cas de ce travail, des références qu'ils contiennent.

En effet, les références jouent un rôle important dans la dissémination de l'information. Ainsi, il est courant qu'un auteur inclue tout au long ou à la fin de son texte des références à d'autres sources qu'il juge pertinentes ou susceptibles d'intéresser le lecteur. Ce dernier, suite à la lecture d'un texte intéressant peut décider de suivre les références de l'auteur afin d'approfondir le sujet. Il peut également décider de rechercher les documents qui font référence à ce texte de manière à trouver de l'information plus récente. Cette hypothèse à l'effet que les références et les liens qu'elles établissent peuvent servir de base à un système de repérage d'information guide la présente recherche.

L'exploration de cette idée impose deux tâches. D'abord, il faut élaborer le modèle du système de repérage basé sur les références où les documents sont représentés par les liens bibliographiques (références, citations) qui le caractérisent. Le travail de conception suppose aussi la consultation des rapports de recherche qui font état des travaux antérieurs qui se sont intéressés à la question. Il nous faut ensuite développer un prototype permettant de vérifier notre hypothèse.

L'outil qui a été développé dans le cadre de cette recherche permet à un utilisateur de naviguer à l'intérieur d'un corpus en lui proposant les documents similaires selon notre modèle fondé sur les liens bibliographiques. La plus grande partie du prototype a été implantée en Java. Ce choix s'explique en partie par la portabilité de ce nouveau langage. Pour le reste, la conception orienté objet nous est apparue idéale car, comme nous le verrons plus loin, elle offre la souplesse nécessaire à l'exploration future des diverses

avenues qu'offre le modèle. Alors la possibilité de réutiliser le code sera très utile pour permettre les améliorations et les ajouts. Le prototype a été testé avec un corpus de décisions de la Cour suprême du Canada. Ce corpus possède des caractéristiques intéressantes que nous aborderons plus loin dans ce texte.

Le présent mémoire décrit et présente les conclusions de ce travail. D'abord nous établissons le cadre conceptuel du projet et les caractéristiques du modèle que nous avons élaboré. Ensuite, nous présentons le prototype implanté ainsi que les résultats obtenus. Finalement, nous précisons certaines limites de ce travail ainsi que certaines améliorations que pourraient lui apporter des futurs travaux.

Chapitre 1

Les références, les citations et le repérage

1.1 Notions préliminaires et définitions

Bien que dans le langage courant les termes *références* et *citations* soient plus ou moins interchangeables, nous allons donner ici à ces termes un sens plus strict et légèrement différent de leur sens habituel. Nous appellerons *item bibliographique* la chaîne de caractères désignant un document référé. Ici, nous rompons avec la tradition du langage courant d'appeler cette chaîne "référence" ou "citation". Pour les fins de ce mémoire, un item bibliographique établit une relation entre deux documents. Les termes *référence* et *citation* viennent préciser le rôle de chaque document dans cette relation. Pour nous, la phrase « B est une référence de A » indique que le document A évoque B au moins une fois. Ceci implique que A contient au moins un item bibliographique qui désigne B. Inversement, « B est une citation de A » signifie que B évoque A au moins une fois et que, par conséquent, le document A est une référence de B. La "Figure 1" met en évidence la dualité entre citations et références. En effet, à chaque référence correspond une citation. Dans les figures de ce type, par convention, la pointe de la flèche indique la référence, c'est-à-dire, le document référé. Ainsi, dans la "Figure 1" le document A est une référence de B et, par le fait même, nous déduisons que le document B est une citation de A, c'est-à-dire que B évoque A au moins une fois.

Figure 1 Dualité des références et citations

Les items bibliographiques établissent donc des liens, que nous désignerons par *liens bibliographiques*, entre différents documents d'un corpus. Ces liens sont induits soit directement par un item bibliographique comme dans le cas des *liens de référence* et des *liens de citation* introduits plus haut, soit par un ensemble d'items bibliographiques partagés entre deux documents comme dans le cas du *couplage bibliographique* et de la *cocitation* que nous définirons plus loin. Ces liens bibliographiques peuvent être mis en évidence par un graphe dont les arcs indiquent les documents qui participent aux différentes relations. La "Figure 2" nous donne un exemple d'un graphe de liens de références.

Finalement, ajoutons que pour un document initial quelconque, il est possible d'établir l'ensemble de ses références en parcourant son texte et en notant les items bibliographiques qu'il contient. La construction de l'ensemble de ses citations, par contre, exige plus de travail. Il faut consulter tous les documents du corpus afin de noter ceux qui citent le document initial. De façon alternative, nous pouvons avoir recours à un index de

citations, c'est-à-dire, un recueil dans lequel nous retrouvons certains documents accompagnés de leurs citations.

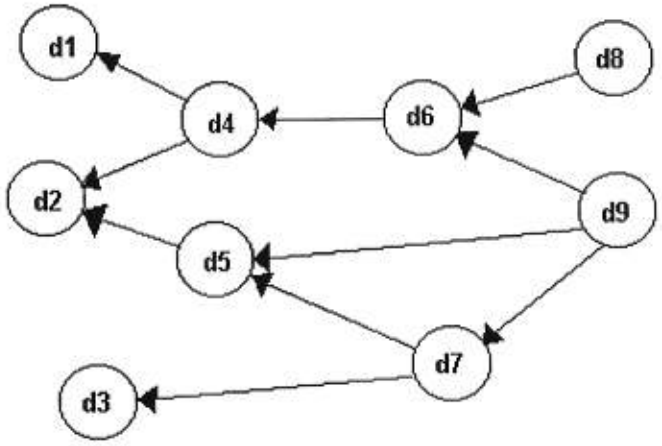

Figure 2 Graphe de liens de références

Avant de s'intéresser à l'utilisation des liens bibliographiques dans un système de recherche d'information, il est bon de s'interroger sur le rôle que jouent les items bibliographiques dans les corpus documentaires. En effet, à première vue, il n'est pas évident qu'ils peuvent servir de descripteurs du contenu des documents. En général, nous ne pouvons affirmer connaître le contenu d'un document alors que nous avons uniquement consulté les documents qu'il cite ou ceux qui le citent. Considérons, par exemple, le cas extrême d'un corpus où les items bibliographiques, inclus un peu partout dans les textes, ont été choisis au hasard, sans aucun but particulier autre que d'avoir des textes ayant des références. Il nous semble qu'un tel corpus se prête mal à un système de recherche d'information basé sur ces références « fantaisistes ». À l'opposé, si nous considérons le cas d'un corpus où leur utilisation est régie par des règles strictes, toujours respectées, comme par exemple « indiquer tous les travaux précédents aboutissant à la

recherche actuelle », ces items bibliographiques paraissent assez prometteurs pour alimenter un système de recherche d'information. Intuitivement donc, il nous semble que le rôle plus souvent implicite qu'explicite tenu par les items bibliographiques détermine le succès de leur utilisation comme descripteurs de contenu pour des fins de repérage. Dans les prochains paragraphes nous examinerons certaines facettes de liens bibliographiques à prendre en compte lorsque nous nous intéressons à leur utilisation dans la recherche d'information.

1.1.1 Le rôle des items bibliographiques dans la littérature scientifique

Plusieurs études ont été faites sur le rôle joué par les items bibliographiques dans la littérature scientifique et les chercheurs ne semblent pas être d'accord sur la question. Nous pouvons observer deux tendances. La première affirme qu'ils sont utilisés dans un cadre bien précis et jouent un rôle spécifique, c'est l'hypothèse normative. La seconde prétend, au contraire, qu'un auteur peut avoir différents motifs pour citer et que ceux-ci, et par conséquent le rôle des items bibliographiques, ne sont pas encore bien connus. Regardons en détails le point de vue de ces deux écoles.

En général, les chercheurs de l'école normative associent l'utilisation des items bibliographiques à la nature cumulative de la science. Dans l'une des premières études faites sur la question, Price établit que chaque résultat scientifique est basé sur les travaux précédents entrepris sur le même sujet, et que depuis 1850, les chercheurs ont pris l'habitude d'inclure systématiquement des références dans leurs textes afin d'indiquer aux lecteurs les points de départ de leur recherche [Price 86]. Les études de Merton [Merton 73] et de Kuhn [Kuhn 73] appuient les idées de Price. Elles suggèrent, entre autres, que la

reconnaissance du travail d'un scientifique se fait principalement à travers les citations subséquentes qu'il engendre.

Au-delà la structure cumulative de la science, Kaplan dans son étude sur les normes de l'utilisation des items bibliographiques estime que ces derniers constituent un mécanisme de contrôle au sein du monde scientifique. Ce mécanisme contribue à établir la propriété individuelle des auteurs sur leurs publications [Kaplan 65]. Ravetz, de son côté, introduit les concepts de propriété intellectuelle et de droit à la propriété intellectuelle basés sur l'interprétation que les processus de publications et de références combinent la récompense et la reconnaissance du travail d'un chercheur [Ravetz 71].

Mitra résume le rôle des items bibliographiques dans la littérature scientifique en affirmant qu'ils témoignent du progrès cumulatif de la science, qu'ils permettent la reconnaissance scientifique ainsi que l'établissement des droits des auteurs relativement à la contribution scientifique qu'ils apportent [Mitra 70]. Mitra relève encore plusieurs autres facettes du phénomène. Il affirme que les items bibliographiques constituent une importante source d'informations pour un chercheur car ils aident à identifier les sources d'informations des scientifiques et, par conséquent, qu'ils indiquent la littérature indispensable à la compréhension de leurs travaux.

Mitra ne se trompe pas quant au rôle social des citations. En effet, l'émergence de l'analyse des citations comme champs d'études vient appuyer le fait que les citations jouent un rôle dans la reconnaissance scientifique. Entre autre, les praticiens de ce domaine élaborent différentes mesures de la qualité des travaux scientifiques basées sur le nombre de citations qu'ils engendrent. Partant de là, différentes études utilisant l'analyse des citations sont régulièrement entreprises pour évaluer l'influence de publications, de

revues scientifiques ou d'auteurs. Dans son travail de synthèse prônant l'analyse des citations [Narin 76], Narin passe en revue une vingtaine d'études montrant que le nombre de citations engendrées et d'autres mesures bibliométriques correspondent à différents niveaux de distinction. L'apparition des « Science Citation Index » et « Journal Citation Report », illustre l'épanouissement de l'analyse de citations ainsi que les études qui en découlent. Les outils de ce type sont régulièrement utilisés comme outils d'évaluation ou mesure d'influence et de productivité.

Dans leur étude pour valider l'analyse de citations, Lawani et Bayer ont trouvé que les publications valables sont citées fréquemment durant les cinq premières années suivant leur parution [Lawani & Bayer 83]. Leur étude montre clairement que l'estime des contemporains et le nombre de citations engendrées sont très reliés. Il n'y a pas de doute que l'analyse de citations est devenue une méthode très utilisée comme mesure de qualité. L'hypothèse implicite qui la fonde est cependant questionnée.

Il en va ainsi pour Gilbert qui propose une théorie alternative de l'utilisation des items bibliographiques. Il considère les publications comme étant des outils de persuasion; un auteur ayant obtenu un résultat qu'il considère important doit persuader la communauté scientifique de partager son opinion sur la valeur de son travail [Gilbert 77]. De ce point de vue le calcul du nombre de citations engendrées et les différents procédés d'analyse de citations font fi de la raison pour laquelle les publications sont citées. Dans sa synthèse sur l'analyse de citations, Smith conclut avec prudence que l'on connaît peu la façon de procéder de l'auteur qui cite [Smith 81]. Martyn, de son côté, met l'emphase sur l'aspect subjectif et personnel du processus de citation [Martyn 75]. Il semble, que l'accessibilité physique des documents sources influence ce processus. Kochen observe

qu'il n'est pas surprenant que l'arbitraire prime dans la sélection des références chez les auteurs et que, sans doute, des documents qui auraient dû être cités, ne le sont pas, alors que d'autres qui le sont, ne sont pas pertinents [Kochen 74]. Certains critiques vont même plus loin et accusent les auteurs de fraude en ce qui à trait aux références à d'autres publications. Broadus pointe un document citant à tort un travail précédent et ses données semblent appuyer l'idée que les auteurs tirent sans les consulter des références à partir de la bibliographie d'autres publications[Broadus 83].

Garfield [Garfield 79] met en garde contre une utilisation naïve de l'analyse des citations afin d'établir la valeur d'une publication ou d'un chercheur en offrant quinze raisons pour lesquelles un auteur peut citer :

1. Rendre hommage aux pionniers ;
2. Donner crédit pour des travaux courants (hommage aux contemporains) ;
3. Indiquer une méthodologie, préciser l'équipement ;
4. Identifier des lectures de mise en contexte ;
5. Corriger ses erreurs ;
6. Corriger les erreurs des autres ;
7. Critiquer des travaux précédents ;
8. Appuyer des affirmations ;
9. Indiquer d'autres résultats imminents ;
10. Fournir des liens vers des travaux peu répandus, peu indexés ou non cités ;

11. Établir l'authenticité des données et des classes de faits, constantes physiques, etc.;
12. Indiquer les premières publications dans lesquelles une idée ou un concept a germé ;
13. Indiquer les premières publications ou d'autres travaux décrivant des termes ou concepts éponymes ;
14. Rejeter les travaux ou idées des autres ;
15. Contester les droits d'auteurs des autres.

Avec une liste aussi impressionnante, il est clair que des données quantitatives brutes ne suffisent pas à elles seules pour établir la valeur ou la qualité d'une publication ou d'un chercheur. De plus, même si la liste est une bonne indication des raisons pour lesquelles un auteur cite, elle ne donne aucun indice sur la motivation personnelle ni sur la façon de procéder de ce dernier lorsqu'il s'agit de construire une liste de références. Ce sont encore deux facteurs qui, selon les critiques de l'analyse des citations, viennent semer le doute sur les résultats obtenus par ce procédé.

Plusieurs autres études s'étendent d'avantage sur le sujet, chacune insistant sur un aspect qui témoigne de la complexité de la pratique des citations. Nous n'allons pas nous attarder plus longtemps sur le sujet, car il est assez vaste pour constituer à lui seul un mémoire de maîtrise. Toutefois, nous référons le lecteur intéressé à l'article de Liu [Liu 93] qui est une revue assez complète des études faites sur les citations.

Comment donc prétendre construire un système de recherche d'information basé sur les références et les citations, alors que, selon la littérature, celles-ci sont déterminées

par un grand nombre de variables que nous n'arrivons pas à bien saisir, voir à maîtriser ? Deux remarques s'imposent.

Tout d'abord, si les critiques désapprouvent l'utilisation des données bibliographiques brutes, tel le nombre de citations engendrées, ils ne disent pas pour autant que les liens bibliographiques ne sont d'aucune utilité. Bon nombre d'études s'entend sur le fait que chaque item bibliographique reflète la décision de l'auteur d'attirer l'attention sur un travail qu'il considère pertinent au sujet traité. Dans ce sens, les liens induits par les items bibliographiques continuent à garder leur intérêt pour un système de repérage. En fin de compte, les études sur la complexité des citations ne font que soulever des points qui, lorsque pris en compte, améliorent les outils existants déjà. Déjà en 1965, Lipetz cherchait à améliorer la sélection dans les index de citations en tenant compte du contexte et du contenu des items bibliographiques.

Ensuite, s'il est vrai qu'un auteur a plusieurs raisons pour inclure un item bibliographique dans son texte, il n'en existe pas non plus une infinité. En effet, certains chercheurs ont réussi à classer les items bibliographiques dans diverses catégories et ce, même si ces catégories diffèrent un peu dépendant du type de l'étude et du type de corpus sur lequel elle a porté. De plus, si l'on considère le cas d'une collection de textes écrits dans un contexte particulier et relativement homogène, il est possible d'énumérer plus ou moins facilement les facteurs déterminant le rôle des items bibliographiques.

Tout au long de ce mémoire, nous apparenterons le concept de contexte à celui de corpus, groupe de textes auquel appartient un texte particulier. Si un corpus est hétérogène, c'est-à-dire s'il est constitué de textes venant de différents types de littérature : sciences naturelles, science sociales, religieuses, etc., si les textes qui le

composent ont été écrits par des auteurs de différents niveaux, visant différents types de lectorat, il est alors naturel de penser que l'énumération des facteurs influençant les items bibliographiques sera difficile, car la liste de ces facteurs sera longue et contiendra beaucoup de cas particuliers. À l'opposé, si un corpus est homogène, nous pouvons nous attendre à ce que les items bibliographiques puissent être facilement classés en un nombre restreint de catégories. Nous pensons que les corpus juridiques, spécialement les corpus jurisprudentiels, sont de bons exemples de corpus homogènes et qu'ils se prêtent bien au genre de système que nous voulons construire. La jurisprudence de la Cour suprême du Canada illustre bien ce type de situation. Voyons le rôle tenu par les items bibliographiques dans de tels corpus.

1.1.2 Le rôle des items bibliographiques dans les corpus juridiques

Les études sur le rôle des items bibliographiques dans la littérature juridique sont plus rares que celles menées pour la littérature scientifique. Pourtant, ils tiennent un rôle important dans le milieu juridique, spécialement dans le cas de ceux s'apparentant à la tradition de la *common law*, comme les systèmes canadien, américain et anglais. Dans ces systèmes, le respect du « stare decisis » impose, entre autre, que les cas qui se ressemblent soient décidés de la même façon [Waddams 92]. Un tel principe reflète la volonté que la justice soit stable et équitable. Aussi, fidèles à ce principe, les tribunaux ont pris l'habitude de référer à leurs décisions passées. Celles-ci constituent la jurisprudence d'un tribunal. Lorsqu'un tribunal décide d'un cas, il doit donner les raisons pour lesquelles il décide de la sorte pour respecter un autre principe, celui du *ratio dicidendi* [Kiralfi 90]. Ces raisons constituent des « précédents » que les autres tribunaux de même niveau doivent considérer pour garder une certaine cohérence et que les

tribunaux de niveau de juridiction inférieure sont obligés de suivre. En pratique, ce sont les avocats des parties concernées qui ont la responsabilité de présenter au juge toute la jurisprudence pertinente afin qu'il puisse prendre une décision. Par la suite, ce dernier indique les raisons qu'il retient en citant la jurisprudence appropriée. Remarquons que, quelques fois, les juges utilisent les items bibliographiques à des fins rhétoriques. Cependant, en général, les références juridiques jouent le rôle informationnel que nous venons de décrire.

Dans un tel contexte, il n'est pas surprenant que les références à la jurisprudence aient un rôle capital. Ce rôle est depuis longtemps reconnu. En effet, de façon plus importante que dans la littérature scientifique, des outils de recherche de citations se sont développés afin de permettre aux avocats de retrouver les précédents qui font autorité en regard des causes qu'ils ont à plaider. Il n'est pas étonnant non plus que les étudiants en droit soient obligés de suivre quelques cours où leur sont présentées les différentes sources de jurisprudence existantes et où leur sont expliqués les outils de recherche juridique leur en facilitant l'accès.

1.2 Les avantages d'un système de recherche d'information basé sur les liens bibliographiques

Nous avons vu donc que les items bibliographiques jouent un rôle important et fort spécifique dans le milieu juridique. De façon plus précise, les juges les utilisent dans leurs jugements pour indiquer les précédents sur lesquels ils s'appuient. C'est ce rôle primordial et spécifique qui favorise leur utilisation dans un système de repérage de la jurisprudence. Cependant, même lorsque nous considérons un contexte plus général, nous constatons que les liens bibliographiques induits par les références dans un corpus

possèdent des caractéristiques qui les avantagent par rapport à l'utilisation des termes comme descripteurs du contenu des documents.

1.2.1 La relation entre référence et décision

Rappelons que l'un des désavantages de l'utilisation des termes pour décrire le contenu des documents vient du fait qu'un concept n'est pas représenté de façon unique par un terme. Les synonymes, les variantes grammaticales et les homographes sont des phénomènes qui nuisent autant aux concepteurs de systèmes de recherche d'information qu'aux utilisateurs de ces systèmes. Les items bibliographiques sont plus avantageux à ces égards.

Considérons tout d'abord la synonymie. En droit, la synonymie résulte de la coexistence de plusieurs normes pour construire les items bibliographiques. Dans le milieu juridique canadien par exemple il existe essentiellement deux normes pour désigner les références jurisprudentielles : la norme de Lluelles[Lluelles 95] et celle de McGill[McGill 92]. Il existe donc deux chaînes de caractères différentes pour référer à un même document juridique. Toutefois, en général, un rédacteur choisit une norme et y adhère de façon stricte. Nous retrouvons donc rarement l'utilisation des deux normes dans un même document. Au surplus, étant donné que les différences entre les normes sont invariables, il est possible de construire des tables de conversion permettant d'obtenir les références dans le style voulu. Un autre cas plus subtil de synonymie vient du fait que les normes permettent l'utilisation d'abréviations du style « op. cit. » pour des références précédemment mentionnées dans le texte. Là encore, étant donné que les règles d'abréviation sont fixes, il est plus ou moins facile d'établir un mécanisme permettant de retrouver les références complètes.

En ce qui à trait aux variantes grammaticales qui entravent le fonctionnement des systèmes d'information basés sur les termes, il est clair que ce phénomène ne s'applique pas aux items bibliographiques. Il n'est donc pas nécessaire d'avoir recours à des algorithmes de standardisation comme c'est le cas si l'on utilise des termes. Remarquons également que même s'il faut parfois transformer les références pour régler les problèmes de synonymie, les transformations que l'on effectue n'introduisent pas des ambiguïtés telles que celles causées, par exemple, par la « lemmatisation » des termes.

Finalement, les items bibliographiques n'ont jamais des homographes au même sens que les termes. Chaque item bibliographique est exclusif et désigne toujours la même décision. Cependant, comme nous le verrons plus loin, il se peut qu'une décision contienne l'exposé de plusieurs principes de droit et que, pour cette raison, elle soit référée dans des jugements subséquents traitant de sujets totalement différents.

1.2.2 La disponibilité et le format standard des items bibliographiques

Les références sont largement disponibles dans les décisions sous un format standard. Étant donné l'importance de la jurisprudence dans le milieu juridique, des normes pour la construction des références ont été élaborées et sont scrupuleusement respectées. Aujourd'hui encore, des propositions de normes de construction d'items bibliographiques plus universels et compacts sont étudiées par les autorités concernées.

Les items bibliographiques auxquels nous allons nous intéresser pour ce projet ont la forme suivante :

R. c. Bernshaw, [1995] 1 R. C. S. 254

Cette référence contient deux parties distinctes. La première partie, l'intitulé, *R. c. Bernshaw*, est généralement construite à partir du nom des parties impliquées. Dans ce cas ci, il s'agit d'une affaire impliquant un certain Bernshaw et la Reine, en d'autres termes, le procureur de la Couronne. La deuxième partie de la référence donne des indications sur la façon dont le jugement a été répertorié. Dans ce cas ci, nous voyons que la décision se trouve dans le premier volume du Recueil de la Cour suprême (R. C. S.) de 1995 à la page 254. Remarquons que la décision est entièrement identifiée par la deuxième partie de la référence. En effet, si nous connaissons cette deuxième partie pour une décision, nous pouvons aller consulter le recueil et la retrouver. Ceci n'est pas vrai pour l'intitulé, car il se peut fort bien que les mêmes parties soient impliquées dans plusieurs jugements différents ou que différentes parties ayant les mêmes noms soient impliquées dans des jugements distincts. De plus, le nom des parties ressemblant à du texte libre, il est plus difficile de le distinguer du texte de la décision. Pour toutes ces raisons, nous avons décidé de n'utiliser que la deuxième partie des items bibliographiques jurisprudentiels. Ce faisant, nous tirons également avantage du fait qu'il est très facile de construire des expressions régulières qui repéreront nos références à l'intérieure du texte des décisions.

1.2.3 Les caractéristiques pondérables

Un autre avantage attribuable à l'utilisation des items bibliographiques plutôt que des termes tient au fait que les premiers possèdent plus de caractéristiques pondérables que les derniers. En effet, dans le cas des termes, l'unique critère quantifiable de manière objective est la fréquence d'occurrence. Un terme qui est répété plusieurs fois à l'intérieur d'un document peut être considéré plus important pour ce document qu'un terme qui y

apparaît rarement. Bien sûr, ce schéma de pondération peut être amélioré en considérant la fréquence des termes dans le corpus en général. Ainsi, un terme usuel, souvent utilisé dans le corpus, est considéré moins important par rapport à un autre qui n'apparaît que sporadiquement.

La fréquence d'occurrence est également un critère pondérable important pour les items bibliographiques. Si une décision réfère fréquemment une autre, il est fort probable que cette dernière est importante pour la cause en question. Il faut avouer cependant, que ceci soulève un problème de reconnaissance différent de celui associé au décompte des termes. Savoir si un mot apparaît dans un texte et, le cas échéant, son nombre d'occurrences ne pose pas de difficulté majeure. Bien qu'il soit aussi facile de déterminer la présence d'items bibliographiques dans un texte, le cas des références implicites vient compliquer le calcul de leur fréquence d'occurrence. Supposons qu'un juge discute énormément d'une décision tout au long de son jugement et qu'à un moment donné, parce qu'il estime que ce n'est plus nécessaire, il arrête de la citer de façon explicite. Il est clair que cette référence est plus importante pour le jugement en question qu'une autre décision à laquelle le juge réfère en deux ou trois fois, peut être juste pour faire une remarque. Ce genre de références implicites se présente sous deux formes. Le juge peut soit décider d'utiliser des formules abrégées telles que « dans l'arrêt *Bernshaw* » ou bien ne pas mentionner la référence en question. Dans le premier cas, il s'agit de développer des algorithmes qui vont repérer ces items bibliographiques partiels et les associer aux références. De tels algorithmes sont réalisables parce qu'en général, les items bibliographiques partiels sont précédés d'au moins une instance sous sa forme explicite.

Dans le second cas, il n'existe pas un moyen permettant de contourner le problème. Il s'agit d'une situation où la fréquence des items bibliographiques est moins utile.

Les items bibliographiques possèdent d'autres caractéristiques pondérables qui ne posent pas autant de problèmes. L'une des plus évidentes est l'âge, c'est-à-dire l'intervalle de temps écoulé entre la décision citée et la décision qui cite (respectivement référence et citation). Une référence plus éloignée dans le temps est considérée être plus importante pour la décision qui cite qu'une référence proche car d'habitude les juges font référence à des décisions récentes et par conséquent s'ils prennent la peine de citer une vieille décision c'est que celle-ci est très importante pour la cause en question.

Le niveau de juridiction est également une caractéristique pondérable intéressante dans un contexte juridique. Si nous voulons tenir compte de l'autorité des références sur les décisions, alors nous attribuerons une plus grande importance aux références vers les instances judiciaires les plus élevées.

Nous pouvons également vouloir considérer l'éloignement géographique entre les deux décisions. En effet, si un tribunal au Canada prend la peine de citer un jugement rendu en Angleterre, c'est sûrement parce que ce jugement-là est très pertinent pour la cause qu'il est en train d'étudier.

D'autres caractéristiques sont plus controversées. Il arrive parfois que dans un jugement on accorde différents niveaux d'accord ou de désaccord aux décisions référées. Les juges font la différence entre les décisions qui sont mentionnées, suivies, approuvées ou rejetées et l'indiquent parfois de façon explicite dans leurs jugements. Cependant, il n'est pas clair laquelle de ces catégories est plus importante pour le jugement en cause.

De plus, il est possible également qu'une référence appartienne à plus d'une catégorie si, par exemple, les juges qui ont entendu la cause n'ont pas la même opinion sur sa nature.

1.2.4 La familiarité du repérage de citations et de références

Les méthodes de recherche bibliographique sont très intuitives et, en général, la plupart des gens y sont familiers. Les juristes ne sont pas des exceptions et la méthode de représentation d'une décision par les liens bibliographiques qu'elle établit avec les autres décisions du corpus correspond à une approche intuitive de la recherche juridique. Les outils qu'ils apprennent à utiliser dans les premières années de leur carrière sont basés sur ces principes. En effet, une fois qu'une personne a fini de consulter un document initial, intéressant ou pertinent à son besoin d'information, elle a le choix entre deux stratégies de recherche pour trouver d'autres documents.

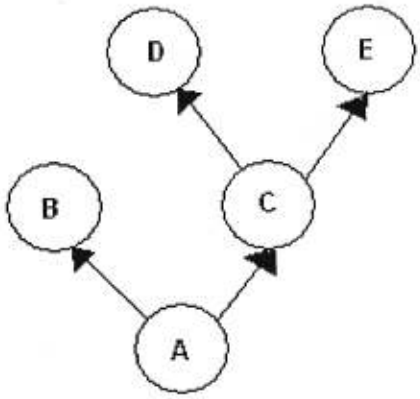

Figure 3 Exemple de recherche de références

Tout d'abord, elle peut effectuer une recherche de références. Cette stratégie consiste à consulter la liste des références du document initial pour aller consulter les documents auxquels il réfère. La "Figure 3" donne un exemple de ce type de recherche.

Le document initial A permet de trouver les documents B et C. Puis, en supposant que le document C est lui aussi intéressant, il servira de point de départ pour accéder aux documents D et E. En général, il faut s'attendre à ce que les documents fournis par cette méthode soient plus vieux que le document initial. Cette approche semble recommandée lorsque l'on veut remonter aux origines d'un sujet particulier.

La deuxième stratégie, la recherche de citations, permet de retrouver des documents plus récents. Elle s'impose lorsque l'on souhaite étudier l'évolution ultérieure d'un sujet dans le temps. En général, c'est la stratégie la plus utilisée parce qu'elle permet d'obtenir de l'information plus à jour. Toutefois, il fallait s'y attendre, si cette méthode est préférable à la première, elle exige plus de travail. La "Figure 4" illustre une recherche de citations. Le document A est utilisé comme document initial et en consultant le corpus, on trouve que les documents B et C le citent. Encore une fois, ces documents peuvent être utilisés comme nouveaux points de départ et générer d'autres documents D, E et F encore plus récents. Remarquons que, dans la pratique, nous ne sommes pas obligés de consulter les documents du corpus afin de découvrir ceux qui réfèrent au document initial. Il existe des ouvrages spécialisés, appelés index de citations, qui énumèrent les citations de différents documents initiaux. Au cours des années, différents types d'index de citations plus ou moins spécialisés ont vu le jour : Science Citation Index, Compu-Math Citation Index... Le Recueil de jurisprudence citée (R.J.C.) est un exemple d'index de citation spécialisé pour la jurisprudence des tribunaux canadiens. Finalement, disons qu'il est également possible de mélanger les deux stratégies afin d'obtenir une plus grande couverture du sujet.

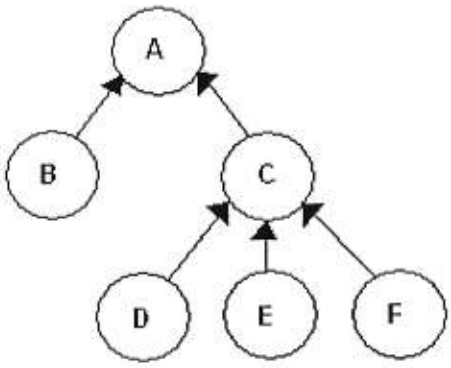

Figure 4 Exemple de recherche de citations

1.3 Liens bibliographiques et similarité

Dans la section précédente, nous avons énuméré les caractéristiques qui nous poussent à croire que les items bibliographiques sont de bons candidats pour être utilisés comme descripteurs de documents dans un système de repérage d'information. Cet intérêt des items bibliographiques et des liens qu'ils établissent a donné lieu à divers travaux. Des auteurs ont tenté d'exploiter l'information qu'ils constituent en utilisant des graphes flous, le couplage bibliographique, la cocitation ou des vecteurs de références.

Notre projet ne consitue pas la première tentative d'inclure les items bibliographiques dans un système de recherche d'information. En effet, l'idée n'est pas complètement nouvelle. Salton, dans son « Introduction to Modern Information Retrieval », reconnaît qu'il est possible de constituer un vecteur descripteur de documents constitué de mots clés et des informations bibliographiques tels que liens de citation réliés aux documents en question. Cependant, il ne s'attarde pas sur le sujet[Salton 83]. Savoy dans son étude sur la combinaison de multiples sources d'évidence de similarité, a indiqué que la prise en compte des liens bibliographiques suggérés par les items

bibliographiques augmentait la précision de son système [Savoy 96]. Toutefois, dans son système, il fallait ajouter manuellement ces informations. Une approche plus systématique fut tentée par Nomoto et ses collaborateurs.

1.3.1 L'utilisation de graphes flous pour exprimer les liens bibliographiques

En supposant que dans le contexte des sciences et de la technologie chaque document représente une partie de la recherche et que les items bibliographiques qu'il contient indiquent les résultats d'autres documents sur lequel il s'appuie, Nomoto et ses collaborateurs utilisent le réseau de citations engendré pour bâtir un système de repérage se basant sur les graphes flous [Nomoto 90]. Ce système est simple d'utilisation et modélise parfaitement les stratégies de recherche bibliographiques susmentionnées. L'utilisateur choisit son document initial et demande au système de récupérer soit les documents qui ont engendré le document initial (recherche de références), soit ceux qui s'appuient sur lui (recherche de citations). Le système répond alors en transformant le réseau de citations en un ensemble flou qui reflète, dans le cas de la recherche de références, le degré avec lequel les documents du corpus sont référés par le document initial, ou, dans le cas de la recherche de citations, le degré avec lequel ils citent le document initial.

Un des aspects intéressants dans ce système tient à la façon dont sont construits les ensembles flous. À ces ensembles sont associées des fonctions d'appartenance qui reflètent le degré avec lequel les documents du corpus sont reliés au document initial. À quelques nuances près, la fonction d'appartenance associée à l'ensemble des références est similaire à celle associée à l'ensemble des citations et donc seule la première sera décrite ici.

Prenons le cas d'un corpus constitué de N documents : $d_1, d_2, \cdots, d_N$. Dénotons d_r le document initial de la requête. Nomoto considère que d_m est cité en m étapes par d_r s'il existe *m-1* documents, disons $d_1, d_2, \cdots, d_{m-1}$, tels que d_r cite d_1, d_1 cite $d_2, \ldots, d_{m-1}$ cite d_m. Cette approche transitive permet de transformer un graphe de liens de citations en un ensemble flou et permet d'établir une relation entre deux documents même sans un lien direct de l'un à l'autre. La relation en m étapes énoncée plus haut est représentée par la valeur $\mu_m^R(d_i, d_r)$qui exprime le degré avec lequel d_i est cité par d_r en m étapes. Ici, l'exposant R sert à indiquer que ces valeurs sont utilisées pour les recherches de références. Encore une fois, ces valeurs sont à distinguer de celles utilisées pour la recherche de citations dénotées $\mu_m^C(d_r, d_i)$. Les valeurs $\mu_1^R(.,.)$sont attribuées manuellement lors de la phase initiale d'enregistrement des documents de telle sorte que :

- $\mu_1^R(d_i, d_j) = 0$, si d_i n'est pas une référence de d_j
- $\mu_1^R(d_i, d_j) = a$ (où a est une constante définie arbitrairement entre 0 et 1), sinon

Une fois ces valeurs enregistrées, les relations en m étapes sont calculées de la façon suivante :

$$\mu_m^R(d_i, d_r) = \mu_1^R(d_i, d_1) \bullet \mu_{m-1}^R(d_1, d_r) \oplus \cdots \oplus \mu_1^R(d_i, d_N) \bullet \mu_{m-1}^R(d_N, d_r)$$

où $\bullet$ et $\oplus$ sont les opérations algébriques classiques définies sur les ensembles flous :

$x \bullet y = xy$ et

$x \oplus y = x + y - xy$

Remarquons que nous pouvons facilement vérifier que les opérations algébriques définies plus haut sont associatives. Par conséquent l'ordre d'évaluation de la somme dans la définition de $\mu_m^R(d_i, d_r)$n'a pas d'importance. Également d'un point de vue théorique, la valeur de a n'a pas d'importance. Cependant, en pratique, cette valeur peut être choisie de façon judicieuse afin de donner les meilleurs résultats possibles. Dans son article[Nomoto 90], Nomoto discute de différentes alternatives pour le choix d'une valeur adéquate.

Pour comprendre ce qui se passe derrière cette formule, considérons les relations en deux étapes. Deux cas peuvent se présenter. Considérons d'abord le cas des références en série illustrées par la "Figure 5". Étant donné que les valeurs $\mu_1^R(.,.)$ sont comprises entre 0 et 1, la relation en deux étapes entre d_i et d_k $(\mu_2^R(d_i, d_k))$est définie de telle sorte qu'elle ne soit pas plus forte qu'aucune des relations en une étape $\mu_1^R(d_i, d_j)$ ou $\mu_1^R(d_j, d_k)$.

$$\mu_2^R(d_i, d_k) = \mu_1^R(d_i, d_j) \bullet \mu_1^R(d_j, d_k).$$

Remarquons que ceci est le résultat souhaité, puisque d_k ne cite pas d_i directement.

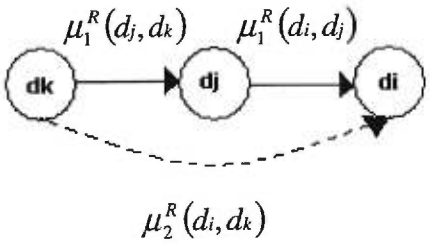

Figure 5 Références en séries

Le deuxième cas pouvant se présenter est lorsque plusieurs « chemins » relient les deux documents. La "Figure 6" montre un cas de références parallèles où d_i et d_l sont

reliés par deux références en série différentes. Dans un tel cas, la relation en 2 étapes est définie de telle sorte que sa valeur ne soit pas inférieure à l'une ou l'autre de celle des références en séries.

$$\mu_2^R(d_i,d_l)=\mu_1^R(d_i,dj)\bullet\mu_1^R(d_j,d_l)\oplus\mu_1^R(d_i,d_k)\bullet\mu_1^R(d_k,d_l)$$

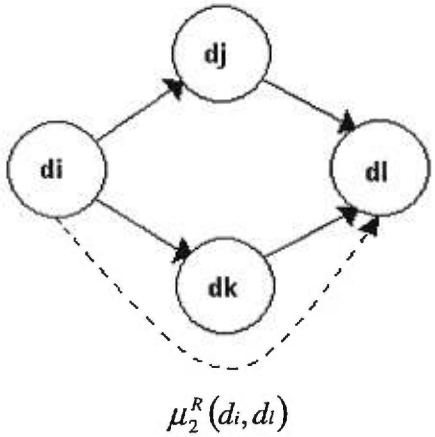

Figure 6 Références en parallèle

Encore une fois, c'est le résultat souhaité. Puisqu'il y a deux chemins reliant les deux documents, la relation entre eux doit être plus forte que s'il n'y en avait qu'un seul.

Finalement, Nomoto définit ce qu'il appelle la relation synthèse évaluant le degré de relation en au plus M étapes entre deux documents :

$$\overline{\mu}_M^R(d_i,d_j)=\mu_1^R(d_i,d_j)\oplus\mu_2^R(d_i,d_j)\oplus\cdots\oplus\mu_M^R(d_i,d_j)$$

C'est cette relation qui sert à définir la fonction d'appartenance de l'ensemble flou des documents référés en au plus M étapes par le document initial. Cet ensemble est défini de la manière suivante :

$\Gamma_M^R(d_r)=\left\{\overline{\mu}_M^R(d_i,d_r)/d_i : i\in[1,N]\right\}$, où M est un paramètre du système choisi de telle façon à obtenir le meilleur compromis entre les temps de calcul et la qualité des

résultats. L'ensemble des documents cités par le document initial, dénoté $\Gamma_M^C(d_r)$, est construit de façon similaire, la seule différence étant dans l'ordre du calcul des relations en M étapes. Sans revenir sur les mêmes détails du calcul des relations en M étapes pour la recherche de références vu plus haut, nous donnons ici la formule pour le cas de la recherche des citations :

$$\mu_m^C(d_r,d_i)=\mu_{m-1}^C(d_r,d_1)\bullet\mu_1^C(d_1,d_i)\oplus\cdots\oplus\mu_{m-1}^C(d_r,d_N)\bullet\mu_1^C(d_N,d_i)$$

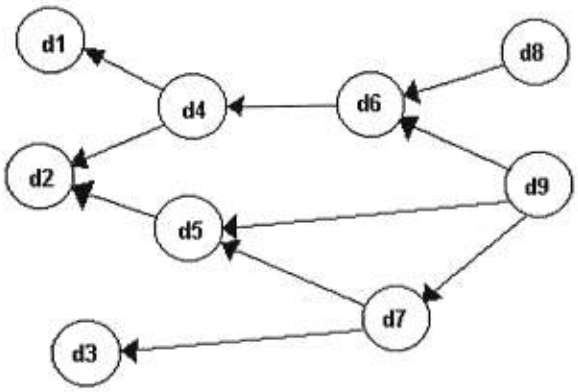

Figure 7 Graphe de liens de références

Illustrons un peu la méthode. Considérons le graphe de liens de références de la "Figure 7". Trouvons le sous-ensemble flou généré par une recherche de références à partir du document d_9. Nous avons donc N=9, $d_r=d_9$, et nous choisissons M=3. Pour la simplicité de l'exemple, nous ne considérons que les relations non nulles. Également, nous avons volontairement laissé tomber l'exposant R afin de ne pas alourdir le texte. Nous avons tout d'abord les relations suivantes :

$\mu_1(d_5,d_9) = \mu_1(d_6,d_9) = \mu_1(d_7,d_9) = a$, où a est une constante choisi à l'avance.

Les relations en 2 étapes sont alors données par :

$\mu_2(d_2,d_9) = \mu_1(d_2,d_5) \,.\, \mu_1(d_5,d_9) = a \,.\, a = a^2$

$\mu_2(d_3,d_9) = \mu_1(d_2,d_7) \,.\, \mu_1(d_7,d_9) = a \,.\, a = a^2$

$\mu_2(d_4,d_9) = \mu_1(d_4,d_6) \,.\, \mu_1(d_6,d_9) = a \,.\, a = a^2$

$\mu_2(d_5,d_9) = \mu_1(d_5,d_7) \,.\, \mu_1(d_7,d_9) = a \,.\, a = a^2$

Les relations en trois étapes sont données par

$\mu_3(d_1,d_9) = \mu_1(d_1,d_4) \,.\, \mu_2(d_4,d_9) = a \,.\, a^2 = a^3$

$\mu_3(d_2,d_9) = \mu_1(d_2,d_4) \,.\, \mu_2(d_4,d_9) \oplus \mu_1(d_2,d_5) \,.\, \mu_2(d_5,d_9)$

$= a \,.\, a^2 \oplus a \,.\, a^2 = 2a^3 - a^6$

Finalement les relations synthèses sont :

$$\overline{\mu}_3(d_1,d_9) = \mu_3(d_1,d_9) = a^3$$

$$\overline{\mu}_3(d_2,d_9) = \mu_2(d_2,d_9) \oplus \mu_3(d_2,d_9) = a^2 \oplus (2a^3 - a^6) = a^2 + 2a^3 - 2a^5 - a^6 + a^8$$

$$\overline{\mu}_3(d_3,d_9) = \mu_2(d_3,d_9) = a^2$$

$$\overline{\mu}_3(d_4,d_9) = \mu_2(d_4,d_9) = a^2$$

$$\overline{\mu}_3(d_5,d_9) = \mu_1(d_5,d_9) \oplus \mu_2(d_5,d_9) = a \oplus a^2 = a + a^2 - a^3$$

$$\overline{\mu}_3(d_6,d_9) = \mu_1(d_6,d_9) = a$$

$$\overline{\mu}_3(d_7,d_9) = \mu_1(d_7,d_9) = a$$

$$\overline{\mu}_3(d_8,d_9) = \overline{\mu}_3(d_9,d_9) = 0$$

En choisissant a = 0.5, on obtient la fonction d'appartenance du sous-ensemble flou résultant :

$\Gamma^R{}_3(d_9) = \{ 0.125/d_1 \,,\, 0.426/d_2 \,,\, 0.250/d_3 \,,\, 0.250/d_4 \,,\, 0.625/d_5 \,,\, 0.500/d_6 \,,\, 0.500/d_7,$

$0/d_8 \,,\, 0/d_9 \}$

Remarquons que le système a trouvé que d_5 est le document le plus relié à d_9. Ce résultat n'est pas étonnant puisqu'en consultant le graphe, nous constatons que d_5 est le seul document à être relié à d_9 par deux chemins : d_9 le cite directement et également à travers d_7.

Ce système a un avantage indéniable, son utilisation est très simple. Il modélise parfaitement les stratégies de recherche bibliographique que nous avons déjà mentionnées et, avec lesquelles la majorité des gens sont familiers. De plus, dans la génération de ses résultats, il considère tous les documents du corpus, ce qu'un humain ne pourrait pas accomplir pour la plupart des corpus à cause du nombre élevé de références avec lequel il faudrait travailler.

Pourtant, une tache persiste dans le tableau. Avant de pouvoir utiliser le système, un utilisateur doit enregistrer manuellement les informations à propos des documents, c'est-à-dire attribuer une valeur initiale aux différentes relations en une étape entre les documents. Quoique Nomoto maintienne que ce traitement initial s'accomplit aisément – dans son article il mentionne qu'un utilisateur sans aucune connaissance préalable d'un corpus a réussi à enregistrer les 606 documents qui le constituaient en trois semaines – il n'en demeure pas moins vrai que de nos jours, nous nous attendons à ce que les systèmes soient capables d'accomplir eux-mêmes leurs tâches d'initialisation.

De plus, il ne s'agit pas d'un système de recherche d'information à proprement parler. Comme l'indique l'hypothèse de départ des auteurs, c'est un système permettant d'évaluer l'influence qu'exercent différents documents scientifiques entre eux. Remarquons tout de même qu'un tel système a beaucoup de potentiel, car, même si Nomoto ne semble pas s'attarder sur les détails de la complexité des liens

bibliographiques, il est possible à notre avis d'assigner aux relations initiales lors de l'enregistrement des documents des valeurs judicieusement choisies qui reflètent une classification des items bibliographiques utilisant les différentes valeurs pondérées que nous avons mentionnées au début de la section précédente. Le système deviendrait alors un outil intéressant pour l'analyse des citations. Dans un contexte juridique également, ce système pourrait connaître des applications intéressantes. Il pourrait, par exemple, permettre d'étudier quelles décisions ont joué un rôle dans la décision initiale (recherche de références) ou encore, dans quels jugements la décision initiale a été considérée (recherche de citations).

Ce n'est pas l'utilisation que nous voulons faire des références. Nous voulons construire un système capable d'évaluer la similarité du contenu des documents plutôt que leur interdépendance. Nous pensons que le système de Nomoto n'est pas apte à évaluer la similarité entre documents parce que, à la lumière des études sur l'utilisation des items bibliographiques, force nous est de constater que ces derniers, quoique suggérant un lien entre les documents, ne garantissent pas que les documents traitent du même sujet. Il nous faut donc concevoir d'autres types de mesures permettant d'évaluer de façon plus spécifique le rapprochement du contenu de différents documents. Il existe des mesures de ce type qui sont utilisées en bibliométrie : le couplage bibliographique et la cocitation.

1.3.2 Le couplage bibliographique

Deux documents sont en couplage bibliographique lorsqu'ils partagent certaines références, c'est-à-dire, lorsque leur bibliographie respective contient des items en communs. Cette mesure fut d'abord proposée par Kessler qui la définit comme étant le

nombre de références communes aux deux documents [Kessler 63]. La "Figure 8" nous donne un exemple de deux documents en couplage bibliographique. Dans ce cas, les documents ont un couplage bibliographique de 2 puisque les deux citent $d_1$et d_2. Plusieurs expériences ont montré que plus le couplage bibliographique entre deux documents est élevé, plus ces documents présentent une grande similarité de contenu.

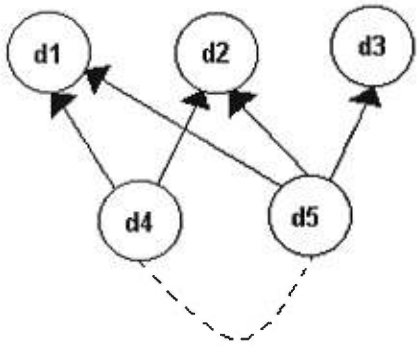

Figure 8 Couplage bibliographique

1.3.3 La cocitation

Deux documents peuvent être également reliés par un pattern commun de citations. En effet, un lien de cocitation est établi entre deux documents lorsqu'ils sont communément cités par un troisième. Cette mesure fut proposée par Small qui la définit comme étant le nombre de citations communes de deux documents [Small 73]. La "Figure 9" nous donne un exemple de cocitation. Dans ce cas, les documents d_1 et d_2 ont un lien de cocitation de 2 parce qu'ils sont cités conjointement par d_4 et d_5, alors que d_2 et d_3, ainsi que d_1 et d_3, sont reliés par une cocitation de 1. Les diverses expériences menées par Small ont indiqué

qu'en général la cocitation est une mesure de similarité plus efficace que le couplage bibliographique.

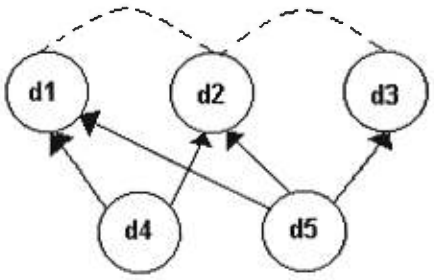

Figure 9 Cocitation

1.3.4 Le vecteur d'items bibliographiques

Tapper propose une autre utilisation des items bibliographiques [Tapper 84]. Dans son article, il décrit les expériences qu'il a menées avec des corpus juridiques américains et anglais en utilisant des vecteurs d'items bibliographiques. Comme nous pouvons nous y attendre, ces vecteurs diffèrent des vecteurs conventionnels utilisés en repérage d'information en ce que leurs composantes sont associées aux items bibliographiques contenus dans les textes indexés plutôt qu'aux termes que ces derniers contiennent. Des poids sont associés à ces descripteurs de façon à mesurer l'importance de ces derniers pour la décision à laquelle ils sont associés.

Le schéma de pondération utilisé par le système de Tapper tire profit des diverses caractéristiques pondérables des liens bibliographiques dont nous avons discuté plus haut. Ainsi la plus grande importance est attribuée à une vielle décision rendue par un petit tribunal situé dans une région reculée référée très fréquemment dans le document alors

que la plus petite valeur est attribuée à une décision très récente de la plus haute instance juridique d'une juridiction proche à laquelle le document réfère juste une fois. Cette pondération repose sur l'hypothèse que le premier cas n'arrive que lorsque la décision citée est très importante pour la cause en question alors que dans le second cas, la référence peut s'appliquer pour plusieurs raisons. Pour tenir compte de l'âge des décisions, la technique utilisée fut de choisir 1974, l'année des décisions des corpus tests, comme année pivot. Lorsqu'une décision est citée, sa date est soustraite de 1974, divisée par 10 et arrondie à l'entier le plus proche. De façon réciproque, 1974 est retranchée de la date d'une décision qui cite et le même procédé est appliqué par la suite. Le niveau de juridiction est mis en évidence en attribuant des constantes spécifiques aux décisions rendues par des tribunaux de différents niveaux. Ainsi, pour le corpus américain, une valeur de 10 est accordée aux décisions de la Cour suprême, 20 aux décisions des cours d'appel et 30 aux décisions des cours fédérales. De façon similaire, pour le corpus anglais, une valeur de 10 est attribuée aux décisions de la « House of Lords », 20 à celles de la « Appeal Court » et 30 à celles de la « Superior Court ». Des règles similaires sont appliquées pour tenir compte de l'éloignement géographique. La fréquence est représentée par un entier correspondant au nombre d'occurrences de la référence dans le texte. Pour tenter de contourner le problème posé par les références implicites, le système ne compte qu'une occurrence par page de texte imprimé, indépendamment du nombre de références effectivement présentes dans la page en question. La somme de ces valeurs fournit le poids accordé aux différentes composantes du vecteur de référence d'une décision.

Un autre aspect intéressant du système réside dans la mesure de corrélation qui y est proposée. En effet Tapper juge nécessaire d'introduire une mesure de corrélation mieux adaptée aux vecteurs d'items bibliographiques. Il estime que la mesure du cosinus utilisée traditionnellement pour comparer deux vecteurs attribue trop d'importance au fait qu'un item n'appartienne qu'à un des vecteurs et pas assez au fait qu'un autre soit commun aux deux. Pour pallier ce problème, deux techniques sont utilisées.

La première consiste à différencier le poids d'un item selon qu'il appartient aux vecteurs ou non. S'il est commun aux deux vecteurs, le poids qui lui est associé prend toute la valeur décrite plus haut (hit-value), sinon, il ne prend que la valeur associée à sa fréquence (miss-value). Tapper motive cette approche en expliquant qu'une décision fréquemment citée, apparaissant dans un seul des vecteurs comparés est importante non pas à cause du niveau de juridiction ou de son âge mais simplement parce qu'elle est citée fréquemment. Ce n'est que lorsqu'elle est une référence commune aux deux vecteurs que les caractéristiques telles qu'âge, niveau de juridiction, éloignement géographique prennent toute leur importance.

La deuxième technique sert à tenir compte du fait que, dans le cas des décisions juridiques, même un petit nombre de références communes indiquent un rapprochement au niveau du contenu. En effet, Tapper estime que si deux décisions discutent de façon détaillée un point A, elles ont un degré de similarité très élevé qui ne doit pas être réduit même si l'une de ces décisions traite également d'un point B qui n'apparaît pas dans l'autre. La technique utilisée constitue donc à faire en sorte que la similarité des deux vecteurs soit proportionnelle aux nombres de descripteurs qu'ils ont en commun en

multipliant le degré de similarité par un coefficient égal au nombre d'items bibliographiques apparaissant dans les deux décisions.

Illustrons ces techniques par un exemple. Prenons le cas où deux décisions, d_1 et d_2, contiennent les items bibliographiques suivants : *a*, *b*, *c*, *d*, *e*, *f*, *i*, *j*, *k*. Le tableau 1 montre les différents « hit-value » et « miss-value » associés à chaque item bibliographique pour chacune des décisions.

	d_1		d_2	
	« hit-value »	« miss-value »	« hit-value »	« miss-value »
a	35	5	0	0
b	21	1	25	2
c	17	5	11	1
d	14	2	13	4
e	31	4	0	0
f	22	2	0	0
i	0	0	35	5
j	0	0	12	6
k	0	0	31	1

Tableau 1

Suivant les techniques mentionnées plus haut, les « hit-value » de *b*, *c*, *d* et les « miss-value » de *a*, *e*, *f*, *i*, *j*, *k* seront utilisés dans la formule puisque seuls les items

bibliographiques *b*, *c*, *d* sont communs aux deux décisions. L'évaluation de la similarité des deux décisions donne donc :

$$\frac{3*(21+13+14)*(25+11+13)}{(21+13+14+5+1+2)*(25+11+13+5+6+1)} = 2.065$$

Une fois calculées de la sorte, ces mesures de similarité sont utilisées pour construire des agglomérats de décisions similaires. Cette technique permet non seulement de récupérer une décision parce qu'elle est similaire à une décision initiale mais aussi parce qu'elles sont toutes les deux similaires à une troisième. Un algorithme très simple est utilisé pour la construction des agglomérats. Supposons que six documents soient reliés de la façon suivante :

d_1 , d_2 .9	d_5 , d_6 .87
d_2 , d_3 .89	d_1 , d_3 .86
d_4 , d_5 .88	d_2 , d_6 .85

Le seul agglomérat existant au niveau .9 comprend d_1 et d_2. Au niveau .89, d_3 vient se rajouter à l'agglomérat. Un nouvel agglomérat composé de d_4 et d_5 se forme au niveau .88 et d_6 vient s'y rajouter au niveau .87. Au niveau .85, les deux agglomérats se fusionnent en un seul. Cet algorithme simple peut fournir de fâcheux résultats en présence de décisions discutant de plusieurs points. En effet, supposons qu'un agglomérat soit composé de décisions discutant d'un point A et qu'un autre soit composé de décisions traitant d'un point B complètement différent de A. Alors, ces deux agglomérats composés de documents assez différents se fusionneront dès qu'une décision se met à discuter des deux points. Pour pallier ce problème, la cohérence des agglomérats est mesurée pour déterminer leur stabilité à différents niveaux. La cohérence est définie comme étant le

nombre de liens effectifs dans l'agglomérat sur le nombre de liens possibles. Dans l'exemple précédent, la cohérence de l'agglomérat formé par d_1, d_2 et d_3 au niveau .89 est de .67 puisqu'il n'existe que deux liens (d_1,d_2 et d_2,d_3) sur les trois liens possibles. Au niveau .86, par contre, le nouveau lien entre d_1 et d_3 rétablit la cohérence à l'unité. Cet agglomérat est donc stable aux niveaux .89 et .86. Remarquons qu'au niveau .85, l'agglomérat formé de toutes les décisions n'est pas stable puisque seulement 6 des 15 liens possibles sont établis induisant une cohérence de .4.

Nous déplorons le fait que, dans son article, Tapper ne présente pas les résultats qu'il a obtenus avec son système, surtout ceux concernant les expériences de comparaison entre sa mesure de similarité et celle du cosinus qu'il semble avoir menées. L'article de Tapper ne fournit que des affirmations générales telles que le système semble remplir ses promesses théoriques quant à l'utilisation des vecteurs de références pour évaluer la similarité des décisions. Selon Tapper, les modifications apportées aux techniques traditionnelles se sont avérées efficaces et, chose surprenante, pas seulement pour les vecteurs de références pour lesquels elles s'appliquaient, mais également pour les vecteurs basés sur les termes. Finalement, Tapper mentionne qu'il ne semble pas être nécessaire d'inclure les techniques d'agglomération dans un système fonctionnel.

Les trois sections précédentes nous ont montré que les liens bibliographiques peuvent servir à mesurer la similarité de contenu entre documents. En particulier, nous pouvons maintenant conclure que des documents exhibant des similarités au niveau des liens bibliographiques qu'ils établissent dans un corpus aient une affinité au niveau du sujet qu'ils traitent. Ceci suggère donc qu'il soit possible de caractériser un document par

ses liens bibliographiques. Le prochain chapitre va décrire le système que nous avons implanté ainsi que le modèle duquel il découle.

Chapitre 2

Tripòt: un outil de navigation

2.1 Présentation générale

Nous avons implanté un prototype de système de recherche d'information, Tripòt, afin de vérifier que le modèle que nous avons élaboré fonctionne. Ce prototype nous a permis de valider notre modèle, d'en étudier la faisabilité et découvrir les difficultés et les avantages qu'il offre.

Tripòt diffère des outils de recherche d'information « traditionnels » que l'on retrouve habituellement sur le Web. La différence fondamentale, bien entendu, réside dans le fait que les liens bibliographiques et non les termes se trouvent au cœur du système. La seconde différence, découlant de la première, réside au niveau de l'interface entre l'utilisateur et le système. Comme il est peu souhaitable d'obliger l'utilisateur à formuler des requêtes composées de références explicites qui peuvent être de très longues chaînes de caractères, nous avons opté pour un autre type d'interface, bien qu'une interface traditionnelle demeure envisageable.

Dans Tripòt, l'utilisateur fournit au système un document initial. Le système utilise ce document pour générer un certain nombre de documents qui lui sont similaires et les propose à l'utilisateur. Cette interface s'est imposée pour plusieurs raisons. Tout d'abord il facilite l'implantation du prototype. En effet, il nous évite de programmer un module chargé d'analyser les requêtes de l'utilisateur afin de les transformer en un vecteur bibliographique dans un format compréhensible par le système. Un tel module constitué essentiellement de tâches de saisie et de vérification de données est bien sûr

réalisable, mais nous éloigne de la tâche primordiale consistant à évaluer la similarité entre document de la manière la plus efficace que possible. De plus, en cachant les détails de requêtes et de vecteurs bibliographiques à l'utilisateur, nous améliorons la convivialité du système. Finalement, l'interface retenue convient parfaitement pour l'expérimentation et la validation du modèle. En effet, comme nous le verrons plus loin, le succès du modèle sera évalué suivant sa capacité à regrouper des documents qui sont effectivement similaires.

Il s'ensuit que Tripòt est un logiciel permettant à l'utilisateur de naviguer à travers un corpus. Nous avons déjà mentionné que la navigation peut être considérée comme une forme de recherche d'information [Frisse 89]. C'est d'ailleurs le moyen de recherche immédiat qu'offre le Web considéré comme un vaste espace informationnel. Cependant, dès que le nombre de documents constituant le corpus devient grand, la navigation simple devient inefficace et l'utilisateur perd souvent du temps précieux à suivre des liens qui le conduisent à des documents non pertinents à son besoin d'information. Pour augmenter l'efficacité de la méthode, Tripòt lui propose un ensemble de documents susceptibles de l'intéresser parce que similaires au premier. L'usager peut alors consulter ces documents et, s'il trouve un document intéressant, il peut demander au système de lui retourner les documents similaires à ce dernier. Dans la section suivante, nous allons décrire notre prototype tout en précisant le modèle sur lequel il s'appuie.

2.2 L'architecture du système

Lors de l'étape de conception, l'une de nos préoccupations était de doter le système d'une architecture modulaire afin de mettre en valeur la flexibilité du modèle en facilitant la modification ou l'ajout d'un module pour adapter le système à une application

particulière. Le système comporte cinq composantes essentielles. La "Figure 10" illustre l'architecture du système et les relations entre les composantes.

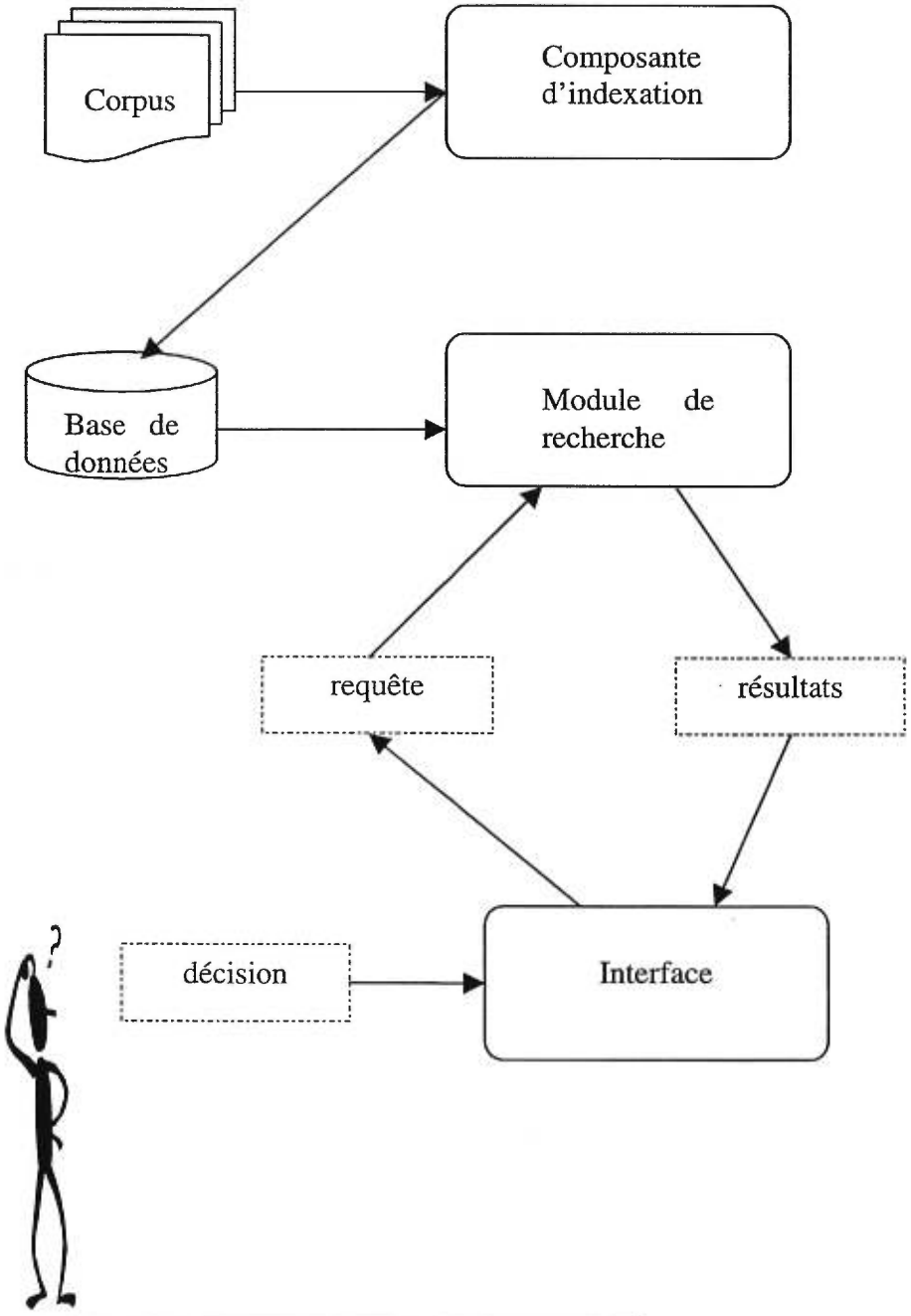

Fig. 10 architecture du système.

2.2.1 Le corpus

La première composante du système est, bien sûr, le corpus utilisé. Pour ce projet, notre choix s'est arrêté sur un corpus des décisions de la Cour suprême du Canada. Ce dernier est constitué de 460 décisions relatives à la *Charte canadienne des droits et libertés* s'étalant sur la période de 1983 à 1995. Les fichiers électroniques utilisés sont dans un format pseudo-HTML. Il s'agit de fichiers HTML (Hypertext Markup Language) lisibles par les différents fureteurs, dans lesquels nous avons inséré quelques balises supplémentaires n'appartenant pas à la norme HTML afin de mettre clairement en évidence les items bibliographiques. Ces fichiers occupent un total de 49.8 MB d'espace disque, soit environ 25 532 pages si l'on compte à peu près 2 KB par page.

Plusieurs raisons ont motivé notre choix d'utiliser ce corpus. La principale a trait à sa nature juridique. En effet, comme nous l'avons expliqué précédemment, ce type de corpus se prête bien à la conception d'un système basé sur les liens bibliographiques à cause du rôle important des références jurisprudentielles.

De plus, les décisions de la Cour suprême offrent un moyen assez simple et fidèle pour connaître leur contenu. En effet, au début de chaque décision apparaît une section contenant une liste de termes vedettes qui résument les faits et les principes de droit en cause. Cette section contient également l'énumération des différents articles de loi et de règlement utilisés dans le jugement en question. Ce travail des arrêtistes de la cour est fait de façon systématique en puisant d'un thesaurus construit spécialement à cette fin. Cette caractéristique est intéressante puisqu'elle permet à l'utilisateur de vérifier rapidement le contenu du jugement en question. Ainsi, il peut juger facilement de la ressemblance de

deux décisions. Ces diverses caractéristiques nous permettent de vérifier la similarité des décisions retournées par Tripòt et ainsi d'évaluer la capacité du système à regrouper les décisions similaires.

Finalement, le corpus de la Cour suprême offre plusieurs autres avantages qui ont facilité son traitement. En effet, les fichiers que nous avons utilisés sont issus d'une procédure de conversion assez complexe ayant pour but de rendre les décisions originales, fournies par la cour dans un format Wordperfect 5.1, disponibles sur le Web. Au cours de ces transformations, les fichiers ont été soumis à différents algorithmes d'épuration et de mise en forme qui, en bout de ligne, garantissent une certaine intégrité des informations. D'ailleurs, les différentes analyses préliminaires que nous avons effectuées sur le corpus n'ont révélé aucune anomalie majeure qui entraverait le bon fonctionnement du système à part le fait que le corpus choisi ne comporte que les décisions traitant de la *Charte des droits et libertés*. Bine que ceci ne constitue pas une entrave au fonctionnement du système, il peut toutefois avoir une influence négative sur les résultats obtenus étant donné que chaque décision manquante contient des informations permettant de catégoriser les autres documents du corpus.

Le système est muni d'un module d'enregistrement qui se charge d'identifier les décisions qui seront indexées. Ce module est également responsable de la sauvegarde des informations « administratives » du corpus. Ainsi, pour un corpus de N décisions, chaque décision est identifiée par un entier choisi entre 1 et N (ici N vaut 460). D'autres informations telles que le nom du fichier électronique ainsi que la chaîne de caractères représentant une référence à chaque décision sont sauvegardées afin que le système puisse faire le lien entre la référence, l'identificateur et le fichier associé à une décision.

2.2.2 La composante d'indexation

La composante d'indexation constitue le cœur du système. Elle se charge d'extraire les informations bibliographiques des décisions, de construire leurs représentations vectorielles et de les sauvegarder dans une base de données de sorte que les décisions similaires puissent être retrouvées par le module de recherche. Avant de décrire plus complètement cette composante, nous allons préciser le schéma de pondération utilisé par le système.

Un document sera représenté par un vecteur construit de la façon suivante :

$$d_i = (l_{i1}, l_{i2}, \cdots, l_{iS}) \text{où les } l_{ij} \text{ sont les poids associés au lien bibliographique j dans } d_i$$

C'est ce que nous appelons vecteur bibliographique. L'indice S indique le nombre de composantes constituant le vecteur. Dans ce modèle, les composantes du vecteur peuvent être associées à divers types de liens bibliographiques tels que le couplage bibliographique ou la cocitation et pas seulement à la référence ou à la citation tel que c'est le cas dans le système présenté par Tapper. Le choix des types de liens dépend de l'application à implanter.

Remarquons que le système proposé par Tapper avec les vecteurs de références se rapproche du modèle que nous venons de décrire. En fait, nos vecteurs bibliographiques peuvent être considérés comme une généralisation des vecteurs de références de Tapper. En effet, tout d'abord, ils apportent un nouvel éventail de possibilités qui, d'après nous, permet de l'adapter à d'autres types de corpus. Évidemment, nous pouvons nous attendre à ce que, dépendant du contexte dans lequel ils sont utilisés, certains types de liens bibliographiques soient préférables à d'autres aux fins de décrire les documents. En effet,

comme il a été montré qu'en général la cocitation est une meilleure mesure de la similarité des documents scientifiques que la citation, nous pouvons supposer que des vecteurs bibliographiques dont les composantes représentent le degré de cocitation des différents documents d'un corpus scientifique produiront un système aussi efficace, sinon plus, que si les composantes sont reliées aux citations. Bien entendu, ce genre de suppositions ne se base que sur l'intuition et des expériences de comparaisons s'imposent si nous voulons en avoir le cœur net.

Par ailleurs, notre modèle apporte un formalisme plus rigoureux que celui trouvé dans les articles de Tapper. Toutefois, la description que nous avons faite jusqu'à présent du vecteur bibliographique est trop générale et ne peut servir à décrire un système fonctionnel. Elle ne fait que donner une idée de la nature des composantes du vecteur bibliographique. Pour définir complètement un tel système il faut encore préciser trois aspects :

- La représentation des documents et des requêtes. Nous savons déjà que, dans notre modèle, documents et requêtes sont représentés par un vecteur bibliographique. Cependant, il faut mentionner quels types de liens bibliographiques sont pris en compte dans ce système et quel schéma de pondération est utilisé, c'est-à-dire, quel algorithme attribue les poids associés aux différents liens contenus dans les vecteurs bibliographiques.

- La méthode d'évaluation de la similarité. Puisque dans notre modèle il est question de vecteurs, nous pouvons nous attendre à ce qu'une méthode algébrique soit utilisée. Cependant, les expériences de Tapper nous ont montré qu'il n'existe pas qu'une seule façon de procéder.

- La façon d'ordonner et de présenter le résultat de la requête à l'utilisateur.

Ce n'est qu'après avoir défini ces aspects que nous aurons complètement décrit le système faisant l'objet du travail. Aussi, dans le prochain chapitre commençons-nous par décrire ces trois aspects pour le prototype que nous avons implanté.

Rappelons que dans Tripòt, une décision est représentée par un vecteur bibliographique. Pour ce prototype, les vecteurs sont associés aux liens de référence et de citation établis dans le corpus. Ainsi, une décision i, dénotée d_i, est représentée par un vecteur

$d_i = (r_{i1}, \cdots, r_{iM}, c_{i1}, \cdots, c_{iN})$, où r_{ij} est le poids associé à la $j^{ème}$ référence de d_i, tandis que c_{ik} est le poids associé à la $k^{ème}$ citation de d_i.

Ici, l'indice N représente le nombre de documents du corpus. Nous attirons l'attention sur le fait que le deuxième indice attaché aux poids des références varie jusqu'à M. Deux éléments déterminent la taille de M. Tout d'abord le corpus est constitué de décisions qui furent rendues à partir de 1983. Or plusieurs décisions font référence à des jugements qui ont eu lieu avant cette date. De plus, comme nous l'avions mentionné plus haut, le corpus est constitué uniquement des décisions durant la période de 1983 à 1995 se rapportant à la *Charte des droits et libertés*. Or il existe d'autres décisions de la Cour suprême de la période qui sont référées dans les décisions indexées. Fatalement donc, le nombre de références répertoriées par le système sera plus grand que le nombre de décisions indexées et ce, même si nous choisissons de ne tenir compte que des références aux décisions de la Cour suprême du Canada. Dans le cas des citations par contre, le deuxième indice varie jusqu'à N car nous ne pouvons évidemment pas savoir si une décision a été citée dans un jugement lorsque le texte de ce dernier ne nous est pas

disponible. Ainsi, les informations sur les liens de citations ne peuvent être recueillies que pour les N décisions constituant le corpus.

Le schéma de pondération que nous avons retenu est très simple. Les r_{ij} représentent la fréquence avec laquelle d_i réfère à d_j, c'est-à-dire, le nombre de fois qu'une référence à d_j apparaît dans d_i. Réciproquement, les c_{ik} représentent la fréquence à laquelle d_i est citée par, c'est-à-dire, le nombre de fois qu'une référence à d_i apparaît dans d_k. Remarquons que $c_{ik} = r_{ki}$.

Nous avons privilégié ce schéma de pondération simple dans le cadre de notre projet. Des travaux subséquents pourraient cependant permettre de tirer profit des autres caractéristiques pondérables des liens bibliographiques. En effet, étant donné que nous avons choisi d'indexer que les références aux décisions de la Cour suprême, les critères de pondération tels que le niveau de l'instance judiciaire et l'éloignement géographique ne sont plus pertinents pour nos fins. Par ailleurs, malgré que le critère de l'âge soit applicable, nous avons retenu le schéma de pondération simple afin de mieux étudier le comportement du système. De plus, nous n'avons tenu compte que de la fréquence des items bibliographiques explicites dans les textes sans faire attention aux items implicites. Ces choix ne limitent cependant pas la valeur de l'approche qu'illustre notre prototype.

L'algorithme de traitement se subdivise en trois étapes correspondant chacune à un module de la composante d'indexation. Le premier module est responsable de l'indexation proprement dite des décisions. Il se charge de parcourir le texte des décisions afin d'en extraire les items bibliographiques. Ces items sont identifiés par un entier qui, dans le cas où ils désigneraient une décision du corpus, correspond à l'identificateur de la décision en question. Différents corpus peuvent exiger différentes variantes de ce module

à cause du caractère spécifique du travail d'identification des items bibliographiques. Dans notre cas, il s'agit de repérer les balises qui avaient été placées au cours des conversions antérieures et qui délimitent les items. Par la suite, le programme récupère la chaîne comprise entre les balises et y enlève toutes autres balises de formatage afin de ne conserver que le texte de l'item bibliographique. Remarquons que le repérage des items nécessite l'utilisation d'expressions régulières ou d'un outil équivalent. Si la présence des balises simplifie le travail du module d'indexation dans notre cas, il n'en demeure pas moins que pour analyser les fichiers et y insérer les balises autour des items bibliographiques nous avons du recourir à un outil spécialisé dans la manipulation de textes et d'expressions régulières. Il s'agit du langage Omnimark™, spécialisé dans le traitement des documents SGML. Le traitement initial ayant pour l'insertion des balises délimitant les items bibliographiques consiste à reconnaître le pattern suivi par les références jurisprudentielles de la Cour suprême. Ce traitement équivaut à peu près à celui de normalisation des termes que nous retrouvons dans les autres outils de recherche d'information à indexation automatique. Enfin, le premier module est aussi responsable de construire la première partie des vecteurs bibliographiques. C'est-à-dire qu'il calcule la fréquence d'apparition des items dans chaque décision et sauvegarde ces informations dans un fichier de vecteurs de références dont la structure sera présentée dans la section suivante.

Le second module rattaché à la composante d'indexation sert à construire la seconde partie des vecteurs bibliographiques, à savoir, celle concernant les citations. Cette partie des vecteurs bibliographiques, rappelons-le, indique le degré avec lequel les décisions du corpus citent une référence. Elle est donc construite en utilisant le fichier de

vecteurs de références produit par le premier module et une technique bien connue dans le domaine de la recherche d'information, c'est-à-dire, l'inversion. Cette technique consiste dans notre cas à prendre l'information du fichier de vecteurs de références que nous pouvons imaginer sous la forme,

décision_1	*référence_1*	...	*Référence_M*
décision_2	*référence_1*	...	*Référence_M*
...	...	...	...
décision_N	*référence_1*	...	*Référence_M*

et à la réorganiser de la façon suivante :

Référence_1	*décision_1*	...	*Décision_N*
...	...	...	...
référence_M	*décision_1*	...	*Décision_N*

Remarquons qu'avec notre procédé d'identification des décisions et des références, les N premières lignes de cette matrice représentent les vecteurs de citations des N décisions du corpus.

Notre module d'inversion utilise l'algorithme FAST-INV (fast inversion) proposé par Edward A. Fox et Whay C. Lee [Frakes & Baeza-Yates 92]. Cet algorithme tire profit de deux principes : la grande quantité de mémoire vive disponible sur les ordinateurs modernes et l'ordre inhérent aux données d'entrée. Le premier principe est important, car même si la quantité de données à traiter est très grande, le travail à effectuer peut être réalisé de façon très efficace en divisant les données en plusieurs chargements de tailles raisonnables qui sont traités rapidement en mémoire principale. Le second principe est crucial car il permet d'éviter de trier les fichiers de vecteurs de références afin d'obtenir le

fichier inversé. En effet, même avec de bons algorithmes de tri de l'ordre de n log n, les opérations de disques détériorent la performance de telles méthodes.

L'algorithme FAST-INV s'exécute en trois étapes. La première étape consiste à parcourir le fichier des vecteurs de références afin de construire une table de chargement et un fichier de positions. La table de chargement contient les informations permettant de diviser les données en plusieurs groupes de petite taille de sorte qu'ils puissent être chargés et traités en mémoire principale l'un à la suite de l'autre. Le fichier de positions contient des informations qui permettent au module de savoir où insérer les données du fichier de vecteurs de références dans le fichier inversé de sorte qu'aucun tri ne soit nécessaire. À la seconde étape de l'algorithme, le fichier de vecteurs de références est divisé suivant les informations contenues dans la table de chargement. Finalement, à la troisième étape, le module utilise le fichier de positions et charge les différents groupes générés à l'étape précédente afin de produire le fichier inversé final contenant les vecteurs de citations.

Le troisième module de la composante d'indexation se charge de calculer la similarité des N décisions entre elles. La formule de la mesure de similarité que nous avons utilisée est la suivante :

$$Sim(d_i, d_j) = \frac{\sum_{k=1}^{M} r_{ik} \cdot r_{jk} + \sum_{k=1}^{N} c_{ik} \cdot c_{jk}}{\left(\sum_{k=1}^{M} r_{ik}^2 + \sum_{k=1}^{N} c_{ik}^2 \right) \cdot \left(\sum_{k=1}^{M} r_{jk}^2 + \sum_{k=1}^{N} c_{jk}^2 \right)}$$

Remarquons que cette formule n'est autre que la formule du cosinus de l'angle formé par deux vecteurs. La formule est légèrement différente de sa forme habituelle parce qu'elle a été adaptée à la notation des vecteurs bibliographiques utilisés dans notre système. Nous

n'avons pas jugé nécessaire d'élaborer une nouvelle mesure de similarité mieux adaptée aux vecteurs bibliographiques. En effet, le schéma de pondération choisi ne l'exigeait d'aucune façon. Notons cependant que la nature modulaire du prototype autorise l'utilisation d'autres mesures de similarité.

2.2.3 La base de données

La base de données du système comporte cinq fichiers.

- Le *fichier des décisions* regroupe les informations « administratives » sur les décisions du corpus. Ces informations s'organisent à l'intérieur d'un enregistrement pour chaque décision. Les enregistrements se composent des attributs suivants :
  - Intitulé (référence complète désignant la décision);
  - Nom de fichier (le chemin complet);
  - Identificateur (entier compris entre 1 et N);
  - Taille (en octets du fichier);
- Le *fichier des items bibliographiques* contient les informations permettant d'identifier les items apparaissant dans les décisions. Comme pour les décisions, ces informations sont regroupées à l'intérieur d'un enregistrement composé de l'intitulé et de l'identificateur associé aux différents items.
- Le *fichier des vecteurs de références* des décisions du corpus est composé d'une série de triplets (i, j, k) qui indiquent que la référence j apparaît k fois dans la décision i. La "Figure 11" donne un exemple d'un fichier de vecteurs de références.
- Le *fichier des vecteurs de citations* possède une structure identique à celui des vecteurs de références sauf que l'ordre des identificateurs a été inversé. Ainsi, un

triplet (i, j, k) dans ce fichier indique que la référence i apparaît en k fois dans la décision j.

- Le *fichier des degrés de similarité* est lui aussi composé de triplets (i , j , k) qui indiquent cette fois que le cosinus de l'angle formé par les vecteurs bibliographiques des décisions i et j est k. Les décisions concernées sont d'autant plus similaires que le nombre k est proche de 1.

décision id	*Référence id*	*Fréquence*
1	134	2
1	135	6
2	17	2
2	135	1
2	466	3
3	12	1
4	21	5

Figure 11 Structure du fichier de vecteurs de référence

La gestion des données du prototype comporte enfin un module d'accès. Ce module permet l'accès simple et uniforme aux informations contenues dans les différents fichiers constituant la base de données. Ainsi, après avoir créé un enregistrement, un client (le module d'indexation, par exemple) peut simplement demander au module d'accès d'ajouter l'information contenue dans le nouvel enregistrement à la base de données sans se soucier du fichier où sauvegarder l'information ni de son format. Enfin,

ce module permet à un client de ne charger que les fichiers de la base de données dont il a besoin de sorte à minimiser l'utilisation des ressources systèmes.

2.2.4 Le module de recherche

Le module de recherche se charge d'identifier l'ensemble des décisions les plus similaires à la requête de l'utilisateur. Pour les fins du module de recherche, une requête n'est rien d'autre qu'un entier identifiant une des décisions du corpus. Puisque les similarités ont déjà été calculées lors de l'indexation, le travail du module recherche est très simple. Il lui suffit d'interroger la base de données afin de récupérer l'ensemble des documents du corpus ordonnés préalablement suivant leur similarité à la décision de la requête. Autrement dit, chaque requête de l'utilisateur génère une permutation des documents du corpus et le module de recherche s'occupe d'aller récupérer cette permutation de la base de données et de la fournir au module d'interface qui s'occupera de la présentation des résultats à l'utilisateur.

En partie à cause de l'application à laquelle nous destinions notre prototype et en partie pour des raisons d'efficacité, nous avons également implanté un serveur chargé de gérer la communication entre le module de recherche et le module d'interface. Le travail du serveur consiste à accepter les connexions du client (le module d'interface), à recevoir et à interpréter sa requête, à passer cette requête au module de recherche et, enfin, à retourner les résultats au client. Ces échanges d'informations se font suivant un protocole que nous avons conçu afin d'assurer la synchronisation de la communication entre le client et le serveur. De plus, le serveur se charge de formuler les résultats dans un format permettant au module client de les décoder en vue de leur présentation à

l'utilisateur. Enfin, le serveur peut également, à la demande du client, retourner l'identificateur d'une décision ou énumérer toutes les décisions qui constituent le corpus.

2.2.5 Le module d'interface

Le module d'interface sert d'intermédiaire entre le module de recherche et l'utilisateur. Il se charge principalement de la présentation des résultats sous la forme de pages HTML. Le client gère deux types de pages HTML. Tout d'abord, il s'occupe de récupérer les fichiers du corpus permettant ainsi à l'utilisateur de consulter le texte d'une décision. Ensuite, il génère de façon dynamique des pages présentant les résultats d'une requête sous la forme d'une liste constituée d'intitulés des décisions similaires et de leur degré de similarité. La liste contient les vingt décisions qui sont les plus similaires à la décision initiale ordonnées suivant leur degré de similarité. Les pages générées contiennent également des liens permettant de consulter les décisions qui y sont énumérées ou de choisir une de celles-ci pour une nouvelle requête.

La "Figure 12" illustre l'aspect visuel de l'interface du prototype suite à une requête. Le module d'interface présente les résultats dans trois fenêtres. Celle d'en haut, la fenêtre principale, contient la décision initiale, c'est-à-dire, celle qui a été utilisée comme requête. Dans ce cas-ci, il s'agit de la décision *R. c. Hufsky,* [1988] 1 RCS 621, tel indiqué par l'intitulé se trouvant au haut de la page. La fenêtre en bas à gauche présente les résultats produits par le module de recherche. Il s'agit de la fenêtre de navigation qui contient tout d'abord un lien « Redémarrer » qui ramène l'utilisateur à une page contenant la liste de toutes les décisions indexées par le système de telle sorte qu'il puisse lancer une nouvelle recherche. Pour le reste, elle présente les résultats générés par le module de recherche. Comme mentionné précédemment, ces résultats sont proposés sous la forme

d'une liste contenant les vingt décisions les plus similaires à la décision initiale. Chaque élément de cette liste est composé de l'intitulé d'une décision qui est aussi un lien sur lequel l'utilisateur peut cliquer pour en consulter le texte. Le degré de similarité est arrondi à trois décimales. Chaque décision similaire est enfin accompagnée d'un lien « Requête » permettant de lancer une nouvelle requête où cette décision deviendrait le document initial. Quant à la fenêtre en bas à droite, que nous appellerons la fenêtre de visualisation, elle permet à l'utilisateur de consulter les décisions similaires proposées par le module de recherche. Ces trois fenêtres définissent pour chaque requête ce que nous appellerons un contexte de navigation caractérisé par la décision initiale et ses décisions similaires.

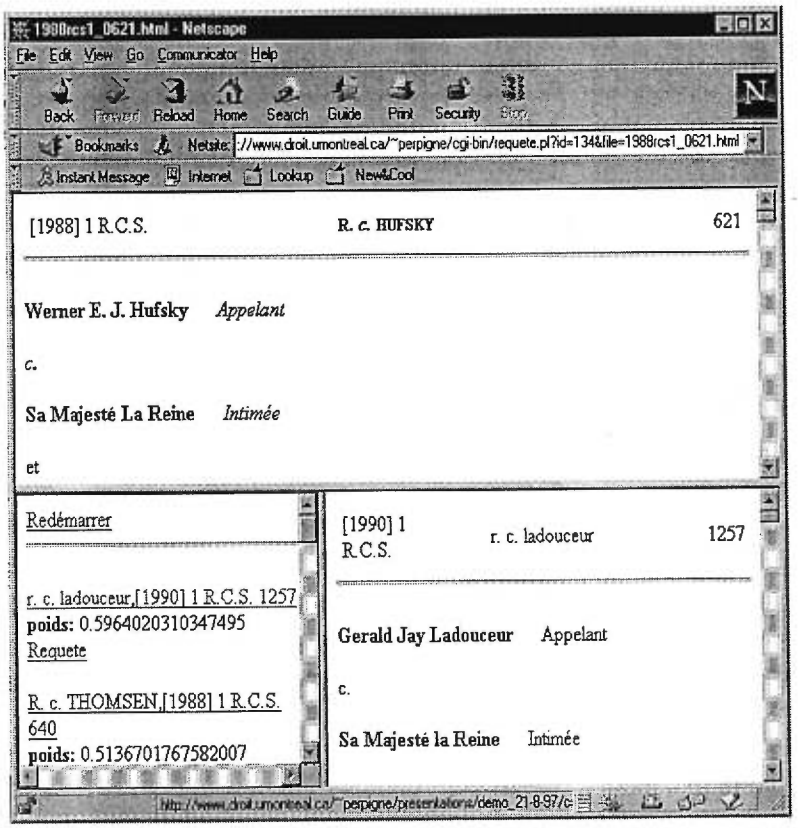

Figure 12 Module d'interface

Il est nécessaire de bien noter les fonctions différentes des deux types de liens apparaissant dans la fenêtre de navigation. Tout d'abord l'activation du lien associé à l'intitulé d'une décision a pour effet de charger le fichier de la décision en question dans la fenêtre de visualisation afin que l'utilisateur puisse rapidement en consulter le contenu. Dans la "Figure 12", nous voyons apparaître dans la fenêtre de visualisation la décision intitulée *R. c. Ladouceur*, [1990] 1 RCS 1257 après avoir activé le premier lien de la liste des décisions similaires à la requête apparaissant dans la fenêtre de navigation. Par ailleurs, l'activation d'un des liens « Requête » amène un changement du contexte de navigation. En effet, lorsque l'utilisateur clique sur un lien de ce type, le client (un script CGI) construit une nouvelle requête, avec pour décision initiale, celle associée à cet élément « Requête » et passe cette information au module de recherche. Par la suite, le client utilise les résultats produits pour générer un nouveau contexte de navigation, cette fois, avec la décision de la nouvelle requête apparaissant dans la fenêtre principale. Bien sûr, la liste de décisions similaires apparaissant dans la fenêtre de navigation est, elle aussi, mise à jour. La "Figure 13" montre le nouveau contexte de navigation obtenu après l'activation du lien requête associé à la décision « R. c. Ladouceur » de la "Figure 12". Les liens « Requête » jouent donc un rôle important. Ils permettent à l'utilisateur de changer de contexte de navigation sans trop s'éloigner du contexte initial puisque le nouveau contexte obtenu a été généré par une décision similaire à celle ayant généré le précédent. C'est ainsi que l'utilisateur arrive à naviguer à travers le corpus en consultant des agglomérats de décisions similaires.

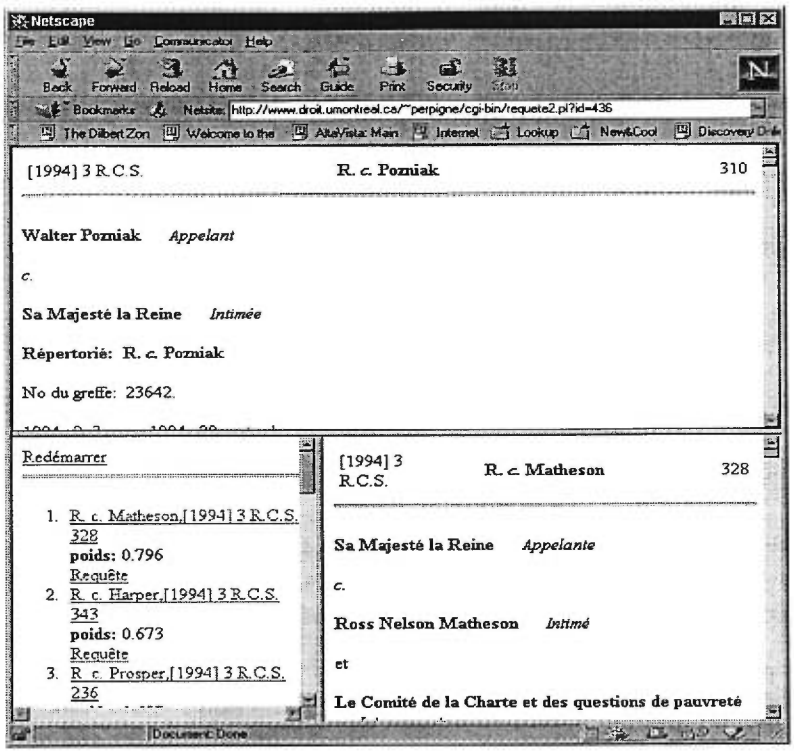

Figure 13 Contexte de navigation

2.3 L'Implémentation

Cette section a pour but de présenter brièvement divers aspects du cadre conceptuel du projet.

2.3.1 Le SGML

Comme mentionné au début de la section précédente, le corpus utilisé par notre système est constitué de fichiers dans un format pseudo-HTML. Ce format est obtenu à la suite d'une série de conversions opérées sur les fichiers originaux fournis dans un format Wordperfect. Le SGML, acronyme anglais de « Standardized General Markup Language », est la colle de ce processus de conversion. Ce langage permet, à l'aide de balises (marqueurs spéciaux délimités par les caractères '<' et '>'), de faire ressortir la

structure d'un texte. Le paragraphe suivant montre un exemple de texte dans un format SGML. Dans ce cas, le balisage identifie les éléments constituant la référence comprise entre les balises <REF> et </REF>, à savoir la clé qui contient entre autre une année, un numéro de volume, un nom de recueil et un numéro de page.

```
<REF>
      <CLE>
            [<ANNEE>1987</ANNEE>] <VOL>1</VOL>
            <RECUEIL>R.C.S.</RECUEIL> <PAGE>1233</PAGE>
      </CLE>
</REF>
```

Le SGML permet ainsi de définir des classes de documents possédant une structure particulière. Cette structure est décrite au moyen d'une grammaire appelée DTD (de l'anglais « Document Type Definition »). À l'aide d'une grammaire de ce type, un analyseur grammatical peut vérifier si un document balisé appartient à une classe spécifique de documents ou reconnaître les différentes parties constituant un document SGML valide. Les fichiers constituant le corpus utilisé par notre système furent créés par un outil de ce genre. Les fichiers originaux ont d'abord été convertis selon une DTD SGML décrivant la classe des décisions de la Cour suprême. Nous avons ensuite transformé ces fichiers en HTML en prenant soin de garder les balises SGML que nous avons jugées utiles pour notre projet, à savoir les balises délimitant les références. Les fichiers hybrides résultant de ce processus sont très utiles, car ils peuvent servir directement à la publication sur le Web et se prêtent bien su travail d'indexation de notre système.

2.3.2 L'approche orientée objet (JAVA)

À l'exception du module d'interface, toutes les composantes du système ont été implantées avec la version 1.0.2 du langage Java. En plus des avantages usuels qu'il offre, tels l'indépendance de plate-forme et une bonne intégration avec le Web, nous avons opté pour ce langage à cause de son orientation objet. En effet, les paradigmes orientés objet semblaient tout à fait appropriés pour le type de projet que nous voulions entreprendre. En particulier, les principes tels que l'encapsulation et l'abstraction furent très utiles lors du développement et de la mise au point de l'architecture modulaire du système. De plus, l'héritage et la réutilisation de code favorisent la réalisation de notre objectif de production d'un prototype flexible permettant les améliorations et modifications futures.

2.3.3 L'architecture client serveur, CGI

Comme nous l'avons mentionné, le prototype est conçu selon une architecture client serveur de sorte qu'il puisse être utilisé sur le Web. Le lien entre l'utilisateur et le système est assuré par des scripts CGI. « Le Common Gateway Interface » est un protocole permettant aux programmes s'exécutant sur la machine hôte du serveur Web d'obtenir de l'information d'un utilisateur visionnant des pages Web. De façon plus précise, l'interface de notre système est gérée par deux scripts Perl. Le premier s'occupe de générer les contextes de navigation. Il s'assure de charger le fichier correspondant à la décision initiale de la requête dans la fenêtre principale, de vider la fenêtre de visualisation, d'appeler l'autre script et d'afficher le résultat de ce dernier dans la fenêtre de navigation. Le second script se charge de fournir la requête au serveur du module de recherche et de générer la page de navigation avec les liens appropriés.

Chapitre 3

La Réalisation du projet

Nous avons conçu Tripòt afin de pouvoir vérifier notre hypothèse quant à l'utilisation des liens bibliographiques pour le repérage de l'information. Nous souhaitions en particulier vérifier leur valeur comme des descripteurs du contenu des documents. Dans ce chapitre, nous présentons les différents résultats obtenus lors de nos expériences avec Tripòt.

Habituellement, la performance des systèmes de recherche d'information est évaluée suivant deux critères, à savoir, leur capacité à retrouver les documents pertinents aux requêtes des utilisateurs ainsi que leur capacité à rejeter ceux qui ne le sont pas. Ces deux critères sont respectivement mesurés par le *rappel* et la *précision*. Pour une requête donnée, ces mesures ont été définies de la façon suivante :

$$\text{Rappel} = \frac{\text{Nbre de documents pertinents retournés}}{\text{Nbre de documents du corpus}}$$

$$\text{Précision} = \frac{\text{Nbre de documents pertinents retournés}}{\text{Nbre de documents retournés}}$$

Un système idéal combine des niveaux de rappel et de précision très élevée. C'est-à-dire, qu'il retrouve la majorité des documents pertinents du corpus et qu'une très grande partie des documents retournés sont pertinents. Remarquons qu'il est, en pratique, très difficile d'atteindre simultanément un haut niveau de rappel et de précision. En effet, afin d'avoir de bonnes chances de récupérer tous les documents pertinents à une requête et atteindre ainsi un niveau de rappel très élevé, les systèmes doivent retourner un grand de documents et ceci au détriment de leur niveau de précision. Aussi pour comparer des

systèmes, les chercheurs fixent un niveau de rappel et compare la précision obtenue avec les différents systèmes ou vice versa. De toute évidence, ce genre d'évaluation nécessite un corpus pour lequel il existe un ensemble de requêtes déterminées pour lesquelles les résultats sont connus. Tel est le cas des corpus utilisés traditionnellement dans les diverses expériences en recherche d'information tel le corpus du CACM. Pour de tels corpus, évaluer le rappel et la précision d'un système est un calcul simple et direct. Dans le cas des corpus pour lesquels ces ensembles ne sont pas connus, les chercheurs ont recours à des méthodes d'évaluation statistiques afin de déterminer le nombre de documents pertinents à une requête. Une fois ces paramètres déterminés, des estimations du rappel et de la précision des systèmes peuvent être calculées.

Bien que ces méthodes constituent l'approche habituelle d'évaluation d'un système, nous avons dû retenir un autre type d'évaluation pour Tripòt. En effet, notre but n'est pas tellement de mesurer l'efficacité de ce prototype ou de le comparer avec d'autres systèmes existants, mais plutôt de valider l'hypothèse ayant conduit à son implémentation. Valider cette hypothèse revient à vérifier que le prototype, basé sur le modèle que nous avons élaboré au Chapitre 2, réussit à regrouper les décisions similaires entre elles.

Comme nous l'avions mentionné au Chapitre 3, la similarité entre les décisions sera évaluée à l'aide de la liste des termes vedettes présente au début de chaque décision. Aussi, pour les fins de notre évaluation, avons-nous défini deux niveaux de similarité. Nous disons que deux décisions sont fortement similaires lorsqu'elles possèdent des termes vedettes en commun et que les même articles de loi ou de règlement y sont

évoqués. Par contre, nous dirons que deux décisions sont faiblement similaires quand seul l'un ou l'autre de ces cas survient.

Évidemment, deux décisions fortement similaires sont également faiblement similaires. Nous avons défini le deuxième niveau de similarité, car il est assez exceptionnel que deux décisions traitent de faits identiques même lorsque les mêmes concepts juridiques, par conséquent les mêmes articles de loi, sont en cause. De façon semblable, il peut arriver également que deux décisions traitent des même faits et que les mêmes concepts juridiques soient en cause mais qu'elles ne mentionnent pas les mêmes articles de loi pertinents ou que l'une d'entre elles traite de concepts ne se trouvant pas dans l'autre. Dans de tels cas, nous ne pouvons pas nier les liens existant entre ces décisions, et il se peut qu'un juriste intéressé par l'une estime, dépendant de ses besoins, l'autre tout à fait pertinente. Finalement, remarquons que le terme « faible » utilisé pour qualifier le second niveau de similarité n'indique pas forcément que les décisions sont peu similaires. Ce terme a été choisi afin de distinguer les niveaux de similarité et faire ressortir le caractère exceptionnel de la similarité forte. La prochaine section présente les expériences que nous avons entreprises avec Tripòt.

3.1 L'expérimentation et les résultats

Afin de tester le prototype, nous lui avons soumis 20 requêtes choisies de façon aléatoire. Pour chaque requête, nous avons consulté les vingt décisions retournées par le système afin de pouvoir déterminer la similarité de leur contenu à celui de la décision initiale selon le cadre établi au paragraphe précédent. Le *Tableau 1* présente les résultats obtenus. La première colonne fournit l'identificateur de la décision initiale. Les deuxième et troisième colonnes donnent respectivement le nombre de décisions fortement et

faiblement similaires à la décision initiale dans la liste des vingt décisions retournées par le système. Les deux dernières colonnes indiquent respectivement le degré de similarité associé à la première et à la vingtième décision. Dans les paragraphes suivant nous allons commenter les résultats présentés.

id.	forte	Faible	maximum	minimum	Intitulé
39	1	6	0.667	0.092	R. c. Jewitt ,[1985] 2 R.C.S. 128
58	0	0	0	0	Ville de Brossard c. Pelletier,[1986] 1 R.C.S. 53
70	0	3	0.369	0.097	E. (MME) c. EVE,[1986] 2 R.C.S. 388
79	0	0	0	0	R. c. Dawson,[1987] 2 R.C.S. 461
81	1	2	0.441	0.123	C.(G.) c. V.-F.(T.),[1987] 2 R.C.S. 244
105	3	4	0.328	0.118	R. c. Robertson,[1987] 1 R.C.S. 918
128	2	5	0.254	0	WASHINGTON (ÉTAT DE) c. JOHNSON,[1988] 1 R.C.S. 327
142	9	13	0.526	0	R. c. Upston,[1988] 1 R.C.S. 1083
170	0	5	0.230	0.053	Moysa c. Alberta (labour relations board),[1989] 1 R.C.S. 1572
180	13	19	0.634	0.328	R. c. Black,[1989] 2 R.C.S. 138
237	2	7	0.328	0.102	Mahe c. Alberta,[1990] 1 R.C.S. 342
246	3	4	0.750	0.166	R. c. Wilson,[1990] 1 R.C.S. 1291
247	3	13	0.600	0.152	rudolph wolff & co. c. canada,[1990] 1 R.C.S. 695
298	0	10	0.367	0.110	R. c. Gruenke,[1991] 3 R.C.S. 263
324	4	4	0.707	0.073	R. c. Martin,[1992] 1 R.C.S. 838
353	16	16	1	0.059	R. c . D. (S.),[1992] 2 R.C.S. 161
359	2	8	0.333	0.067	Vidéotron ltée c. Industries microlec produits électroniques inc.,[1992] 2 R.C.S. 1065
379	1	7	0.355	0.067	Hunt c. T&N plc,[1993] 4 R.C.S. 289
436	18	20	0.790	0.192	R. c. Pozniak,[1994] 3 R.C.S. 310
453	3	10	0.526	0.083	R. c. Durette,[1994] 1 R.C.S. 469

Tableau 2

Tout d'abord nous remarquons que les requêtes 79 et 58 n'ont généré aucune décision similaire. Dans les deux cas, la consultation des vecteurs bibliographiques nous a permis de vérifier qu'ils sont nuls et que par conséquent il s'agit de décisions ne possédant pas de références ni de citations. Ce type de décisions « isolées » n'apporte aucune information utile au système pour les fins de repérage. Il peut exister d'autres types de décisions isolées. C'est notamment le cas des décisions qui sont les seules à citer et/ou à être citées par d'autres décisions. Remarquons que, même si elles sont isolées, ce deuxième groupe de décisions isolées peut contenir des informations utiles au système pour regrouper d'autres décisions. N'ayant pas fait une analyse exhaustive du corpus, nous ne savons pas combien de décisions isolées sont présentes dans la collection et l'influence qu'elles exercent sur les résultats. Toutefois remarquons qu'une inspection des deux décisions « isolées » retrouvées lors de nos expériences nous a montré qu'il s'agit de jugements oraux exprimés en quelques lignes dans les fichiers. Ces jugements ne comportent aucun item bibliographique et leur brièveté explique sans doute qu'ils n'ont pas été cités. Il s'agit de cas où un système fondé sur les liens bibliographiques est moins utile.

La requête 353 constitue un autre cas exceptionnel. En effet, mis à part le fait que cette requête ait généré beaucoup de décisions similaires, nous remarquons que l'une de celles-ci possède un vecteur bibliographique identique à celui de la décision initiale car son degré de similarité est de 1. En consultant le texte de ces deux décisions, nous constatons qu'elles apparaissent effectivement fort similaires. La "Figure 14" montre le contexte de navigation généré par la requête 353. Sur cette figure, nous pouvons constater la similarité de la liste des termes vedettes des deux décisions. L'analyse de leur vecteur

bibliographique a révélé qu'ils ne sont constitués que de liens de références, c'est-à-dire que toutes les composantes associées aux liens de citations sont nulles. En effet, ces deux décisions ne sont citées par aucune autre du corpus. Cet exemple nous permet de déduire que la similarité des liens de références qu'établissent les décisions juridiques est une bonne indication de la similarité de leur contenu.

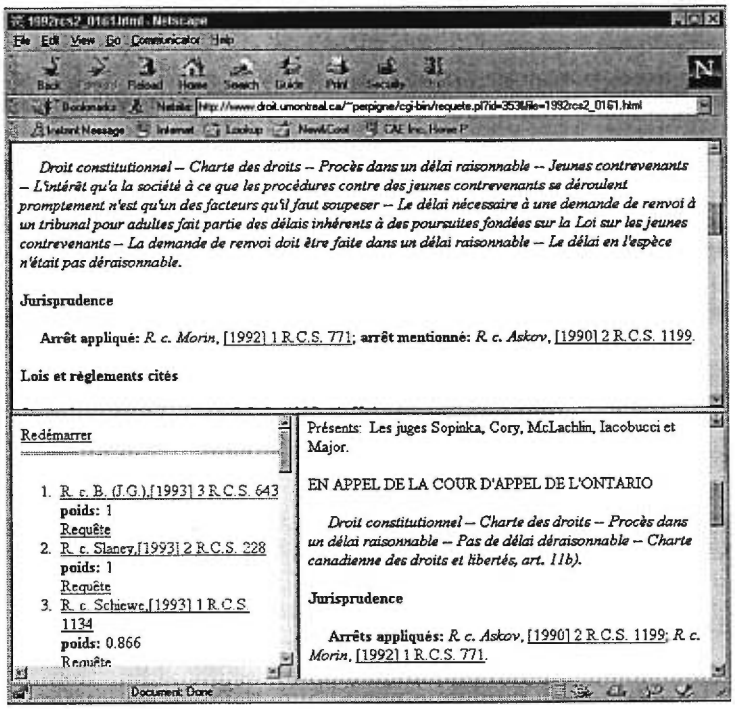

Figure 14 Références et similarité

Qu'en est-il du coté des liens de citations? Nous n'avons pas trouvé dans le corpus des décisions possédant exactement les mêmes liens de citations. Cependant, nous avons trouvé des décisions qui n'ont pas de références mais qui sont citées et, par conséquent, sont jugées similaires à d'autres décisions uniquement parce qu'elles partagent avec celles-ci un certain nombre de citations. Tel est le cas de la requête 142 dont le contexte de navigation est montré à la "Figure 15". Nous pouvons à nouveau constater la très grande similarité des termes vedettes et des articles de loi en cause entre la décision

initiale et la décision intitulée *R. c. Manninen*, [1987] 1 RCS 1233 affichée dans la fenêtre de visualisation. Nous en déduisons donc que la similarité des liens de citations est également une bonne indication de la similarité du contenu des décisions.

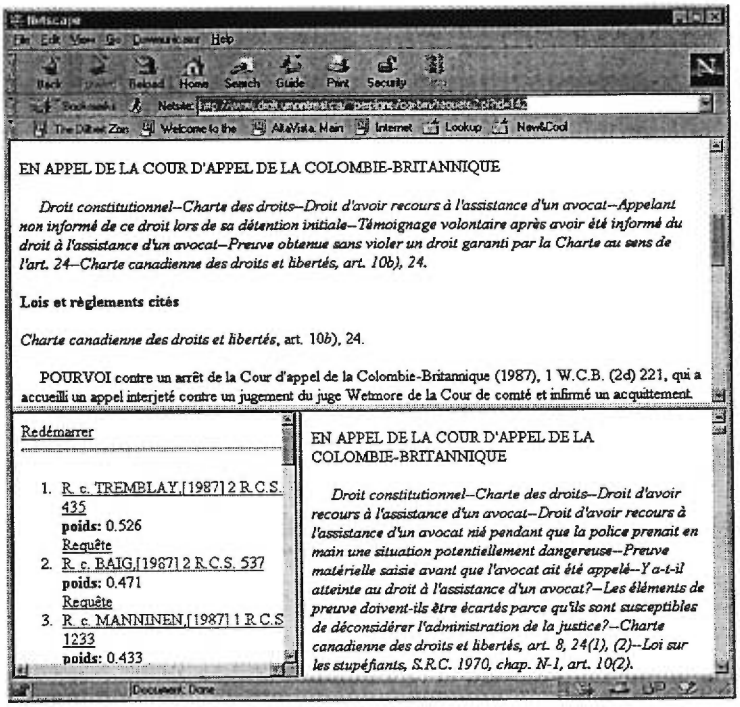

Figure 15 Citations et similarité

Lorsque les références et les citations rentrent en jeu simultanément, il plus difficile de mesurer la performance du système. En effet, les résultats des tests que nous avons effectués ne suivent pas une tendance définie. Les requêtes 81, 70 et 170, n'ont pas retourné beaucoup de décisions similaires, tandis que d'autres, telles que la 180, la 353 et la 436, au contraire, ont généré plusieurs décisions similaires parmi lesquelles au moins 70% sont fortement similaires à la décision initiale. Alors que ces dernières requêtes sont des preuves quasi irréfutables que les liens bibliographiques peuvent servir à identifier le rapprochement du contenu des décisions, l'interprétation des résultats obtenus avec les premières est plus problématique. Il est vrai que les degrés de similarité des décisions

retournées en réponse aux requêtes 81, 70 et 170 sont faibles; en moyenne, ces décisions ont un degré de similarité de 0.166 tandis que celle des requêtes 180, 353 et 436 atteint 0.422. En général, nous avons remarqué que le degré de similarité des vecteurs bibliographiques ne dépasse 0.26 pour des décisions dont le contenu ne se rapproche pas. Toutefois, le nombre limité d'expériences effectuées jusqu'à présent ne nous permet pas d'établir un seuil relativement au degré de similarité des vecteurs bibliographiques à partir duquel nous pourrions être certains de la similarité des décisions. En effet, nous avons remarqué que les décisions *R. c. Pozniak*, [1994] 3 RCS 310 et *R. c. Smith*, [1991] 1 RCS 714 partagent une certaine similarité mais que le degré de similarité de leur vecteur n'est que de 0.192. Un examen attentif révèle toutefois que dans l'affaire *Smith* d'autres problématiques étaient en cause et qu'ainsi la similarité révélée par le premier groupe de termes vedettes (droit à un avocat) se trouvait atténués par la présence d'autres groupes de termes vedettes référant à des questions fort différentes (droit criminel, meurtre).

Nous avons mené d'autres expériences afin de tester un autre aspect également important de notre outil, à savoir la navigation. Nous voulions connaître l'effet produit par la génération d'une suite de contextes de navigation à partir d'une décision initiale. Nous entendons par génération de suites de contexte de navigation le fait de choisir l'une des décisions retournées par le système en réponse à une requête comme nouvelle décision initiale pour une autre requête et ainsi de suite. Nous voulions déterminer dans quelle mesure ce procédé éloigne l'utilisateur du contexte initial, c'est-à-dire génère de nouvelles décisions qui ne sont pas similaires à la toute première décision initiale, ou sinon le rapproche en fournissant des décisions de plus en plus similaires à la décision

d'origine. Nous présentons au paragraphe suivant les résultats que nous avons obtenus lors de deux expériences de navigation.

Nous avons choisi les requêtes 81 et 436 comme points de départ pour chaque navigation. Ces deux requêtes représentent respectivement le pire et le meilleur cas obtenus lors de notre première expérience. Nous avons constaté que la navigation issue des deux requêtes produisait des résultats assez stables dans le sens qu'un certain nombre de décisions revenaient toujours dans les réponses du système, même si elles n'apparaissaient pas forcément dans le même ordre. Ceci est une indication que le système regroupe les décisions partageant certains concepts juridiques en agglomérations. Cependant, la structure de ces agglomérations semble différente.

Partant de la requête 81, nous avons constaté que plusieurs concepts rattachés à l'article 2 de la Charte tels le « droit à la vie », la « liberté d'expression », la « liberté de religion » et la « liberté d'association » coexistaient si bien que nous ne sommes pas arrivés à trouver un concept principal sous lequel nous pouvions regrouper toutes les décisions générées. Lorsque nous choisissions une décision traitant du « droit à la vie » ou du « droit familial », concepts abordés dans la décision 81, la constitution de l'agglomération générée ne changeait pas et nous ne nous éloignons pas trop de la décision originale. Cependant, les degrés de similarité demeuraient faibles. Par ailleurs, si nous suivions les décisions traitant de la « liberté d'association », concept absent des termes vedettes de la décision 81, nous constations que nous nous éloignons petit à petit des concepts de la décision originale au profit de ceux des nouvelles décisions que nous choisissions tandis que les degrés de similarité augmentaient.

La requête 436 donne lieu à une navigation totalement différente. La quasi totalité des décisions retournées traitait du « droit à l'assistance à un avocat », thème principal de la décision originale. Pour tous les résultats générés, le degré de similarité des décisions demeure relativement élevé. Nous ne pouvons pas vraiment nous éloigner de la décision initiale étant donné que la majorité des décisions générées lui étaient similaires. Cependant, nous avons remarqué qu'en suivant les décisions possédant les degrés de similarité les moins élevés, de nouvelles décisions venaient s'ajouter à l'agglomération tandis qu'en suivant celles qui affichaient les degrés de similarité les plus élevés la composition de l'agglomération ne changeait pas vraiment et dans bien des cas nous tournions en rond. La "Figure 16" montre le contexte de navigation issu à partir de celui de la requête 436 en choisissant la décision *R. c. Matheson*, [1994] 3 R.C.S. 328. On constate que la décision *R. c. Harper*, [1994] 3 R.C.S. 343 affiche un degré de similarité plus élevé que celui généré pour la requête 436 dont le contexte de navigation est montré à la "Figure 13".

Cette exploration des possibilités de navigation nous amène à conclure que le système exprime fidèlement les agglomérations de décisions similaires. Certains groupes de décisions entretiennent des liens étroits comme le montre la navigation à partir de la décision 436, alors que d'autres, comme illustrés par la navigation issue de la décision 81, froment des regroupements plus lâches, offrant la possibilité de bifurquer vers d'autres agglomérations.

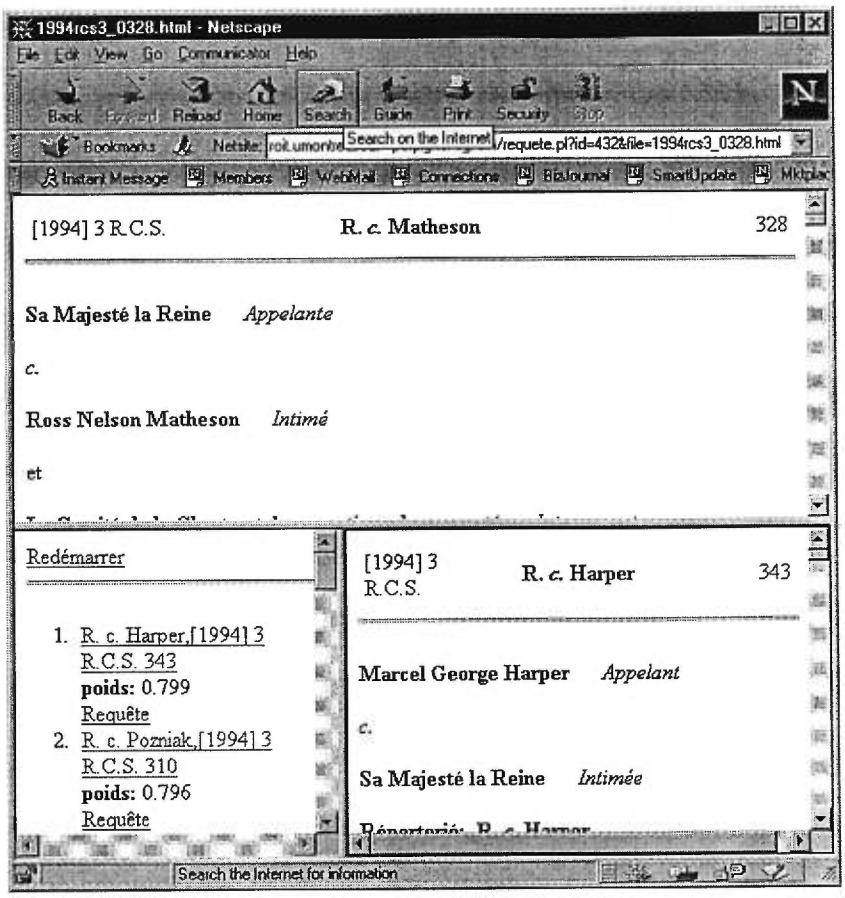

Figure 16 Contexte de navigation de la décision Matheson

3.2 Les travaux futurs

Malgré les résultats très satisfaisants obtenus avec le prototype, nous ne pouvons nous empêcher de penser aux améliorations qui peuvent y être apportées. Deux types viennent spontanément à l'esprit. Il y a d'un côté celles visant à améliorer les différentes composantes du système. De l'autre, se trouvent celles visant à expérimenter avec des liens bibliographiques autres que les références et les citations.

La composante d'indexation est, sans doute, celle à laquelle il y aurait le plus d'améliorations à apporter. Tout d'abord, il arrive souvent que les items bibliographiques n'apparaissent pas sous leur forme complète, mais sont plutôt constitué du nom d'une des parties impliquées dans le jugement précédé d'expressions du type : « l'arrêt » ou « dans l'arrêt ». Par exemple, une référence au jugement *R. c. Pozniak,* [1994] 1 R.C.S. 310 peut être tout simplement désigné par l'expression « dans l'arrêt Pozniak », lorsque la référence complète a été fournie précédemment dans le texte. Nous pensons qu'il est possible de repérer et d'identifier ces items bibliographiques incomplets et d'en tenir compte lors de l'indexation. Il s'agit d'extraire le mot suivant l'expression indiquant une référence, dans le cas de l'exemple l'indicateur et le mot sont respectivement « dans l'arrêt » et « Pozniak », et ensuite de vérifier si la liste d'items bibliographiques rencontrés jusqu'à présent dans le texte contient un item contenant le mot extrait. Dans des cas plus compliqués où, par exemple, le nom des partis comporte plus d'un mot, l'algorithme d'identification devra être plus élaboré. Ces améliorations permettront au système de tirer le maximum de profit des informations bibliographiques contenues dans chaque décision du corpus et de représenter plus fidèlement les décisions. Ensuite, les décisions de la Cour suprême du Canada ne contiennent pas uniquement des références à des décisions de cette même cour. Nous y trouvons également des références à des jugements d'autres tribunaux canadiens et parfois même à des tribunaux étrangers. Elles contiennent enfin des références à la législation et à la doctrine pertinente. Ces items bibliographiques constituent eux aussi de bonnes sources d'informations dont le système actuel ne tient aucun compte. Reconnaître ces autres types d'items bibliographiques ne pose aucun problème de principe puisqu'il suffit de construire les patterns permettant de les identifier au moment du traitement de mise en forme des fichiers.

Encore au plan de l'amélioration des modules actuels du prototype, nous pourrions envisager d'offrir à l'usager outre la navigation à partir d'une décision du corpus, la formulation d'une requête, c'est-à-dire, de construire un vecteur bibliographique en spécifiant les liens références et les citations que contient, selon lui, une décision répondant à ces attentes. Avec un tel ajout, le système pourrait récupérer de sa base de données les décisions dont les vecteurs bibliographiques sont similaires à celui fourni par l'utilisateur. Ce nouveau type de requête permettrait à l'utilisateur de combiner à loisir les liens bibliographiques et le libérerait des requêtes prédéterminées du corpus. Toutefois, l'interface du système doit demeurer conviviale de telle sorte que l'utilisateur ne soit pas obligé de mémoriser complètement les items bibliographiques. Pour ce faire, il nous semble possible de concevoir des mécanismes assez simples permettant à l'utilisateur de fournir, par exemple, le nom des parties en cause et qui, à partir de ces indices, déduiraient les items bibliographiques. Dans le cas où plusieurs items bibliographiques correspondraient aux indices, le système pourrait proposer une liste à partir de laquelle l'utilisateur choisirait les items qui lui apparaissent important.

Toujours au plan des améliorations des modules existants, nous envisageons des changements plus fondamentaux touchant le modèle lui-même. Les expériences menées semblent suggérer que l'apport des liens de références dans la mesure de similarité doive être séparé de celui des liens de citations. Plus formellement, en supposant qu'une décision est caractérisée par deux vecteurs r_i et c_i associés respectivement aux liens de références et de citations, la nouvelle mesure de similarité pourrait prendre la forme suivante :

$$sim(d_i, d_j) = \frac{\alpha \cdot \cos(r_i, r_j) + \beta \cdot \cos(c_i, c_j)}{\alpha + \beta}$$

où α et β sont des paramètres compris entre 0 et 1 que l'utilisateur spécifierait au moment de fournir sa requête. Ce schéma lui permettrait d'accorder plus d'importance à l'un ou à l'autre des types de liens selon qu'il s'intéresse davantage à l'histoire du traitement d'un concept ou plutôt à son traitement subséquent. De plus, la fonction du cosinus pourrait éventuellement être remplacée par une autre mesure de similarité plus appropriée aux liens bibliographiques. Les travaux antérieurs déjà mentionnés fournissent quelques pistes dans ce sens.

De façon alternative, nous pourrons également utiliser d'autres liens tels que le couplage bibliographique et la cocitation. Bien qu'ils aient été utilisés avec succès comme mesures de similarité, il n'est pas sûr a priori que leur utilisation dans un système comme le nôtre produira de meilleurs résultats. Aussi nous faudra-t-il concevoir des procédés d'évaluation et de comparaison plus objectives afin d'estimer les bénéfices ou désavantages engendrés par leur introduction dans le système.

Conclusion

Dans ce mémoire nous avons examiné une autre approche pour la représentation des documents à des fins de recherche d'information. Cette approche consiste à utiliser les items bibliographiques, couramment appelés références, comme descripteurs de document à la place des termes. Ce faisant, nous caractérisons chaque document par les liens bibliographiques que ces items établissent entre les différents documents constituants le corpus. C'est le rôle primordial que jouent ces liens dans certains corpus, tels que les corpus juridiques, qui nous a poussé à croire qu'une telle approche est réalisable et qu'elle recèle un grand potentiel pour la recherche d'information. De plus, les liens bibliographiques sont avantagés par rapport aux termes, car leur traitement est plus simple. Enfin, et c'est sans doute le plus important, les items bibliographiques possèdent davantage de caractéristiques pondérables.

Nous avons donc élaboré un modèle qui, à l'exception du choix des descripteurs, est identique au modèle vectoriel classique de Salton. Le prototype que nous avons développé à partir de ce modèle a donné des résultats positifs. Les décisions retournées par le système sont similaires entre elles et forment des agglomérats autour des concepts juridiques traités dans le corpus. Ces agglomérats, à leur tour, permettent une navigation plus efficace et cohérente à l'intérieur du corpus. De plus, certaines décisions traitant de plusieurs concepts juridiques permettent de passer d'un agglomérat à un autre dont le concept dominant est différent de celui du premier. Ainsi, les expériences menées nous ont permis de vérifier de façon tout à fait satisfaisante notre hypothèse quant à l'intérêt des liens bibliographiques comme des descripteurs de documents et leur utilité à des fins

de repérage de l'information dans le domaine juridique. L'idée est sans doute valable aussi pour d'autres domaines, cependant nous devons nous attendre à ce que, dépendant du rôle tenu par les items bibliographiques dans ces domaines, l'approche aie plus ou moins de succès.

Les choix techniques que nous avons faits pour la réalisation de ce projet se sont révélés adéquats. Bien que nous n'ayons pas utilisé toutes les propriétés du SGML, les balises insérées dans les fichiers ont énormément facilité l'identification des items bibliographiques par le module d'indexation. Le langage Omnimark™ nous est apparu comme le meilleur choix pour traiter initialement les fichiers afin de les transformer dans le format voulu. Le choix d'utiliser Java comme langage de développement s'est révélé satisfaisant également, malgré que nous ayons eu à programmer beaucoup de classes utilitaires afin de nous doter des structures de données nécessaires à notre projet car celles-ci ne faisaient pas partie des classes expédiées avec le langage. Le CGI nous a permis de construire aisément une interface usager répondant à tous les besoins du système et facilitant les expériences que nous voulions effectuer. Cependant, avec les récents développements du Web, nous pensons que l'utilisation de « servlets » java, types d'application (applet) s'exécutant sur le serveur, améliorerait la performance de notre système. En effet, la technologie sur laquelle se base les servlets permet une gestion plus efficace des ressources (mémoire allouée, entrées sorties) du serveur Web. De plus, le fait que ces « servlets » soient programmés en Java nous assurerait une meilleure intégration avec les autres modules du système.

Par ailleurs, le prototype développé ne tire pas profit du plein potentiel de notre approche et la section précédente mentionne quelques avenues qu'il nous reste à explorer.

Notamment, il est possible de concevoir une mesure de similarité qui permet de distinguer l'apport des références de celui des citations. Un système basé sur une telle mesure donne plus de flexibilité à un utilisateur en lui permettant d'accorder plus d'importance au traitement subséquent d'une cause qu'à son traitement précédent ou vice versa. Il est également possible d'utiliser des liens bibliographiques autres que les références et citations tels que la cocitation ou le couplage bibliographique comme descripteurs de documents. Finalement, un algorithme d'indexation plus méticuleux qui repère les items bibliographiques implicites est envisageable.

Par ailleurs, remarquons que compte tenu du rôle de plus en plus important joué par le Web dans notre vie de tous les jours, nous ne pouvons nous empêcher de penser à l'adaptation de notre approche pour les liens hypertextes. Bien que cet exercice ne pose pas de problèmes techniques insurmontables, il nous apparaît prématuré de prédire pour l'instant le succès ou l'échec d'une telle entreprise. En effet, le Web constitue un très vaste espace informationnel, seule l'expérimentation pourra clarifier le rôle qu'y jouent les liens hypertextes et leur utilisation. Rappelons-le, c'est la cohérence de ces éléments dans les corpus juridiques qui a permis le développement fructueux de notre approche dans le présent projet. Cependant, le fait même que le Web soit si vaste confère une grande importance aux liens bibliographiques entre les documents. Cette grande importance peut se révéler favorable à l'utilisation de notre approche dans le contexte du Web. De plus, le « World Wide Web Consortium », organisme chargé de développer les standard pour le Web a adopté le XML (eXtensible Markup Language) comme successeur au HTML. Le XML [W3C 97] est en quelque sorte une version simplifiée du SGML et promet de combler beaucoup de lacunes du HTML. Entre autre, de concert avec

le XLL (eXtensible Link Language), le XML améliore l'actuel mécanisme des liens hypertextes. Une de ces améliorations consiste à doter ces liens d'un attribut décrivant le type relation impliquant les documents reliés. Nous pensons que notre système adapté pour le Web pourrait tirer profit de ce genre d'attributs. Toutefois, il n'est pas sûr que notre approche sera aussi fructueuse pour le Web en général qu'elle l'a été pour le corpus de la Cour Suprême. Nous ne pourrons avoir le cœur net qu'après avoir adapté le système pour le Web et réalisé des expériences.

Bibliographie

[Broadus 83] Broadus, R. N. « An Investigation of the validity of bibliographic citations », dans Journal of the American Society for Information Science 34(2), 1983, pp. 132-135.

[Frakes & Baeza-Yates 92] Frakes, W. B., Baeza-Yates, R. « Information Retrieval : Data Structures and Algorithms », Prentice-Hall Inc., New Jersey, 1992.

[Kaplan 65] Kaplan, N. « The norms of citation behavior: prolegomena to the footnote », dans American Documentation 16(3), 1965, pp. 178-184.

[Kessler 63] Kessler, M. M. « Bibliographic coupling between scientific papers », dans American Documentation 14(1), 1963, pp.10-25.

[Kiralfi 90] Kiralfi, A. K. R. « The English legal system », Sweet & Maxwell, 1990.

[Kochen 74] Kochen, M. « Principles of Information Retrieval », Los Angeles: Melville, 1974.

[Kuhn 73] Kuhn, T. S. « The structure of scientific revolutions », 2nd edition, University of Chicago Press, Chicago, 1973.

[Garfield 79] Garfield, E. « Citation indexing: its theory application in science, technology and humanities », Wiley, New-York, 1979.

[Gilbert 77] Gilbert, G. N. « Referencing as persuasion », dans Social

Studies of Science 7(1), pp. 113-122.

[Lawani & Bayer 83] Lawani, S. and Bayer, A. E. « Validity of citation criteria for assessing influence of scientific publications: new evidence with peer assessment », dans Journal of the American Society for Information Science 34(1), 1983, pp. 59-66.

[Liu 93] Liu M. « Progress in documentation. The complexities of citation practice: A review of citation studies », dans Journal of Documentation 49(4), 1993, pp. 370-408.

[Lluelles 95] Lluelles, D. « Guide des références pour la rédaction juridique », Éditions Thémis, Montréal, 1995.

[Martyn 75] Martyn, J. « Citation Analysis », dans Journal of Documentation 31(4), 1975, pp. 290-297.

[McGill 93] Revue de droit de McGill, « Manuel canadien de la référence juridique », 3ème édition, Carswell, 1992.

[Merton 73] Merton , R. K. « The sociology of science : theoritical and empirical investigations », Universty of Chicago Press, Chicago, 1973.

[Mitra 70] Mitra, A. C. « The bibliographical reference: a review of its role », dans Annals of Library Science and Documentation 17(3), 1977, p 117.

[Narin 76] Narin, F. « Evaluative bibliometrics: the use of publication and citation analysis in the evaluation of scientific activity », Cherry

Hill : Computer Horizons, 1976

[Nomoto 90] Nomoto, K., Wakayama, S., Kirimoto, T., Ohashi, Y., Kondo, M. A. « Document retrieval based on citation using fuzzy graphs », dans Fuzzy Sets and Systems 38(1), 1990, pp. 207-222.

[Price 86] Price, D. « Little Science, big science... and beyond », Columbia University Press, New York, 1986

[Ravetz 71] Ravetz, J. R. « Scientific knowledge and its social problems », Clarendon, Oxford, 1971.

[Salton 83] Salton, G., McGill, M. J., « Introduction to Modern Information Retrieval », McGraw-Hill, New-York, 1983

[Savoy 96] Savoy, J., « Citation Schemes in Hypertext Information Retrieval », dans Information Retrieval and Hypertext, 1996, pp. 99-120.

[Small 73] Small, H., « Cocitation in the scientific litterature : a new mesure of the relationship between two documents », dans Journal of the American Society for Information Retrieval 24, 1973, pp. 265-269

[Smith 81] Smith, L. C. « Citation analysis », dans Library Trends 30(1), 1981, pp. 83-106

[Tapper 84] Tapper, C., « An experiment with citation vectors », dans Data

Processing and the Law, 1984, pp. 90-104

[W3C 97] http://www.w3.org/

[Waddams 97] Waddams, S. M. « Intrduction to the study of law », Carswel, 1992.