Université de Montréal

Une approche CBR textuel de réponse au courrier électronique

par

Luc Lamontagne

Département d'informatique et de recherche opérationnelle

Faculté des arts et des sciences

Thèse présentée à la faculté des études supérieures
en vue de l'obtention du grade de
Philosophiae Doctor (Ph.D.)
en informatique

Université de Montréal

Faculté des études supérieures

Cette thèse intitulée :

Une approche CBR textuel de réponse au courrier électronique

présentée par :

Luc Lamontagne

a été évaluée par un jury composé des personnes suivantes :

_____

Esma Aïmeur (Président-rapporteur)

_____

Guy Lapalme (Directeur de recherche)

_____

David Leake (Examinateur externe)

_____

Philippe Langlais (Membre du jury)

Thèse acceptée le : _____

Sommaire

L'insertion de systèmes de gestion de courriels dans les entreprises devient une nécessité en raison de la popularité de ce mode de communication et de l'accroissement du volume de messages échangés. Dans cette thèse, nous présentons une approche pour la mise en oeuvre d'un système intelligent de réponse au courrier électronique. Notre approche mise sur la réutilisation de messages antécédents lors de la construction de nouvelles réponses. Notre démarche de recherche s'appuie sur des techniques de raisonnement à base de cas textuels (CBR textuel), un formalisme de l'intelligence artificielle qui vise à résoudre de nouveaux problèmes à partir d'expériences passées décrites dans des documents. La réalisation du module CBR de réponse implique des choix au niveau de la construction de la base de cas, de la phase de sélection de messages antécédents, et de la réutilisation du contenu des messages sélectionnés.

Nous effectuons au chapitre 2 une revue bibliographique des principaux travaux dans le domaine du CBR textuel et nous comparons les différentes approches proposées dans la littérature. Cette étude, portant sur la technologie et sur les tâches qu'elle résout, nous a permis d'identifier des voies de recherche que nous avons poursuivies dans cette thèse.

Au chapitre 3, nous présentons les algorithmes que nous utilisons lors de la phase de recherche CBR. Notre stratégie consiste à incorporer des descriptions de réponses (au lieu des requêtes) pour guider le choix du message antécédent le plus pertinent. Ce choix est justifié par l'uniformité des descriptions de réponses qui rend leurs comparaisons plus pertinentes que celle des requêtes. Pour réaliser cette étape, nous avons étudié et évalué deux approches statistiques basées sur des modèles de cooccurrences et de traduction. Nos résultats indiquent que ces deux approches apportent des performances supérieures à un schéma de type *tf*idf*. D'autres techniques font également l'objet d'une discussion. En complément, nous menons au chapitre 4 une évaluation du processus de construction du module CBR. Nous proposons une démarche pour quantifier l'impact des différentes décisions prises lors de la conception de la base de cas et nous définissons une métrique pour effectuer le choix d'un modèle avant la mise en opération du module CBR.

Finalement, nous explorons au chapitre 5 la réutilisation du contenu des réponses sélectionnées. Nos travaux s'inspirent des systèmes de type "canevas de réponse" basés sur des textes prédéterminés comportant des portions à remplir par l'usager. Notre approche consiste à déterminer dynamiquement les portions de textes qui devraient être retirées (passages optionnels) ou modifiées (passages variables) par l'usager. Nous étudions deux techniques pour la sélection des portions optionnelles et nous comparons leurs performances sur un sous-ensemble de notre corpus. Nous décrivons également des travaux que nous avons menés sur l'identification de portions de textes variables à partir de techniques d'extraction d'entités nommées.

Nous concluons au chapitre 6 en résumant les motivations de ce travail, les objectifs, les contributions et nous proposons des avenues de recherche pour la poursuite de travaux futurs.

Mots clés : raisonnement à base de cas, réponse au courrier électronique, cooccurrences de mots, modèle de traduction, traitement de la langue naturelle.

Abstract

The use of email management systems by organizations is becoming a necessity due to the popularity of this communication mode and the increasing volume of messages being exchanged. In this thesis, we propose an approach for implementing an intelligent email response system. Our approach is based on the reuse of antecedent messages for the synthesis of new responses. The framework underlying our research is textual case-based reasoning (textual CBR), a formalism of artificial intelligence that aims to solve new problems based on past experiences contained in documents. The implementation of a CBR response module involves making choices regarding the authoring of the case base, the retrieval of antecedent messages and the reuse of the content of selected messages.

We present in chapter 2 a review of the main research in the textual CBR literature and a comparison of the various proposed approaches. This technology-based and task-based comparative study allowed us to identify research avenues that we explored in this thesis.

In chapter 3, we present algorithms we used in the retrieval phase of the CBR cycle. Our strategy consists of incorporating the responses' descriptions (instead of the requests) in order to guide the selection of most pertinent message. This strategy is justified by our observation of the uniformity in the response descriptions which makes their comparison more pertinent than that of the requests. To support this strategy, we studied and evaluated two statistical approaches based on word co-occurrences and translation models. Our results indicate that these approaches yielded a performance superior to the *tf*idf* approach. In this chapter, we also discuss other techniques for addressing this problem. As a complement, we conduct in chapter 4 an evaluation of the CBR module authoring process. We propose an approach for quantifying the impact of the various decisions made during the construction of the case base and define a metric to help in choosing a model prior to developing an operational CBR module. Finally, we explore in chapter 5 the possibility of reusing the content of the selected responses. Our work is similar to that of template-based systems that use predetermined texts containing

portions to be filled in by the user. Our approach consists of determining dynamically the text portions to be removed (optional passages) or modified (variable passages) by the user. We investigate two techniques for the selection of the optional portions and compare their performances on a subset of our corpus. We also describe work that we conducted in order to identify the variable text portions using extraction techniques for named entities. We conclude in chapter 6 by summarizing the motivations behind this work, the objective, the contributions and we propose future research topics.

Keywords : case-based reasoning, email response, word co-occurrences, translation models, natural language processing.

Remerciements

Je remercie vivement le professeur Guy Lapalme de l'Université de Montréal d'avoir accepté de diriger cette thèse. Ses conseils et sa confiance tout au long de ce travail m'ont permis de persévérer.

Je remercie les professeurs David Leake du département d'informatique du *Indiana University*, Philippe Langlais, et Esma Aïmeur qui a agi à titre de présidente-rapporteuse du jury, tous deux du département d'informatique et de recherche opérationnelle de l'Université de Montréal, d'avoir bien voulu participer au jury. Les commentaires de tous les membres du jury furent très pertinents et très appréciés, et je les en remercie.

Je garderai un très bon souvenir des deux autres Luc et de Julien Dubois qui ont participé au projet Mercure. Merci également à Oscar Nilo pour les nombreuses discussions et pour ses connaissances inépuisables.

Je remercie mes parents de m'avoir appris par l'exemple, le courage et la persévérance.

Finalement, je suis infiniment reconnaissant envers mon épouse Irène Abi-Zeid, qui m'a encouragé et soutenu tout au long de cette thèse, et sans qui ce travail n'aurait jamais été possible. Merci aussi à Mathieu et Valérie, qui m'ont inspiré et m'ont donné la volonté d'aller jusqu'au bout.

Québec, juin 2004

Le financement de cette thèse a été assuré par les Laboratoires Universitaires Bell (LUB).

Table des matières

Liste des figures

Liste des tableaux

Chapitre 1 . Introduction

La tâche de réponse au courrier électronique consiste à synthétiser un texte suite à une requête. Dans un contexte d'entreprise, l'utilisation de technologies de l'information permet d'alléger cette tâche et de réduire les temps de traitement. Des techniques de l'Intelligence Artificielle (IA) et du Traitement Automatique des Langues Naturelles (TAL) peuvent être mises à contribution pour automatiser certaines fonctionnalités de cette tâche. L'objectif de nos travaux est l'élaboration d'une approche qui permet la réutilisation des messages antécédents pour formuler de nouvelles réponses. Pour y arriver, nous adoptons une approche de l'IA, le raisonnement à base de cas (CBR), qui exploite des solutions passées pour résoudre de nouveaux problèmes. Dans ce chapitre, nous décrivons les motivations et les fondements de nos travaux et nous positionnons notre approche par rapport au domaine du CBR.

La gestion du courrier électronique est une activité importante sur les plans commercial et technique pour les entreprises. On prévoit que plus de six trillions de messages électroniques seront échangés cette année dont plus de 20% échangés pour des fins commerciales, ce qui donne une indication du rôle important joué par ce mode de communication au sein des entreprises.

De plus en plus, on note un accroissement de l'utilisation du courriel en entreprise. En effet, pour pallier l'augmentation du nombre de requêtes dans les centres d'appels, une large part des communications actuellement effectuées par voie téléphonique est redirigée vers des moyens électroniques comme le courriel et les services de *chat*. Étant donné qu'une forte proportion d'entreprises[1] n'est pas préparée pour gérer adéquatement le volume de courrier découlant de l'interaction avec leur clientèle, l'insertion de nouvelles technologies d'information est la solution envisagée pour faire face à ce volume de messages tout en maintenant la qualité du service à la clientèle.

Les technologies de l'information permettent de prendre en charge différentes fonctions de gestion de courriels. Dans nos recherches, nous nous intéressons plus particulièrement à la tâche de réponse, c.-à-d. celle qui consiste à formuler un texte suite à

[1] En 2000, Gartner Group (Gartner Research 2000) estimait que seulement 10% des entreprises étaient adéquatement préparées.

un message de requête. Ces travaux ont été menés dans le cadre du projet Mercure (Lapalme et Kosseim 2003) de l'Université de Montréal qui vise à étudier différentes architectures pour la réponse automatique aux messages électroniques. Durant la première phase du projet (Kosseim *et al*. 2001), des expérimentations ont permis d'apprécier le potentiel et les limitations d'une combinaison de techniques de traitement automatique des langues naturelles, l'extraction d'information et la génération de texte, pour aborder cette tâche. L'approche que nous adoptons pour cette phase est l'application du raisonnement à base de cas (CBR) pour générer des réponses aux nouveaux messages électroniques (Lamontagne et Lapalme 2003a, Lamontagne et Lapalme 2003b). En général, le CBR tente de résoudre de nouveaux problèmes en exploitant des expériences passées. Pour notre application, les expériences correspondent aux messages de requêtes et aux réponses qui leur sont associées. Une approche CBR de réponse au courriel consiste donc à se remémorer des messages similaires et à modifier des réponses utilisées auparavant.

D'un point de vue technique, la réponse au courrier électronique représente un défi intéressant. Cette tâche devrait être idéalement supportée par des systèmes qui combinent une forme de reconnaissance et de génération de texte. Or, la robustesse des techniques actuelles n'est pas suffisante pour aborder un tel problème et pour envisager un déploiement de systèmes commerciaux. D'autres directions devront donc être étudiées. Nous avons choisi le raisonnement à base de cas, une approche que nous jugeons prometteuse pour cette tâche. La conception d'un module de réponse CBR repose sur un corpus de messages antécédents, une ressource qui est représentative du domaine du discours et des différents problèmes traités durant les échanges de courriels. L'approche CBR permet de réduire les efforts d'acquisition et de modélisation de connaissance, des activités qui constituent habituellement un goulot d'étranglement dans la conception de systèmes à base de connaissance. De plus, le cycle de raisonnement CBR repose principalement sur les processus de recherche (*retrieval*) et d'adaptation. Ce découpage du raisonnement correspond bien au processus de réponse, lequel repose sur l'analyse des messages de requête et la synthèse de réponses pertinentes.

1.1 Motivations et résumé de la thèse

Dans notre thèse, nous proposons et vérifions que la réutilisation de messages, par l'entremise d'un processus de raisonnement à base de cas textuels, est une voie efficace pour la synthèse de réponses de courrier électronique.

L'objectif de cette recherche est donc de proposer une approche CBR pour mettre en oeuvre un module de réponse au courriel. Le résultat final de ces travaux est un groupe de techniques pour spécialiser le cycle de raisonnement CBR afin de traiter des cas comportant des descriptions textuelles. Ces spécialisations portent sur a) l'utilisation des descriptions de solutions (c.-à-d. les réponses) durant la phase de recherche, b) une stratégie et des mesures pour la construction de la base de cas, et c) une technique pour la réutilisation de solutions antécédentes. Ces travaux sont décrits plus en détail dans les chapitres suivants.

La base de nos travaux est que le cycle de raisonnement à base de cas offre un paradigme viable et efficace pour la synthèse de nouvelles réponses au courriel. Nous avons observé, à partir d'un corpus de messages pour une application de service à la clientèle, que plusieurs réponses comportent des extraits provenant de messages antécédents. Ceci suggère que les préposés exécutant manuellement cette tâche se servent fréquemment d'exemples passés pour répondre à de nouvelles requêtes, ce qui confirme la pertinence des techniques CBR pour rehausser la réutilisation de portions de texte.

L'acquisition des ressources nécessaires à la construction d'un module CBR de réponse est déjà prise en charge par les logiciels actuels. Ces systèmes de courriel permettent de préserver les messages par l'entremise des boîtes de courrier. Comme l'accumulation des expériences ne pose pas de problèmes particuliers, les efforts de recherche ont porté en un premier temps sur la sélection et la réutilisation de ces messages.

Afin de réaliser un outil d'aide à la réutilisation de messages antécédents, nous avons proposé des schémas de raisonnement adaptés à nos ressources. Les

caractéristiques des courriels ne permettent pas une exploitation directe et facile par le CBR structurel car ces courriels sont des documents peu structurés. Ceci exige que le cycle de raisonnement à base de cas soit spécialisé pour exploiter adéquatement leur contenu. Traditionnellement, les systèmes CBR s'appuient sur des expériences bien structurées et dont la réutilisation est maîtrisée. Toutefois les systèmes et approches traditionnels sont déficients lorsque les expériences, relatées par des textes, se prêtent plus difficilement à une représentation structurée.

Afin de surmonter cette difficulté, nous positionnons nos travaux dans un cadre textuel, afin de préserver et d'exploiter le corps textuel de messages. Des travaux récents, regroupés sous la bannière CBR textuel, ont proposé des extensions aux systèmes CBR traditionnels afin d'exploiter des expériences contenues dans des documents textuels (Lamontagne et Lapalme 2002). Bien que ces travaux servent de base intéressante à notre recherche, nos recherches visent à les enrichir et en combler certaines lacunes. Plus particulièrement, nous utilisons le contenu textuel des solutions pour guider le cycle de raisonnement CBR, i.e. pour estimer la similarité entre les cas, pour réutiliser des portions de cas antécédents et pour construire la base de cas initiale du système de réponse. Cette perspective contribue aux techniques actuelles du CBR textuel qui reposent principalement sur le contenu des descriptions de problème.

Dans les prochains paragraphes, nous présentons des approches préconisées pour la réponse aux courriels. Nous décrivons les propriétés du corpus de messages que nous utilisons pour nos travaux. Nous illustrons par la suite l'approche CBR que nous proposons pour réutiliser des messages antécédents. Finalement, nous décrivons nos principales motivations de recherche portant sur la construction, la recherche et la réutilisation de messages antécédents.

1.2 La réponse au courrier électronique

Le développement de systèmes de gestion du courrier électronique dans les entreprises a connu une progression importante au cours des dernières années. Les premiers systèmes de gestion de réponses sont apparus vers 1997. Récemment, le

domaine a été très volatile avec un grand nombre d'acquisition de compagnies (*Inference*, *Aptex*, *EchoMail*, *Genesys*…) qui œuvraient dans ce champ d'activité. On estime que ce secteur d'activité a atteint en 2002 des revenus de $210 millions répartis sur une clientèle de 25,000 entreprises. Des systèmes commerciaux, dont les systèmes *KanaResponse* et *Primus FirePond*, sont utilisés actuellement pour effectuer le suivi de messages.

Le processus de suivi d'un message, tel que géré par ces logiciels, comporte les étapes suivantes :

- la catégorisation qui permet de classifier les messages et d'identifier les boîtes de courrier correspondantes dans lesquelles les messages seront déposés. Les boîtes correspondent généralement à différents thèmes du modèle d'affaire ;
- le routage qui redirige les messages vers la personne la plus apte à les traiter ;
- la mise en priorité qui détermine le degré d'urgence du traitement des requêtes ;
- la formulation de réponse aux requêtes contenues dans les messages en entrée.

Nous avons répertorié des approches pour automatiser les trois premières étapes (par ex. Segal & Kephart 2000, Dubois 2002, Manco *et al*. 2002, Wang *et al*. 1999, Jordan 2001, Koole & Mandelbaum 2002, Koole *et al*. 2003). Toutefois, nous avons concentré nos travaux sur la quatrième étape du processus de suivi d'un message, car la formulation de réponse a fait l'objet de peu d'étude dans la littérature scientifique. Nous présentons dans la section suivante un survol des approches préconisées par les logiciels pour supporter cette tâche.

1.2.1 L'automatisation de la fonctionnalité de réponse

Une approche populaire dans les logiciels client courriel actuel est la fonction d'auto-réponse, c.-à-d. un système de règles qui gèrent l'envoi automatique de textes prédéterminés. Les règles de réponse sont déclenchées lorsque l'on retrouve dans l'en-tête des messages certaines adresses courriels ou certains mots-clés. Cette fonctionnalité

est une forme élémentaire de réponse automatique. Elle est utile pour acheminer des accusés de réception ou pour indiquer qu'un suivi est en cours. Toutefois elle est trop rigide pour s'adapter à un contenu de requêtes varié ou pour être personnalisée en fonction de l'émetteur du message.

Les systèmes de réponse au courrier électronique visent à supporter plus activement le traitement des messages. Pour ce faire, plusieurs approches sont envisagées dont :

- la sélection de réponses prédéterminées : le contenu du message fait l'objet d'un traitement de type recherche ou de classification. Des réponses rédigées à priori et indexées par des thèmes ou par des regroupements de mots-clés sont utilisées ;

- l'utilisation de canevas de réponse : un canevas est une esquisse de texte qui comporte des portions pouvant varier en fonction du contenu du message ; un canevas offre plus de flexibilité que les réponses complètement prédéterminées mais nécessite une intervention humaine pour compléter la rédaction de la réponse;

- la génération complète de réponse : l'approche générative est théoriquement la plus flexible et la plus complète. Elle est toutefois difficile à mettre en oeuvre et elle est coûteuse en temps de calcul. De plus, elle dépend d'une représentation structurée de la requête qui doit être réalisée par des logiciels de compréhension de texte. L'approche générative demeure expérimentale pour l'instant.

La plupart des systèmes sur le marché appartiennent aux deux premières approches et font usage de classification automatique pour la sélection des réponses/canevas. Le Tableau 1 donne un survol des quelques logiciels que nous retrouvons dans la littérature et des techniques qui y sont préconisées pour aider les utilisateurs à répondre aux courriels.

Le défi d'une automatisation "intelligente" de la fonction de réponse est de limiter l'utilisation de messages dont le contenu est complètement déterminé a priori et de tenter de formuler des réponses qui sont adaptées dynamiquement au contexte du contenu des messages. Par exemple, le système *KanaResponse* mise sur la classification automatique pour choisir les portions d'un canevas qui répondent le mieux à la situation (http://www.kana.com/solution/servicesolution/response/). Le classificateur est constitué d'un ensemble de règles construites manuellement par le gestionnaire du système. Les canevas sont également rédigés manuellement.

ReplyMate	Des réponses prédéterminées sont associées à des thèmes présentés sous forme de question. La sélection des réponses est manuelle. (www.replymate.com)
Cisco	Des réponses prédéterminées sont associées aux feuilles d'un arbre de décision. Les branches correspondent à des mots-clés du texte. L'arbre peut être construit par classification. (www.cisco.com/warp/public/180/email/)
SERiMail	Des réponses prédéterminées sont associées à des thèmes ; la classification du contenu des requêtes permet de faire la sélection d'une réponse. Si une classification ne peut être faite, on cherche une requête similaire et on suggère la solution correspondante. (www.ser.com)
KanaResponse	Des canevas de réponses sont structurés avec des choix de phrases, et la sélection des énoncés est effectuée à partir de règles. (www.kana.com)
YY software	Une analyse linguistique du message permet de reconnaître le contenu des requêtes. Le système utilise des grammaires HPSG (*head-driven phrase structure grammar*) pour faire cette analyse. Des réponses sont associées aux interprétations du système. La compagnie a récemment cessé ses activités.
XtraMind	Une catégorisation des requêtes est effectuée par la combinaison des règles obtenues à partir de leur représentation vectorielle, par des distributions n-grammes et par un apprentissage de type *boost-end*. (www.xtramind.com)

Tableau 1 : Quelques systèmes de réponse aux courriels

1.2.2 Contexte d'utilisation et bénéfices anticipés

En premier lieu, il s'agit de déterminer le contexte dans lequel les systèmes de réponses sont susceptibles d'apporter des bénéfices. Devrait-on miser sur une

distribution grand public ou sur des applications bien ciblées ? Il est évident que l'utilisation de systèmes de réponse n'est pas justifiée pour des fins personnelles dont les situations sont variées et les échanges sont spontanés. Il en résulte des textes improvisés qui sont informels et qui n'observent aucune structuration particulière. Il pourrait s'avérer utile parfois de réutiliser des portions de textes mais il serait difficile de justifier les coûts d'achat du logiciel et l'ajustement de ses composantes pour une utilisation plutôt sporadique.

Par ailleurs, une automatisation du processus de réponse dans les entreprises, principalement pour les applications de service à la clientèle, peut apporter de nombreux bénéfices. On retrouve plusieurs études qui démontrent que le temps d'attente moyen pour obtenir une réponse (estimé en moyenne à 95.38 heures selon un sondage de la firme *Emagicon*) pourrait être grandement réduit par l'utilisation de technologies de l'information.

Le courrier électronique est principalement utilisé par les entreprises pour des fins de promotion (messages *outbound*) et pour le support à la clientèle (messages *inbound*). Les systèmes de gestion des messages *inbound* offrent de nombreux avantages. Ils permettent la réduction du temps de réaction (jusqu'à 1000 messages par jours) et la diminution des coûts (3$ par message au lieu de 53$ pour une réponse complètement manuelle). De plus, leur utilisation entraîne une augmentation de la productivité et une diminution de la redondance car il facilite le traitement des messages répétitifs, la distribution du travail entre les préposés, et un meilleur suivi des messages. Ces systèmes aident les entreprises à réduire le nombre de requêtes qui demeurent sans réponse[2]. Cette solution offre un mode de communication plus flexible pour les requêtes moins urgentes. Elle permet également d'accumuler plus d'information sur la clientèle. Le coût de ces logiciels, qui varie entre $30 000 entre $100 000, est amplement justifié pour les entreprises qui traitent un volume élevé de messages. Toutefois, ces systèmes exigent des efforts considérables pour leur mise en fonction et leur maintenance (acquisition de règles d'affaires, des canevas de messages, etc.).

Lorsqu'on analyse les avantages de ces systèmes tels que décrits dans la littérature, il est difficile de distinguer ceux qui découlent de la fonction de réponse de ceux qui résultent des autres fonctions du système (catégorisation, routage, ordonnancement en priorité). Une des contributions importantes d'un support à la réponse est la qualité des réponses et de leur prestation. La qualité d'une réponse a sûrement une influence sur la perception du client de la qualité du produit ou du service offert par l'entreprise. Un module de réponse offre une plus grande uniformité et une consistance dans le contenu des messages, ce qui favorise également la clarté des réponses. Son utilisation réduit les fautes de frappe. Il permet également de réduire les malentendus et les ambiguïtés qui entraînent une augmentation du nombre de messages échangés. Finalement la préservation des messages favorise la propagation et la réutilisation de "bonnes recettes" au sein de l'entreprise.

---

2 Selon l'Actualité du 1 octobre 2003, le nombre de jours qui s'écoule avant d'obtenir une réponse à un courriel envoyé aux 100 plus grandes entreprises américaines observe la distribution suivante: a) deux jours

D'autres aspects liés à l'insertion de technologie de réponse semblent plus difficiles à gérer. Par exemple, cette approche n'assure pas de la concision ou de la pertinence des textes proposés. De plus, il se peut que des portions d'une requête ne soient pas adéquatement abordées. Finalement, la fréquence de mise à jour d'un module de réponse qui utilise des ressources fixes (textes prédéterminés, canevas, règles) risque d'être déficiente si elle repose sur une maintenance manuelle.

1.3 Cadre applicatif

Nos recherches ont porté sur un corpus de messages échangés dans le cadre du domaine du service aux investisseurs (*investor relations*), c.-à-d. le processus par lequel une compagnie communique avec ses investisseurs. Dans ce domaine, le courrier électronique est utilisé par les entreprises de deux manières. Premièrement, les messages *outbound* sont distribués par la compagnie pour faire la promotion d'événements corporatifs ou de résultats financiers. Deuxièmement, la plupart des compagnies ont, sur leur site web, une section de service aux investisseurs à partir de laquelle les analystes financiers peuvent être contactés pour fournir de l'aide aux investisseurs (messages *inbound*). L'information disponible sur ces sites porte sur différents sujets tels que la valeur des actions et les rapports financiers. Ces services sont d'une importance significative car la qualité de l'information joue un rôle important dans les décisions des investisseurs professionnels. Dans les prochains paragraphes, nous présentons de quelques points reliés au domaine, au corpus et à son exploitation.

1.3.1 Quelques caractéristiques du domaine étudié

Les messages acheminés aux analystes de Bell Canada Enterprise (BCE), l'entreprise qui nous a fourni le corpus, portent principalement sur les indicateurs financiers de BCE, sur la valeur de son titre boursier et sur les dates de divulgation des rapports financiers (par ex. des rapports, appels conférences). Plus particulièrement, on note des questions ayant trait aux aspects suivants :

(58%), b) trois jours (6%), c) quatre jours (6%) et d) aucune réponse (31%).

- Caractéristiques de la compagnie : le lien entre BCE et ses différentes filiales, l'actionnariat de la compagnie et la composition du titre de BCE. Ces messages sont peu fréquents.

- Résultats financiers : pour un trimestre donné, on demande une description des principaux indicateurs financiers de la compagnie, tels que les bénéfices (*earnings*), les dividendes et leurs variations par rapport aux trimestres précédents.

- Performance boursière : des requêtes portent sur la valeur du titre en bourse et la variation de ce titre par rapport aux indices boursiers (par ex. l'indice TSE). Des demandes d'explication de la performance en bourse du titre de BCE ou des filiales sont également effectuées.

- Sources d'informations : il s'agit de questions sur les moyens utilisés par BCE pour communiquer des nouvelles de la compagnie ou les derniers résultats financiers. Par exemple, des investisseurs veulent obtenir des copies de rapports financiers ou se faire ajouter à une liste de distribution. On retrouve aussi des requêtes sur l'obtention de la date et des coordonnées de différents événements (divulgation de rapports, appels conférence téléphoniques, rencontres d'actionnaires).

De plus, notre corpus contient des messages portant sur le niveau de taxation, les débentures, les pratiques comptables, la mise à jour des comptes personnels, la définition de terminologie financière ainsi que des plaintes à propos du site web.

1.3.2 Quelques caractéristiques du corpus de messages utilisé pour les recherches

Notre corpus est constitué de plus de 1500 messages intrants (*inbound*). Ce corpus comporte des requêtes soumises par des investisseurs et des réponses formulées

par les analystes. Nous avons identifié un certain nombre de caractéristiques qui ont un impact sur la sélection d'une approche CBR adéquate :

a) Taille des messages : la plupart des messages ont moins de 100 mots (en incluant la signature des messages électroniques). La longueur des messages varie entre un seul terme[3] (l'adresse électronique) et 178 termes. La longueur moyenne est de 57 termes. La plupart des réponses sont brèves (en moyenne 28 mots), les plus longues comportant des explications fournies par l'analyste. Les réponses sont écrites par un nombre limité d'analystes (5-10) et sont uniformes au niveau du format et des structures des réponses, lesquelles sont répétitives voire même quasiment identiques (par ex. les dates de divulgation de résultats financiers). En revanche, puisque les questions proviennent de différentes personnes, le style de rédaction varie d'une question à l'autre. On retrouve donc différentes formulations pour une même requête (paraphrase). En général, les messages sont bien rédigés et contiennent peu de fautes d'orthographe. Le contenu est clair, le vocabulaire est précis et les phrases sont correctement structurées. On retrouve très peu de négations de propositions dans les messages. La plupart des messages sont rédigés en anglais, et les autres en français.

b) Structure des requêtes : une question est habituellement constituée de trois parties :

- Une en-tête contenant la date, l'adresse de l'expéditeur et le sujet du message ;
- une brève description du contexte, Par exemple, quelques justifications de l'envoie du message comme "*I am considering investing in your company*" ou "*I am conducting an analysis of your stock*". Rarement retrouve-t-on une description détaillée du problème auquel fait face l'investisseur.
- une ou plusieurs questions portant sur l'un des sujets décrits dans la section précédente sur les caractéristiques du domaine,

[3] Nous désignons par *terme* les mots, les symboles, les quantités et les autres unités de base lexicales (par ex. un URL ou une adresse électronique) que l'on retrouve dans un texte.

- les coordonnées de l'investisseur : cette portion contient des informations telles que nom, affiliation, adresses postale et électronique, signature. Les réponses contiennent des explications suivies d'une phrase de courtoisie à la fin du message. Le message est toujours personnalisé (utilisation du nom de l'investisseur).

c) *Séquence de messages* : pour des messages échangés sur une période de 10 mois, on ne retrouve pratiquement aucune suite d'échanges multiples entre un analyste et un investisseur donné (ce phénomène fut observé à une seule reprise). Chaque question est, en principe, posée par une personne différente. La plupart des requêtes individuelles contiennent suffisamment d'information pour que l'analyste puissent formuler adéquatement une réponse. Nous faisons donc l'hypothèse d'indépendance entre les messages.

d) *Temporalité* : le contenu d'un message fait référence à des périodes de temps spécifiques. Par exemple, les trimestres financiers peuvent être décrit explicitement (*the third*) ou implicitement (*next, previous*). La date dans l'en-tête du message aide à expliciter les dates implicites.

e) *Requêtes multiples :* plusieurs messages contiennent plus d'une question. Par exemple, l'investisseur peut demander une copie du dernier rapport financier et, en plus, d'être ajouté à la liste de distribution. D'autres demandent une même information sur les différentes filiales ou plusieurs indicateurs financiers pour d'une même compagnie. Cette multiplicité des requêtes est une caractéristique importante qui distingue les courriels et des contenus de foires aux questions (*frequently-asked questions - FAQ*) lesquels ne comportent qu'un seul sujet.

f) *Spécificité des messages* : le degré de spécificité des requêtes et des réponses varie grandement. Quelques requêtes sont génériques du type "Pourquoi devrais-je investir dans BCE ?". A l'opposé, d'autres sont très spécifiques tel que "*since 1999 earnings for BCE are $8.35 per share and Nortel's are -$0.23, I would assume that after the spin-off, BCE earnings should be something more than $8.50 per share...*". Pour leur

part, les réponses n'abordent pas toujours directement le contenu de la requête. Par exemple, on redirige souvent l'investisseur vers le site web en indiquant que de nombreuses informations y sont disponibles. De plus, les requêtes spéculatives sur les fluctuations du marché boursier font l'objet d'une réponse de courtoisie. On note également des "méta-réponses", c.-à-d. des réponses ne portant pas directement sur le contenu de la question mais plutôt sur la nature du message. Par exemple "ce que vous demandez représente beaucoup d'information".

1.4 Opportunités pour le raisonnement à base de cas

La mise en correspondance du CBR et de la réponse au courriel offre des opportunités d'enrichissement mutuel. Le problème de réponse peut bénéficier d'une approche CBR puisqu'il existe une correspondance entre le cycle de raisonnement CBR et le processus de réponse tel que pratiqué par les préposés dans les centres de service. Nous pouvons donc nous attendre à ce que les modèles que nous proposons soient efficaces pour supporter la tâche de réponse.

Par ailleurs, le CBR peut également bénéficier de son application à des tâches de réponses. Tel que nous le décrivons en détail au chapitre 2 de ce document, les approches CBR présentent des limitations majeures lorsque les expériences sont textuelles. Or la tâche de réponse au courriel constitue une métaphore intéressante pour repousser les frontières du raisonnement à base de cas textuel. Les courriels présentent des caractéristiques qui représentent des défis pour la sélection de cas similaires. Le problème de la réponse est une tâche qui favorise l'étude de la réutilisation de solutions textuelles.

Dans les prochains paragraphes, nous discutons de la pertinence de l'approche CBR, des particularités de son applicabilité à la tâche de réponse et des principales limitations que nous avons eues à confronter.

1.4.1 Adéquation du CBR

Le raisonnement à base de cas est en principe bien adapté à la fonction de réponse car :

a) il est fréquemment appliqué à des tâches connexes comme le service à la clientèle,

b) il peut être intégré sous différentes formes aux opérations de l'entreprise, et

c) de nouvelles réponses peuvent être réutilisées sans modification de connaissances du système.

Un premier point qui motive l'utilisation du raisonnement à base de cas est son application largement répandue pour le service à la clientèle de type *help desk*. Ces systèmes servent souvent à guider les conversations téléphoniques des préposés à la clientèle ou encore à consulter des informations de type "*frequently-asked questions*" (*FAQ*) à partir de sites web. L'extension de ces systèmes au service à la clientèle par courrier électronique est donc naturelle. Plusieurs questions étant répétitives, l'utilisation des réponses précédentes pour répondre à de nouvelles questions est tout indiquée.

Suite à notre recherche bibliographique, nous avons répertorié une seule publication sur la gestion de réponse au courrier électronique dans la littérature CBR scientifique. Un système (Cheetham 2003) développé par *General Electric Global Research* a été proposé pour gérer la sauvegarde du courrier interne et la sélection de réponses antécédentes. Les réponses similaires sont regroupées dans les feuilles d'un arbre de décision. Des mots-clés sont associés aux nœuds de l'arbre et permettent de récupérer les réponses lors du traitement de nouvelles requêtes. Cette approche est intéressante d'un point de vue pratique. L'utilisation de cette approche a amené une réduction de coûts de 40%. Toutefois, d'un point de vue technique, elle se limite à la phase de recherche de message et ne propose par d'approches pour structurer ou traiter le contenu des réponses. La construction de l'arbre de décision, l'affectation des mots-clés

aux réponses et le parcours dans l'arbre de décision (mots clés) sont effectués manuellement.

On note également l'application de techniques CBR pour l'analyse de messages électroniques afin de détecter de nouvelles descriptions de problème (Shimazu et Kusui 2001). Ces travaux misent sur l'accroissement de mots-clé comme indicateurs de problèmes potentiels. Des travaux ont également été menés par cette équipe sur le partage de "mail folders" et de "bulletin boards" convertis en base de cas (Kusui et Shimazu 2001).

Il existe, par ailleurs, dans le domaine commercial deux systèmes de gestion de réponse qui utilisent des techniques CBR : *eGain* et *FirePond*. Ces outils utilisent un modèle de raisonnement CBR conversationnel (modèle présenté à la section 2.2 de ce document). De plus, plusieurs des compagnies œuvrant dans ce domaine utilisent des documents de type FAQ, c.-à-d. des questions fréquentes auxquelles on associe une réponse standard générée manuellement. Malheureusement, leurs documentations techniques fournissent peu de précision sur les mécanismes de gestion des FAQ. Plusieurs logiciels offrent des fonctions d'analyse statistique pour évaluer la performance du processus de gestion de réponse et pour caractériser les besoins et intérêts de la clientèle.

On pourrait considérer différentes conceptions de module qui respectent les principes de l'approche CBR. Le choix d'un modèle de raisonnement dépend de l'étendue du traitement à accomplir et du degré d'intégration entre le module et le processus de réponse. Trois aspects caractérisent le niveau de support offert par le module CBR de réponse :

- L'aspect « glaneur » : le module permet la sélection de textes existants qui peuvent être récupérés lors de la composition de nouvelles réponses. Ces systèmes permettent de sauvegarder, d'indexer et de rechercher des messages passés. Ces fonctions peuvent être complètement manuelles ou reposer sur des fonctions de repérage plus évoluées. La réutilisation des réponses repose

sur des opérations de copier-coller sans toutefois bénéficier de l'assistance du module dans la formulation des réponses. L'aspect glaneur correspond en gros aux différents systèmes avec réponses prédéterminées que nous avons décrits auparavant.

- L'aspect « rédacteur » : le module a la capacité de recommander des réponses ajustées au contexte de la requête. Les ajustements correspondent à des modifications de messages pour différentes fins tels que la personnalisation du contenu, le retrait/ajout de passages ou le maintien de la véracité des énoncés.

- L'aspect interactif : pour apporter ses recommandations, le module mise sur des échanges avec l'usager. Le niveau d'autonomie peut aller d'un raisonnement sans aucune intervention humaine jusqu'à la supervision complète par l'usager de chacune des étapes du raisonnement. Et on peut décliner différentes approches intermédiaires qui varient dans la sollicitation de l'usager et l'aiguillage du processus de réponse.

D'un point de vue CBR, le niveau glaneur correspond à un module qui se limite à repérer des expériences similaires (la fonction de *retrieval*). Le niveau rédacteur apporte un traitement supplémentaire du contenu des solutions (réponses) pour en favoriser une réutilisation efficace. Finalement, il existe dans la littérature CBR des approches qui proposent des interactions de type questions-réponses permettant de créer une synergie entre le système et son utilisateur. Plusieurs approches ont été proposées pour aborder ces aspects et nous passons en revue les principales d'entre elles au chapitre 2.

Le troisième point qui milite en faveur du CBR est l'adéquation de cette approche pour des domaines qui évoluent dynamiquement, contrairement aux approches de l'intelligence artificielle qui reposent sur des modèles prédéfinis de connaissance et qui exigent des modifications au système pour prendre en compte l'occurrence de nouvelles situations ou l'adoption de nouvelles solutions. Ces approches nécessitent des connaissances difficiles à acquérir qui peuvent devenir périmées rapidement pour des domaines qui évoluent vite. Par exemple, un système expert reposerait sur la définition

d'un ensemble de règles pour guider le choix ou la formulation des réponses, ce qui s'avère une tâche ardue. Que ces modèles de connaissance soit générés manuellement ou dynamiquement, il est difficile de prévoir toutes les questions possibles et de développer un modèle du domaine pour y répondre. De plus, une reconstruction du système est nécessaire pour incorporer ces nouveaux éléments. Or, les systèmes CBR offrent des capacités d'apprentissage *online*. Le système n'a pas à détenir toutes les réponses, car si une nouvelle situation survient, la nouvelle paire <*requête*, *réponse*> peut être immédiatement réutilisée par le système pour les requêtes subséquentes.

1.4.2 Particularités de la tâche d'un point de vue CBR

La tâche de réponse au courriel présente des particularités qui méritent d'être soulignées pour comprendre les choix que nous avons faits dans notre travail de recherche (Figure 1).

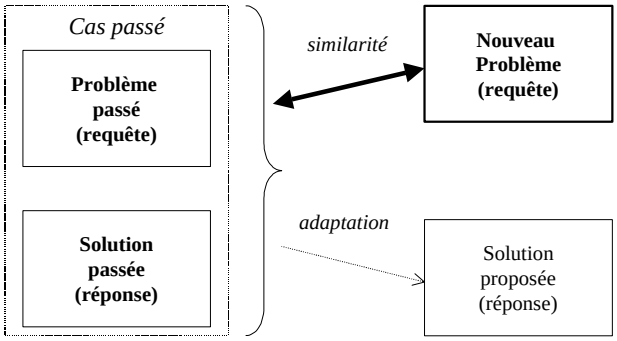


Figure 1 : Similarité et adaptation de cas

Plusieurs caractéristiques des courriels ont retenu notre attention. En premier lieu, on note la correspondance entre les messages et les cas d'un système CBR. La notion centrale d'un système CBR est la base de cas qui contient les différentes expériences passées. Une description d'expérience, c.-à-d. un cas, contient une portion problème et une portion solution. Pour la réponse au courrier électronique, l'approche la plus intuitive est de construire un cas à partir d'une requête (le problème) et d'une réponse (la solution). Toutefois, d'autres correspondances entre cas et messages sont possibles et ont été considérées au début de nos travaux. Par exemple, une solution pourrait être une agrégation des énoncés de plusieurs réponses similaires. L'agrégation permettrait de

réduire le nombre de cas du système. Mais la sélection des phrases devrait éviter les redondances et maintenir la cohérence du texte. Nous avons finalement retenu une relation un à un entre cas et messages car elle nous a semblé la plus naturelle tout en facilitant l'implantation du schéma de réutilisation que nous avons adopté.

Nous avons également observé que les requêtes portent souvent sur plus d'un sujet, ce qui rend la réutilisation de réponse antécédente plus difficile. Le choix d'une réponse passée devrait donc minimiser le déséquilibre entre les thèmes qu'elle contient et ceux de la nouvelle requête à traiter. Il pourrait arriver que la réponse passée omette d'aborder des passages de la nouvelle requête. Également, certains de ces énoncés de réponse pourraient ne pas être pertinents au contexte de la nouvelle requête. Il est donc souhaitable que le schéma de réutilisation permette d'établir une correspondance étroite entre les requêtes et les réponses. Puisque les descriptions de réponses et de requêtes correspondent à des séquences d'énoncés, des liens naturels peuvent être établis entre les énoncés de part et d'autre.

En plus des descriptions textuelles des courriels, on pourrait ajouter aux cas des informations structurées provenant de l'en-tête des messages. Pour notre domaine d'application, les différentes adresses électroniques (*sender*, *cc*, *bcc*) apportent peu d'information. Elles donnent parfois des précisions sur le nom de l'émetteur du message. Le sujet contient habituellement quelques mots qui décrivent à l'occasion l'un des thèmes du message. Ces termes sont souvent redondants avec ceux du contenu de la requête. On en déduit que le contenu de la réponse repose essentiellement sur les énoncés provenant du corps (*body part*) de la requête.

Les requêtes et réponses courriels sont habituellement rédigées par différentes personnes. Ceci peut entraîner des variations au niveau du vocabulaire, de la formulation des énoncés et de l'organisation des descriptions. L'approche préconisée doit donc être assez robuste pour réduire l'impact de ces variations.

La nature de la tâche de réponse nous a conduit à mener une étude sur la modification de solutions textuelles. Tel que mentionné précédemment, le raisonnement

à base de cas s'appuie principalement sur deux processus : la recherche d'expériences antérieures qui sont similaires (*retrieval*) et la modification de ces expériences (l'adaptation). Pour plusieurs systèmes, l'adaptation est manuelle, et n'est habituellement pas envisagée lorsque les expériences sont textuelles. Toutefois, nous croyons que notre application pourrait bénéficier de la modification des cas textuels. Ces modifications pourraient augmenter la qualité des solutions proposées et guider l'usager du système dans la rédaction de ses réponses.

1.4.3 Limitations des approches CBR

Les caractéristiques de notre application nous permettent d'identifier quelques difficultés que les approches actuelles de raisonnement à base de cas doivent surmonter. Ces limitations sont reliées à la similarité des cas lorsque les expériences sont décrites en langue naturelle. Le peu de structuration des descriptions complique l'évaluation des similarités entre problèmes. La richesse de la langue fait que des situations similaires peuvent être exprimées de manière totalement différente. On pourrait aborder ce problème à différents niveaux. La correspondance entre mots permet de capturer certaines similarités, mais elle demeure toutefois limitée par les ambiguïtés liées à l'utilisation des termes. Des approches, où l'on exploite la forme syntaxique où le contenu sémantique des textes, pourraient s'avérer avantageuses. Ces approches permettent d'obtenir une meilleure structuration des cas. Toutefois leur mise en oeuvre est limitée par le problème d'acquisition de connaissance et la disponibilité de ressources du domaine. De plus, il est difficile de préconiser ces approches pour des domaines plus vastes ou qui évoluent dynamiquement dans le temps.

Une autre limitation est que, dans le raisonnement à base de cas, la similarité des problèmes est garante de l'utilité des solutions. Dans un cadre textuel, la similarité des solutions ne découle pas nécessairement de la similarité des problèmes. La monotonicité de la relation problème-solution n'est pas assurée. Des solutions/problèmes similaires peuvent être exprimés différemment et donner lieu à des descriptions très différentes. Les textes peuvent varier selon le nombre d'auteurs, leurs expériences, leur rôle dans le

processus de résolution de problème et le contexte dans lequel les descriptions ont été formulées. Pour qu'un ensemble d'expériences soit exploitable, on s'attend à retrouver une homogénéité dans les descriptions. Toutefois, la bonne conduite de l'approche CBR repose sur l'homogénéité des problèmes. Cette hypothèse nous semble limitative lorsque l'homogénéité des solutions est plus importante. Un phénomène que nous avons observé dans nos descriptions de courriels.

Finalement, comme nous l'avons mentionné précédemment, la littérature CBR ne propose pas de modèle pour la réutilisation de solutions textuelles.

1.5 Approche préconisée

Dans nos travaux, nous avons développé le logiciel LUG[4], un module de raisonnement à base de cas qui permet la synthèse de nouveaux messages à partir de ceux rédigés par des analystes du domaine. Nous avons abordé le développement de ce module du point de vue d'un logiciel client de courrier électronique.

Pour initier le processus de réponse, notre logiciel client offre à l'usager la possibilité de consulter ses messages dans la fenêtre principale du client (voir Figure 2). Pour rédiger une nouvelle réponse, l'usager ouvre une fenêtre de composition et peut soit rédiger manuellement son message, soit obtenir une liste de paires <requête-réponse> qui sont jugées similaires par le module CBR (voir Figure 3). Ces recommandations sont présentées par ordre décroissant de similarité.

[4] LUG est le dieu correspond à Mercure dans la mythologie celtique.

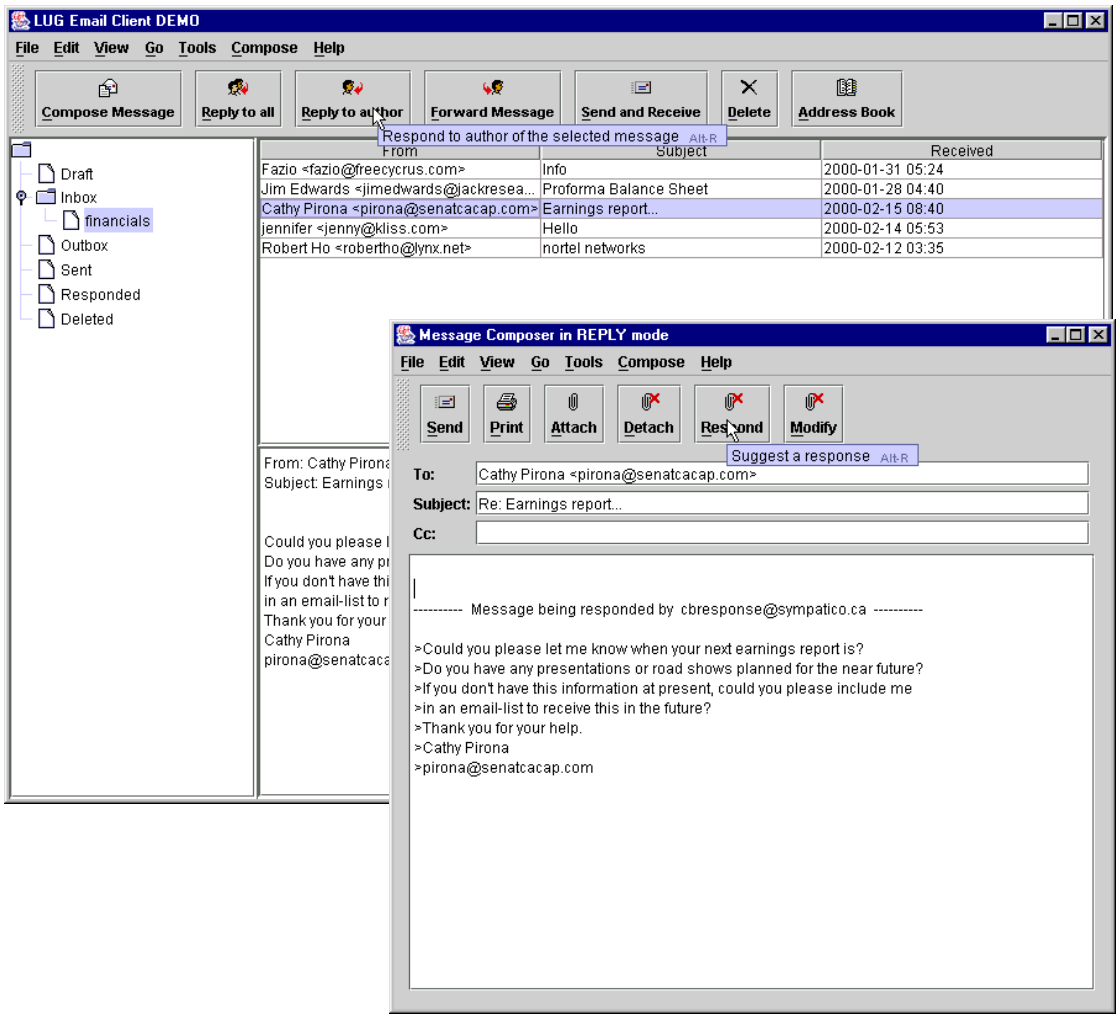


Figure 2 : Fenêtre principale du client courriel LUG

Si l'usager sélectionne l'un des messages recommandés, la portion réponse est recopiée dans la fenêtre de composition. Par la suite, l'usager peut obtenir des recommandations sur les portions de texte de cette réponse qui méritent d'être réutilisées (Figure 4). Par ce schéma d'interaction, l'usager peut continuer d'utiliser un environnement de travail avec lequel il est familier et faire appel au module CBR lorsqu'il en éprouve le besoin.

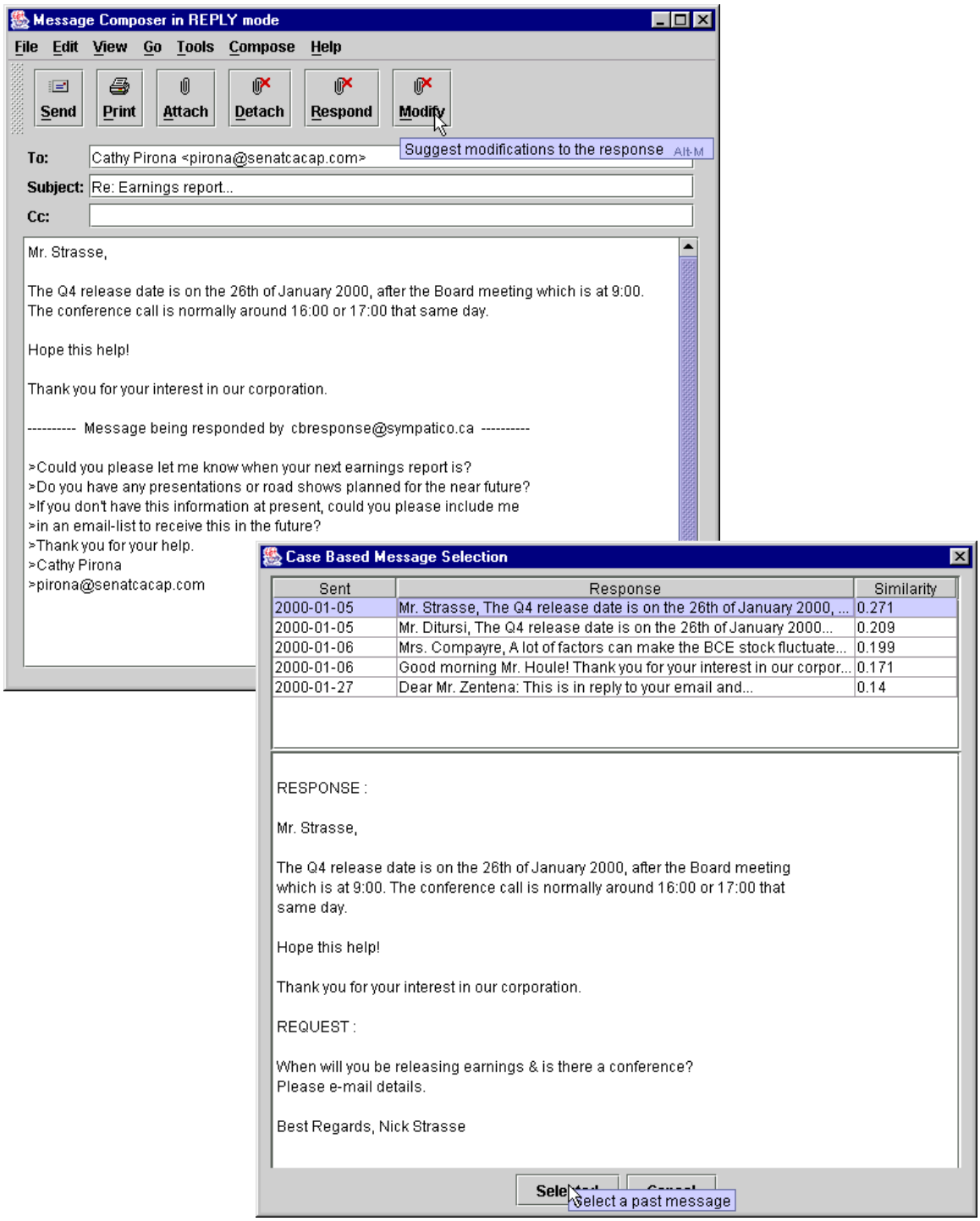


Figure 3 : Recommandations de messages similaires

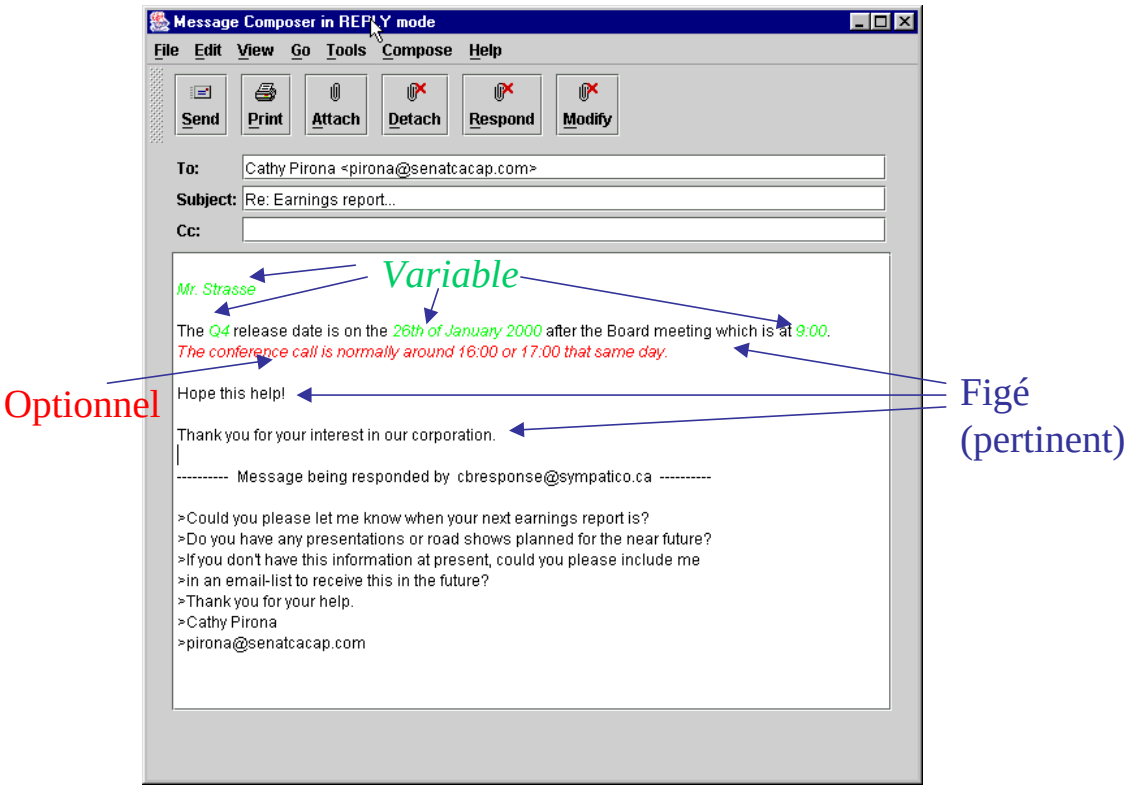


Figure 4 : Proposition de portions de message à modifier

Lorsqu'il est présenté selon cette perspective client, le module CBR tente de répondre à de nouveaux messages arrivant dans le INBOX du logiciel client en réutilisant une réponse ou des portions de réponses contenues dans la boîte de courrier SENT. Tel qu'illustré à la figure 5, nous avons étendu le schéma de raisonnement *search and adapt* pour prendre en compte les tâches suivantes : la recherche de messages à partir de requêtes comportant plusieurs thèmes, l'extraction et la substitution d'informations variables, et l'identification de passages optionnels dans les réponses. Les approches que nous proposons pour aborder ces aspects sont discutés aux chapitres 3, 4 et 5 de ce document. Ces fonctions sont sollicitées par le module client lorsque l'usager désire obtenir des recommandations sur les messages pertinents ou sur les portions de ces messages qui peuvent être réutilisées.

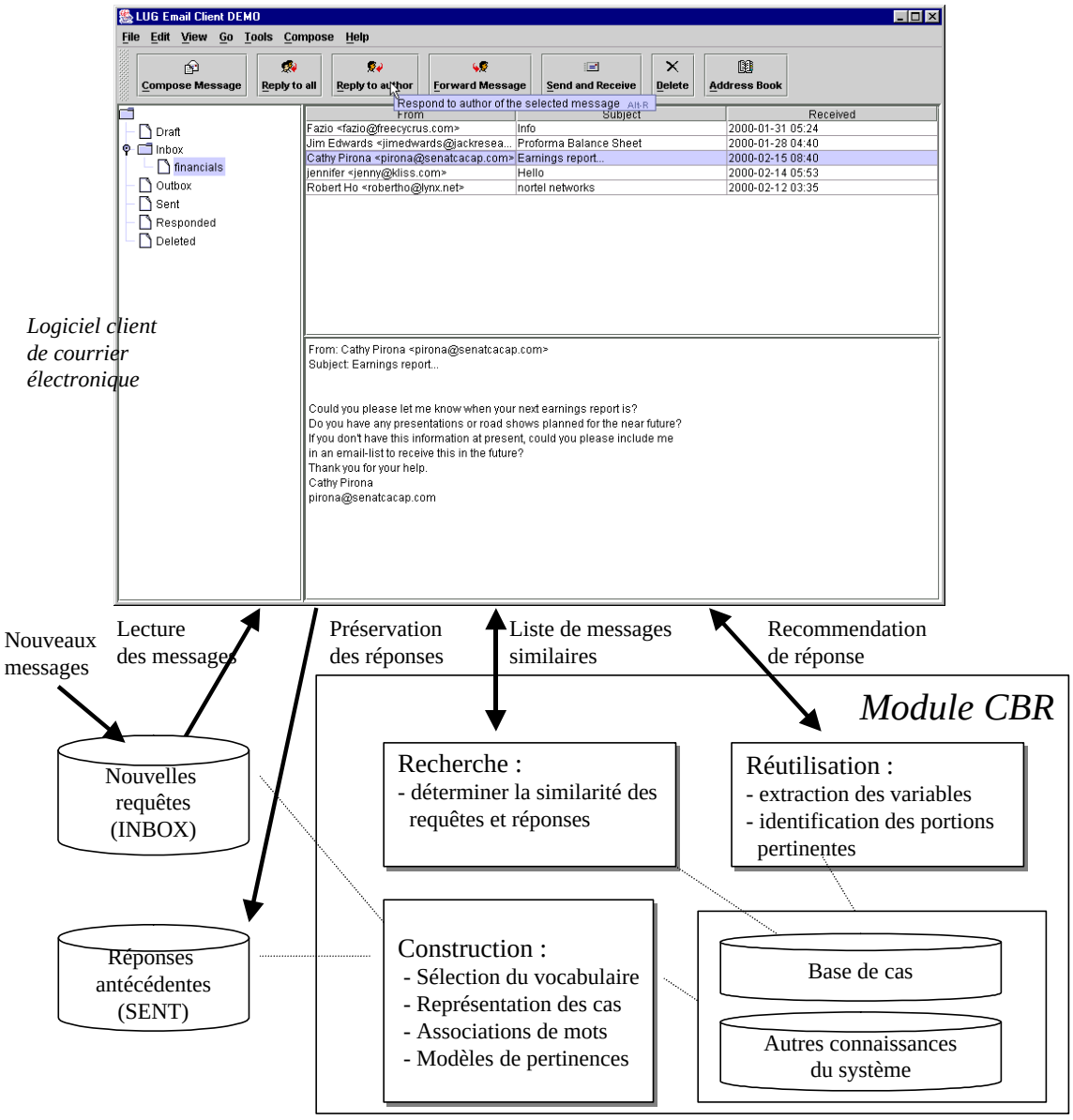


Figure 5 : Schéma de notre approche

Afin d'illustrer notre approche CBR, considérons la requête suivante reliée à la divulgation de résultats financiers :

***Requête*$_1$** : Can you tell me when you are reporting next. Thanks, Elliott.

Le système doit maintenant suggérer une réponse à la *Requête*$_1$. Supposons que le cas *j* (paire requête/réponse) de la figure 6 ait été proposé par le module CBR et sélectionné par l'usager. Étant données les caractéristiques de notre application, nous

devons par la suite identifier les passages pertinents de la *Réponse*$_j$ pour des fins de réutilisation.

Requête$_j$*:*	Hello Can you tell me when you will be releasing your Next earnings report also when your fiscal year ends Best Regards, Mark Strasse
Réponse$_j$*:*	Dear Mr. Strasse, The year ended on 31 December 1999. The release date for the next earnings report is on 26 January 2000. Please, do not hesitate to contact us for any other questions. Sincerely…

Figure 6 : Cas utilisé pour répondre à la *Requête*$_1$[5]

En utilisant le message *j*, le module CBR suggère que le passage portant sur la divulgation des résultats financiers est utile. Il établit également que le passage relié à la fin de l'année fiscale et la date du rapport trimestriel doivent être modifiés ou simplement élagués. Les passages à modifier ou élaguer sont indiqués par la notation «texte». Ainsi, suite aux recommandations proposées par le système, la réponse proposée est la suivante (Figure 7) :

Elliott

Réponse1: Dear «Mr. Strasse»,

«The year ended on 31 december 1999».

24 October 2001

The release date for the next earnings report is on «26 January 2000».

Please, do not hesitate to contact us for any other questions. Sincerely…

Figure 7 : Réponse proposée par le système

Nous avons eu à faire plusieurs choix pour mener notre étude. Nous ne voulions pas d'automatisation complète du processus de réponse pour éviter le piège des approches génératives. Nous avons plutôt cherché à offrir des pistes de réponses à l'usager qui soient de qualité supérieure à l'utilisation de messages pré-rédigés. Nous souhaitions également éviter que le nombre de situations que le système peut traiter soit déterminé a priori. Notre exemple précédent illustre le besoin de permettre à l'usager de repérer facilement une piste de réponse et de repousser autant que possible son intervention lors

[5] Les noms de personnes des messages originaux ont été modifiés.

de la phase de rédaction de réponse. Ceci implique que le module CBR doit donc offrir un niveau de précision permettant à l'usager de sélectionner facilement une base de réponse. Nous avons aussi le défi additionnel d'étendre le processus CBR au-delà de la phase de recherche et d'identifier des modifications pour améliorer la qualité des réponses.

Nous avons également limité le nombre de ressources nécessaires pour construire le module. La principale source fut le corpus de messages antécédents déjà géré par le système de courriel. Cette ressource fut la principale source des connaissances du module et de nos choix de techniques. Nous avons également limité les efforts humains alloués à l'acquisition de connaissance. On souhaite éviter ainsi un goulot d'étranglement qui empêcherait un déploiement à plus grande échelle des approches proposées dans cette recherche. Nous utilisons surtout des techniques qui reposent sur le contenu lexical des cas et que nous ne traitons pas les cas en fonction de leur contenu sémantique.

Nos travaux ont porté sur la construction initiale du système et sa mise en opération avec un premier jeu de ressources et modules. Nous n'avons pas abordé les problèmes de maintenance ou d'adaptation des paramètres du système.

1.6 Les résultats de nos travaux de recherche

Les principaux résultats de nos travaux de recherche se situent au niveau du raisonnement à base de cas textuels et de son application à la tâche de réponse. Premièrement, nous avons développé une application de réponse au courrier électronique dans lequel le raisonnement à base de cas joue un rôle essentiel. Le CBR s'est révélé une approche naturelle et suffisamment flexible. Deuxièmement, nous avons proposé des modèles et des algorithmes pour spécialiser les phases de recherche et de réutilisation du processus CBR textuel. Nos principales contributions sont les suivantes :

- Nous avons démontré qu'il est bénéfique d'estimer la similarité non pas uniquement à partir des descriptions de problèmes mais en tenant également compte des solutions (sections 3.1-3.2).

- Nous avons modélisé les relations entre les descriptions de problème et de solution par des associations de mots. Ces associations sont incorporées dans le processus de raisonnement à base de cas afin d'en améliorer les performances. Nous avons étudié et évalué deux modèles du domaine du traitement automatique de la langue, les modèles de cooccurrence et de traduction, pour capturer cette notion d'association (sections 3.3.-3.5).

- Nous avons proposé des mesures pour quantifier différentes décisions du processus de construction du module CBR. Nous avons proposé une métrique qui permet d'anticiper si l'utilisation des solutions des associations de mots apporte des avantages lors de la phase de recherche. Ces mesures comparent le recouvrement et la similarité relative des composantes d'un cas (chapitre 4).

- La tâche de réponse nous a incité à mener des travaux sur la réutilisation de solutions et nous avons exploré comment aider le rédacteur dans la modification des réponses antécédentes. Pour les cas peu structurés, le principal problème est de déterminer les portions de textes qui méritent d'être réutilisées ou modifiées. Nous avons abordé ce problème en déterminant des portions des messages qui semblent superflues et celles qui pourraient être inexactes. Deux approches pour déterminer les portions superflues ont été évaluées et comparées. De plus, nous avons mené quelques expérimentations pour identifier les portions de réponse qui pourrait en être modifiées en fonction du contexte de la nouvelle requête (chapitre 5).

L'idée principale dans nos travaux est qu'un message de réponse est une séquence d'énoncés construite en fonction des énoncés d'une requête. Un processus de réponse CBR propose des messages antécédents recouvrant adéquatement le contenu de la requête (la phase de recherche) et identifie les portions des messages où le recouvrement semble déficient (la phase d'adaptation). Des relations entre termes de requêtes et termes de solutions permettent de modéliser et de mesurer ce recouvrement. Notre approche CBR est donc une spécialisation des phases CBR dont le défi était d'intégrer ces relations dans les processus de raisonnement.

1.7 Résumé du contenu de la thèse

Nos travaux de recherche sont décrits dans les quatre prochains chapitres de cette thèse. Dans la prochaine section, nous présentons une brève description des principes du raisonnement à base de cas pour le lecteur non familier avec le domaine. Par la suite, nous passons en revue les principaux travaux du CBR textuel. Nous décrivons les différentes approches que nous avons identifiées dans la littérature et nous comparons leurs particularités, leurs avantages et leurs inconvénients. Aux chapitres 3 et 4, nous décrivons comment les descriptions de solutions sont exploitées dans la recherche de messages pertinents et nous élaborons une stratégie pour la construction du module CBR. Le chapitre 5 décrit comment le module guide l'usager dans la réutilisation des solutions antécédentes. Finalement, nous concluons par quelques remarques sur nos contributions et nous proposons quelques idées pour poursuivre les recherches en CBR textuel et son application pour la réponse au courrier électronique.

Chapitre 2 . Travaux en raisonnement à base de cas textuels

Dans ce chapitre, nous présentons les principes de base du raisonnement à base de cas et nous comparons les principaux travaux portant sur des cas provenant de documents textuels.

Traditionnellement le raisonnement à base de cas (CBR) s'appuie sur des expériences décrites dans des formats complètement structurés tels que des objets ou des enregistrements de base de données. Ce formalisme a permis au CBR de prendre un essor important au cours de la dernière décennie grâce, entre autre, à de nombreuses applications commerciales qui se sont avérées fructueuses. Toutefois les praticiens du domaine ont rapidement constaté les limites de cette approche structurelle et ont proposé d'autres modèles pour en surmonter les difficultés et étendre son application à des domaines plus variés.

Dans le cadre de nos recherches, nous nous intéressons plus particulièrement aux extensions du formalisme CBR pour traiter des expériences décrites dans des documents textuels, travaux regroupés sous la bannière CBR textuel. Ces approches sont relativement récentes (Lamontagne & Lapalme 2002) et s'appuient principalement sur des techniques de la recherche d'information, de l'apprentissage automatique et du traitement automatique des langues. La voie du CBR textuel s'avère nécessaire pour des applications, telles que celles du domaine de la jurisprudence légale ou du diagnostic médical, dont le raisonnement s'appuie sur des comptes-rendus textuels. Ces travaux sont également motivés par l'avènement des technologies web et l'émergence de pratiques de gestion de connaissance au sein des entreprises favorisant la préservation et l'exploitation d'expériences corporatives.

Dans ce chapitre, nous présentons l'état actuel des travaux de ce domaine. Nous débutons par une présentation succincte des principes généraux du CBR et des différentes familles de modèles de système CBR. Par la suite, nous décrivons les principaux travaux du CBR textuel et nous établissons un tableau comparatif de ces approches. Finalement

nous passons en revue quelques domaines techniques dont nous nous inspirons dans nos travaux de recherche.

2.1 Principes généraux du raisonnement à base de cas

Le raisonnement à base de cas (CBR) est une approche de résolution de problèmes qui utilise des expériences passées pour résoudre de nouveaux problèmes (Leake 1996). L'ensemble des expériences forme une base de cas. Typiquement un cas contient au moins deux parties : une description de situation représentant un "problème" et une "solution" utilisée pour remédier à cette situation. Parfois, le cas décrit également les conséquences résultant de l'application de la solution (par ex. succès ou échec). Les techniques CBR permettent de produire de nouvelles solutions en extrapolant sur les situations similaires au problème à résoudre. Cette approche est adéquate pour les domaines où la similarité entre les descriptions de problèmes nous donne une indication de l'utilité des solutions antécédentes.

Les fondements du CBR proviennent de travaux en sciences cognitives menés par Roger Schank et son équipe de recherche durant les années 80 (Riesbeck 1989). Leurs travaux ont mené à la théorie de la mémoire dynamique selon laquelle les processus cognitifs de compréhension, de mémorisation et d'apprentissage utilisent une même structure de mémoire. Cette structure, les *memory organization packets* (MOP), est représentée à l'aide de schémas de représentation de connaissance tels que des graphes conceptuels et des scripts.

Au début de la dernière décennie, on a assisté à un regain de popularité du domaine et de nouvelles tendances qui misent sur la simplification de la représentation des cas et sur des applications à plus grande échelle. Le CBR se révèle alors une précieuse technique pour la mise en œuvre d'applications commerciales (Watson 1998) pour différentes tâches telles que la résolution de problèmes (par ex. diagnostic, planification, design), les systèmes d'aide à la décision, les *help desk* et la gestion de connaissances. Ceci en fait l'une des techniques de l'intelligence artificielle les plus largement répandues actuellement.

L'approche CBR offre de nombreux avantages. Pour certaines applications, la démarche CBR est plus simple à mettre en œuvre que celles basées sur un modèle du domaine (par ex. une base de règles) ; elle permet d'éviter les problèmes d'acquisition de connaissance (*knowledge bottleneck*) qui rendent difficile la conception de bases de connaissances de taille importante. Le CBR est particulièrement bien adapté aux applications dont la tâche est accomplie par des humains expérimentés dans leur domaine et dont les expériences sont disponibles dans une base de données, dans des documents ou chez un expert humain. On l'utilise pour les domaines n'exigeant pas de solution optimale et dont les principes sont mal formalisés ou peu éprouvés.

2.1.1 Composantes d'un système à base de cas

Un système CBR est une combinaison de processus et de connaissances (*knowledge containers*) qui permettent de préserver et d'exploiter les expériences passées. Pour simplifier notre présentation, nous nous appuyons sur le modèle générique présenté dans la Figure 8. On y note comme principaux processus[6] la recherche (*retrieval*), l'adaptation (*reuse*), la maintenance (*retain*) et la construction (*authoring*)[7] et comme structures de connaissances le vocabulaire d'indexation, la base de cas, les métriques de similarité et les connaissances d'adaptation.

---

[6] Aamodt et Plaza (Aamodt 1994) ont proposé un modèle qui représente le raisonnement à base de cas comme un cycle comportant 4 processus: recherche-réutilisation-révision-rétention. Pour simplifier notre discussion, nous intégrons la phase de rétention, qui consiste à intégrer une nouvelle paire problème-solution dans la base de cas, dans la politique de maintenance du système. Une autre phase du modèle, la révision de solution, n'est pas discutée dans ce document. Cette activité consiste à vérifier la validité d'une solution soit par consultation avec l'usager du système, par simulation ou par évaluation numérique.

[7] La littérature CBR francophone ne propose pas de terme permettant de bien capturer la notion de "authoring". Nous retenons pour cet article le terme "construction". Toutefois d'autres appellations telles que "édition", "ingénieurie", "acquisition"ou "rédaction" nous ont également été proposées.

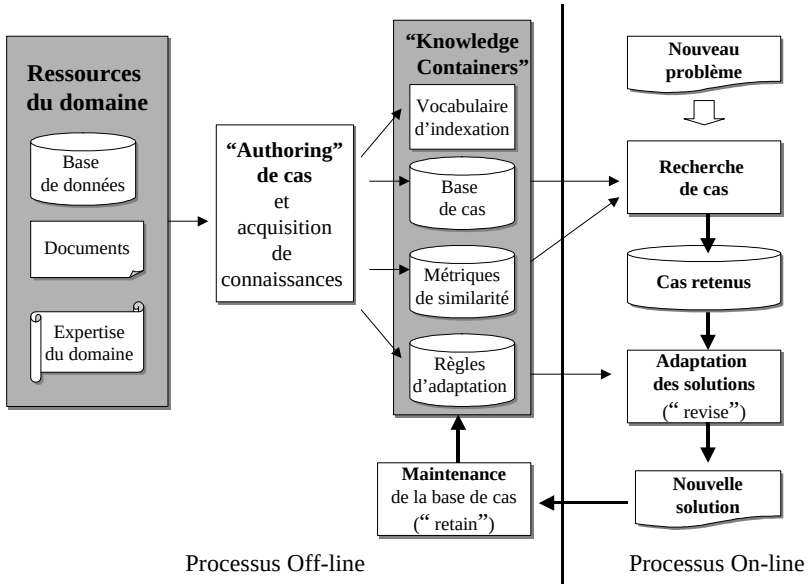


Figure 8 : Modèle générique d'un système CBR

2.1.2 Processus

La recherche : cette phase permet de déterminer les cas de la base qui sont les plus similaires au problème à résoudre. La procédure de recherche est habituellement implantée par une sélection des plus proches voisins (*k-nearest-neighbors*) ou par la construction d'une structure de partition obtenue par induction. L'approche des plus proches voisins utilise des métriques de similarité pour mesurer la correspondance entre chaque cas et le nouveau problème à résoudre. L'approche par induction génère un arbre qui répartit les cas selon différents attributs et qui permet de guider le processus de recherche.

L'adaptation : suite à la sélection de cas lors de la phase de recherche, le système CBR aide l'usager à modifier et à réutiliser les solutions de ces cas pour résoudre son problème courant. En général, on retrouve deux approches pour l'adaptation de cas (Figure 9). Par l'approche transformationnelle (ou structurelle), on obtient une nouvelle solution en modifiant des solutions antécédentes et en les réorientant afin de satisfaire le nouveau problème. Par l'approche générative (ou dérivationnelle), on garde, pour chaque cas passé, une trace des étapes qui ont permis de générer la solution. Pour un nouveau

problème, une nouvelle solution est générée en appliquant l'une de ces suites d'étapes. Des travaux visent également à unifier ces différentes approches d'adaptation (voir Fuchs *et al*. 1999, Fuchs *et al*. 2000 pour une proposition de modèle général).

Peu de systèmes CBR font de l'adaptation complètement automatique. Pour la plupart des systèmes, une intervention humaine est nécessaire pour générer partiellement ou complètement une solution à partir d'exemples. Le degré d'intervention humaine dépend des bénéfices en terme de qualité de solution que peut apporter l'automatisation de la phase d'adaptation.

Maintenance : durant le cycle de vie d'un système CBR, les concepteurs doivent préconiser des stratégies pour intégrer de nouvelles solutions dans la base de cas et pour modifier les structures du système CBR pour en optimiser les performances. Une stratégie simple est d'insérer tout nouveau cas dans la base. Mais d'autres stratégies visent à apporter des modifications à la structuration de la base de cas (par ex. l'indexation) pour en faciliter l'exploitation. On peut également altérer les cas en modifiant leurs attributs et leur importance relative. Cet aspect de recherche est actuellement l'un des plus actifs du domaine CBR (Leake *et al*. 2001).

Construction : ce processus, en amont des activités de résolution de problèmes du système CBR, guide la structuration initiale de la base de cas et des autres connaissances du système à partir de différentes ressources tels des documents, bases de données ou transcriptions d'interviews avec des praticiens du domaine. Ce processus, souvent effectué manuellement par le concepteur du système, se prête moins bien à l'automatisation car il nécessite une connaissance du cadre applicatif pour guider, entre autre, la sélection du vocabulaire d'indexation et la définition des métriques de similarités.

2.1.3 Connaissances

Les différentes connaissances utilisées par un système CBR sont regroupées en quatre catégories (*knowledge containers* – voir Richter 1998) :

- *vocabulaire d'indexation* : un ensemble d'attributs ou de traits (*features*) qui caractérisent la description de problèmes et de solutions du domaine. Ces attributs sont utilisés pour construire la base de cas et jouent un rôle important lors de la phase de recherche.

- *base de cas* : l'ensemble des expériences structurées qui seront exploitées par les phases de recherche, d'adaptation et de maintenance.

- *mesures de similarité* : des fonctions pour évaluer la similarité entre deux ou plusieurs cas. Ces mesures sont définies en fonction des traits et sont utilisées pour la recherche dans la base de cas.

- *connaissances d'adaptation :* des heuristiques du domaine, habituellement sous forme de règles, permettant de modifier les solutions et d'évaluer leur applicabilité à de nouvelles situations.

2.2 Modèles CBR

Il existe plusieurs modèles pour le raisonnement à base de cas. Ces modèles sont regroupés en trois grandes familles : structurelle, conversationnelle et textuelle. Avant de présenter plus en détail les travaux du CBR textuel, nous décrivons dans les sections suivantes les principales différences entre ces trois familles.

2.2.1 Modèle structurel

Le modèle structurel a émergé des premières vagues applicatives de systèmes CBR. Dans ce modèle, toutes les caractéristiques importantes pour décrire un cas sont déterminées à l'avance par le concepteur du système. Ainsi, le concepteur élabore un modèle de données du domaine applicatif. À partir de ce modèle de données, les cas sont complètement structurés et représentés tels des paires <attribut, valeur> (similaire à un *frame* ou à un objet), des arbres ou des graphes. Un exemple de représentation de cas par paires <attribut, valeur> est illustré à la figure 9. D'un point de vue applicatif, un attribut représente une caractéristique importante du domaine d'application. Les échelles de

valeurs les plus fréquemment utilisées pour structurer les attributs sont les entiers/réels, les booléens et les symboles. La représentation des cas peut être sur un seul niveau ou sur plusieurs niveaux (hiérarchie d'attributs).

Cas : *2735*
Entreprise : *BCE*
Date : *22/01/2002*
Dividende annuel : 1,20
Haut_52_semaines : 43.70
Bas_52_semaines : 32.25
Dernier_cours : 35,65
Recommandation : achat

Figure 9 : Exemple de structuration d'un cas en CBR structurel

La similarité entre deux cas est mesurée en fonction de la distance entre les valeurs de mêmes attributs. Cette distance est fréquemment estimée par les mesures euclidienne et de Hamming. La similarité globale entre deux cas est habituellement évaluée par une somme pondérée de la similarité de chacun des attributs. Comme les attributs d'un cas n'ont pas tous la même importance et que cette importance varie d'une situation à l'autre, un poids est attribué à chaque attribut de chaque cas. Ces poids permettent de pondérer la similarité globale entre deux cas en accordant un "vote" plus important aux attributs les plus méritants.

Tous les travaux sur l'adaptation de cas sont menés dans le cadre du modèle structurel. L'adaptation peut varier d'une simple substitution de la valeur d'un attribut jusqu'à la restructuration complète d'une solution. Leake (Leake 1996) identifie environ dix techniques permettant de générer des solutions par substitution, transformation partielle ou dérivation complète. Ces techniques sont habituellement mises en œuvre par des systèmes à base de règles, ce qui nous ramène aux problèmes d'acquisition de connaissance et d'absence de principes généraux. Pour en limiter les difficultés, on retrouve des approches qui évitent l'adaptation en sélectionnant, durant la phase de recherche, des cas qui nécessiteront peu d'adaptation (Smyth et Keane 1995).

2.2.2 Modèle conversationnel

Dans l'approche traditionnelle (le modèle structurel), un problème doit être complètement décrit avant que ne débute la recherche dans la base de cas. Cette exigence présuppose une expertise du domaine d'application permettant de bien caractériser une situation à l'aide de valeurs numériques ou symboliques de sélectionner les principaux facteurs pouvant influencer la résolution de son problème. Toutefois pour quelques domaines comme le service à la clientèle, ces aspects sont difficiles à déterminer à l'avance, surtout pour les usagers novices de systèmes CBR. Le modèle conversationnel a donc été proposé par *Inference Corporation* pour surmonter ces difficultés. Il est actuellement le modèle le plus répandu parmi les applications commerciales du CBR.

Cas : *241*
Titre : *cartouche d'encre endommagée causant des traces noires*
Description : *l'imprimante laisse de petits points noirs sur les deux côtés de la page. Parfois des larges tâches couvrent également la région à imprimer.*
Questions :
Est-ce que les copies sont de mauvaise qualité ? Réponse : oui Score : (-)
Quels types de problèmes avez-vous ? Réponse : trace noires Score : (default)
Est-ce qu'un nettoyage de l'imprimante règle le problème ? Rép : non …
Actions : vérifier la cartouche d'encre et la remplacer si le niveau d'encre est faible

Figure 10 : Exemple de cas pour le modèle conversationnel

Comme son nom l'indique, le modèle CBR conversationnel mise sur l'interaction entre l'usager et le système (d'où la notion de "conversation") pour définir progressivement le problème à résoudre et pour sélectionner les solutions les plus appropriées (Aha *et al*. 2001). Un cas conversationnel consiste en trois parties (voir Figure 10) :

- un problème *P* : une brève description textuelle, habituellement de quelques lignes, de la nature du problème exprimée.
- une série de questions et de réponses Q_A : des index, exprimés sous forme de questions, permettant d'obtenir plus d'information sur la description du problème. Chaque question a un poids représentant son importance par rapport au cas.

- une action *A* : une description textuelle de la solution à mettre en œuvre pour ce problème. Cette description n'est pas structurée (*free-text*).

Cette représentation de cas est donc une extension du modèle structurel avec des attributs de trois types bien précis : description, questions, actions. La notion de trait est étendue à la notion de question afin de pouvoir interroger l'usager.

Dans le schéma de résolution du CBR conversationnel, l'interaction entre le système et l'usager se fait comme suit :

- l'usager fournit au système une brève description textuelle du problème à résoudre et le système calcule la similarité entre cette description et la section "problème" des cas. Le système propose alors à l'usager une série de questions.

- l'usager choisit les questions auxquelles il souhaite répondre. Pour chaque réponse fournie par l'usager, le système réévalue la similarité de chacun des cas. Les questions n'ayant pas reçu de réponse sont présentées par ordre décroissant de priorité.

- lorsqu'un des cas atteint un niveau de similarité suffisamment élevé (i.e. qu'il franchit un seuil), le système propose ce cas comme solution. Si aucun cas n'atteint un degré de similarité suffisant et que le système n'a plus de questions à poser à l'usager, le problème est stocké comme étant non résolu.

Les systèmes CBR conversationnels n'effectuent pas d'adaptation des solutions passées. Une des raisons est que la portion 'solutions' des cas n'est pas structurée (*free-text*), ce qui rend difficile la formulation de connaissances d'adaptation. Également, il semble que, pour les applications de type *help-desk*, les solutions sont relativement faciles à modifier, même par un préposé inexpérimenté. De plus, l'investissement en temps et en efforts consacrés à développer un système d'inférence qui modifie les solutions est difficile à justifier dans ce contexte opérationnel.

2.2.3 Modèle textuel

Les travaux sur le raisonnement à base de cas textuels portent sur la résolution de problème à partir d'expériences dont la description est contenue dans des documents textuels. Dans cette approche, les cas textuels sont soit non-structurés ou semi-structurés. Ils sont non-structurés si leur description est complètement en *free-text*. Ils sont semi-structurés lorsque le texte est découpé en plusieurs portions étiquetées par des descripteurs tels que "problème", "solution", etc. Un cas textuel non-structuré est un cas dont le seul attribut est textuel tandis qu'un cas textuel semi-structuré est un cas dont un sous-ensemble de ses attributs est textuel.

Pour ce modèle, la représentation textuelle des cas joue habituellement un rôle important dans la résolution du problème. Elle peut être une finalité en soi : par exemple, obtenir le texte d'un jugement légal servant de jurisprudence à une nouvelle cause. Elle peut aussi décrire une situation et une solution qui ne peuvent être facilement codifiées selon un schéma de représentation de connaissance.

Cette voie de recherche est relativement récente car les premiers travaux datent du milieu des années 90. A ce jour, aucune représentation standard ne s'est dégagée pour le modèle textuel. Les approches actuelles misent leurs efforts principalement sur la phase de recherche sur la base de cas et ne proposent pas d'avenue pour l'adaptation de solutions textuelles.

Nous pouvons identifier deux pôles importants dans les différents travaux en CBR textuel :

- structuration de cas textuels : on représente les textes selon un nombre limité de traits basés sur des caractéristiques du domaine (concepts, catégories, sujets, mots-clés, etc.). Pour ce pôle de recherche, on vise à structurer le mieux possible les cas textuels afin de tirer profit de techniques développées pour les systèmes CBR structurel. Les efforts sont déployés pour enrichir l'indexation des textes à l'aide de traitements relativement élaborés comme la catégorisation de texte. Cette approche est intéressante pour les applications

dont le domaine est restreint. Le projet SMILE (Brüninghaus 1997) présenté à la section 4.5 en est un exemple.

- extension du modèle de recherche d'information : dans ce pôle de recherche, on élabore des mécanismes de recherche plus sophistiqués tout en gardant le processus d'indexation le plus simple possible. Dans ce cadre, le choix des traits de cas est déterminé à partir de la fréquence de mots-clés ou de syntagmes de référence (*keyphrases*). Les particularités de l'application se reflètent au niveau de la recherche, soit par la définition de mesures de similarité sémantique ou par des extensions au modèle vectoriel de recherche d'information (Salton & McGill 1984). Cette approche semble plutôt valide pour les applications génériques qui veulent conserver une indépendance par rapport au domaine d'application. Le projet FAQFinder (Burke *et al*. 1995) présenté à la section 4.1 en est un exemple.

Ces deux pôles sont en fait des stéréotypes auxquels empruntent la plupart des approches actuelles. Nous présentons à la section 4 divers travaux qui illustrent l'exploitation de connaissances du domaine et le contenu linguistique des textes.

Le CBR textuel diffère de l'approche structurelle dans laquelle les textes sont tout simplement des chaînes de caractères sans syntaxe ni sémantique. De plus, cette dernière impose une structuration complète des attributs d'un cas. Nous considérons également que le modèle conversationnel, présenté à la section précédente, ne fait pas partie des approches textuelles. La phase préliminaire du CBR conversationnel se limite à une comparaison, par mots-clés ou n-grammes[8] de caractères, de courtes descriptions textuelles de problèmes. Durant la phase suivante, l'interaction avec l'usager est guidée par une suite de questions et de réponses. Les échanges lors de l'interaction ne font l'objet d'aucun traitement textuel. La langue y est utilisée uniquement dans le but de rendre les questions plus intelligibles à l'usager du système.

2.3 Principaux travaux en CBR textuel

Dans cette section, nous présentons les travaux que nous jugeons les plus représentatifs de l'état d'avancement du CBR textuel. Ces travaux ont été sélectionnés parce qu'ils apportent des contributions au raisonnement à base de cas et, pour la plupart, ont une influence sur les travaux actuels de la communauté CBR. Ce tour d'horizon donne un aperçu de l'étendu du niveau de structuration des cas, de la complexité des métriques de similarité et des mécanismes de recherche sur la base de cas. Toutefois on retrouve également d'autres travaux combinant CBR et documents textuels pour des domaines tels que le service à la clientèle (par ex. les systèmes *help-desk*, voir Racine & Yang 1997), la gestion de connaissances et les applications du traitement de la langue (par ex. la traduction automatique, voir Macklovitch *et al*. 2000).

2.3.1 FAQ-Finder – exploitation de questions-réponses

FAQFinder (Burke *et al*. 1995) est un système de questions-réponses basé sur les foires aux questions (*Frequently-Asked Questions* - FAQs) de USENET. Un FAQ est une

[8] Une représentation de type *n-gramme* consiste à découper un texte en séquences de *n* caractères. Par exemple, le terme *raisonnement* serait représenté en trigramme par les séquences suivantes : {rai, ais, son, onn, nne, nem, eme, men, ent}.

réponse à une question fréquemment posée dans un groupe d'intérêt (par ex. un groupe de programmation Java). Un FAQ est considéré comme un cas CBR car il contient la description d'un problème (la question) et la description d'une solution (la réponse). Un exemple de FAQ est présenté à la Figure 11.

> ***FAQ # : 241***
> ***Question :*** *où se transigent les actions ordinaires de BCE ?*
> *Réponse : les actions ordinaires de BCE sont négociées aux Bourses de Toronto, New York ainsi qu'à la Bourse de Suisse sous le symbole BCE.*

Figure 11 : Structuration de "foires aux questions" (*frequently-asked questions*)

Le système est conçu pour recevoir en entrée une question en langage naturel et identifier les FAQs de USENET qui sont les plus similaires à cette question. La recherche de réponses pertinentes à ces questions est effectuée en 2 étapes.

La première étape permet de choisir, à partir de tous les fichiers FAQ de USENET, le sous-ensemble qui est le plus pertinent. Chaque fichier contient plusieurs dizaines de questions-réponses. Par exemple, à la question "*What is garbage collection*", on obtiendrait des fichiers de FAQ sur différents langages de programmation tels que Lisp et Java. Cette étape adopte une approche de recherche d'information et utilise le système SMART (Buckley 1985). Les fichiers FAQ sont convertis selon le modèle vectoriel de recherche d'information et le rangement des fichiers est effectué selon des métriques statistiques (*tf*idf*[9]). La comparaison entre la question et un fichier de FAQs est basée sur la correspondance exacte des termes des deux textes. Cette étape permet de filtrer les fichiers de FAQ et de n'en retenir que quelques dizaines.

La deuxième étape tente d'identifier, pour les fichiers jugés pertinents, les FAQs individuels qui correspondent le mieux à la question de l'usager. La correspondance entre la requête et chaque FAQ est évaluée selon trois métriques de similarité :

- métrique statistique : des fonctions du domaine de la recherche d'information sur des vecteurs de poids de type *tf*idf*. Différentes fonctions pour mesurer la

[9] Une mesure indiquant la fréquence d'un terme et sa capacité de discriminer entre plusieurs documents.

distance entre les vecteurs de termes ont été testées dans leurs expérimentations.

- métrique sémantique : en utilisant le thesaurus WordNet, la distance sémantique entre chaque paire de mots est estimée. Pour évaluer cette distance, on utilise un algorithme de type *edge-counting* qui estime la distance entre deux concepts à partir du nombre de liens qui les séparent dans un réseau sémantique.

- métrique de recouvrement : pour quelques expérimentations, les auteurs ont tenté d'utiliser le pourcentage de mots de la requête qui est inclus dans les FAQs. Des résultats expérimentaux indiquent que cette métrique n'apporte pas de progrès significatifs et peut même causer une dégradation du système.

La similarité globale entre la requête et chaque FAQ est une somme pondérée de ces métriques. Les questions-réponses jugées les plus pertinentes sont présentées à l'usager en ordre décroissant de similarité. L'usager peut alors sélectionner les FAQs qu'il juge intéressants et recommencer la recherche.

Des expérimentations ont été menées sur un corpus de 241 questions, dont 138 avaient des réponses dans les FAQs de UseNet. Le système contient plus de 600 fichiers FAQ, et donc quelques dizaines de milliers de FAQs individuels. Les questions sont habituellement courtes (quelques dizaines de mots) mais les réponses sont plus longues et peuvent parfois contenir plus de mille mots.

La performance de la première phase, basée sur le système SMART, est excellente. Le bon fichier FAQ (i.e. le bon thème) est retourné parmi les cinq premières positions dans 88% des cas et en première position dans 48% des cas. La deuxième étape donne des résultats intéressants lorsque les métriques statistiques et sémantiques sont utilisées conjointement. La capacité du système à fournir une bonne réponse parmi les 5 premières recommandations est de 55% pour la métrique statistique seulement, 58% pour la métrique sémantique seulement, et 67% pour une métrique combinée. La qualité des résultats est toutefois limitée par l'utilisation de WordNet qui est une ressource

linguistique trop générale pour ce type d'application. Également le système éprouve des difficultés à identifier les questions qui n'ont pas de réponses dans les FAQ.

Plusieurs tentatives ont été menées pour améliorer la performance du système (Burke *et al*. 1997). L'analyse lexicale (*part-of-speech tagging*) et syntaxique ont été utilisés pour identifier les termes importants des questions. Ces analyses n'ont pas permis d'améliorer significativement le système. Pour la deuxième étape du système, des techniques pour reconnaître le type de question et pour en faire la reconversion ont été appliquées. Les auteurs pensaient pouvoir améliorer les performances du système en restreignant les comparaisons entre questions de même type et en les exprimant sous différentes formes (Burke *et al*. 97). Par exemple, la question "Quand devrais-je changer l'huile de mon automobile ?" peut être remplacé par la question "A quelle fréquence recommande-t-on de faire les changements d'huile ?". Ceci correspond à paraphraser des questions à partir de canevas prédéterminés. Toutefois ces reconversions s'avèrent peu fiables lorsqu'elles sont effectuées uniquement à partir d'informations syntaxiques. Également un désambiguïseur sémantique a été utilisé pour améliorer la performance du système en terme de courbe rappel-rejet (Lytinen *et al*. 2000). Finalement, des expérimentations ont été menées afin de déterminer automatiquement la pondération de chacune des fonctions de similarité par apprentissage automatique (programmation génétique) (Cooper 1996, Burke *et al*. 1997). Par ailleurs, aucune tentative n'a été menée pour combiner ou modifier les réponses des FAQs (donc pas d'adaptation ou de modification de textes).

2.3.2 SPIRE - utilisation de cas pour rehausser la recherche d'information

Ces travaux, de J. Daniels et E. Rissland de l'Université du Massachusetts à Amherst (Daniels 1996) ont mené au développement de SPIRE, un système hybride de CBR et de recherche d'information. Dans ce système, le CBR aide les usagers du système de recherche d'information INQUERY à mieux formuler leurs requêtes et à

identifier les passages pertinents dans les documents. Ainsi le module CBR agit comme pré-processeur et post-processeur pour une tâche de recherche d'information.

Le module CBR contient deux bases de cas : la première base contient des cas structurels décrivant les principaux attributs (*features*) du contenu de documents tirés du corpus. La deuxième base contient des extraits textuels pour chacun des attributs d'un cas. Les bases de cas sont construites manuellement par un analyste humain.

Le traitement d'une requête s'effectue en deux étapes. Premièrement, la première base est utilisée pour sélectionner un nombre restreint de cas que l'on sait pertinents à la requête. Le contenu des cas est utilisé par le système INQUERY pour faire l'expansion de la requête. Ce mécanisme est analogue au *pseudo-relevance feedback* qui permet d'ajouter des termes à la requête initiale et d'ajuster le poids de chacun de ces termes. La requête étendue est alors traitée par INQUERY pour identifier les documents les plus pertinents de la collection.

La deuxième base de cas contenant les extraits sert à formuler une requête pour l'identification des passages. La requête contient soient tous les termes soient seulement les termes communs des passages reliés à un attribut. Cette requête est utilisée par INQUERY pour déterminer les fenêtres de mots pertinentes des documents retenus à la première étape. Une comparaison des extraits obtenus avec des requêtes formulées par le système et par un expert du domaine a été menée pour 10 attributs et 20 documents (Daniels & Risland 1998). Les résultats indiquent que SPIRE offre une précision légèrement supérieure pour un plus grand nombre d'attributs (précision globale d'environ 50% pour chacun). Une analyse des résultats illustre l'avantage des extraits qui donnent un contexte plus diversifié que les requêtes formulées par des humains.

2.3.3 DRAMA – cas partiellement textuels

Le projet DRAMA (Leake & Wilson 1999, Wilson & Bradshaw 2000) a pour but de gérer la connaissance des concepteurs de systèmes aéronautiques. Le système développé dans le cadre de ce projet aide les concepteurs lors du design de nouveaux

avions et permet de préserver les différents aspects du design. Dans ce système, chaque design est décrit à l'aide de cartes conceptuelles (*concept mapping*[10]), d'attributs descriptifs (par ex. les caractéristiques du moteur) et d'annotations textuelles donnant des précisions sur les choix des concepteurs et sur les caractéristiques des composantes de l'avion. Puisque les cas contiennent des parties structurées (les cartes et attributs valués) et des parties non structurées (les annotations), les auteurs les qualifient *de weakly-textual*, i.e. des cas dont une partie est textuelle mais dont la plus importante portion est non textuelle.

Le mécanisme de recherche de ce système combine des fonctions de similarité sur les attributs structurés et sur les annotations textuelles. Afin de simplifier la recherche textuelle, les auteurs utilisent un modèle vectoriel de recherche d'information. Plus précisément, la recherche comporte les étapes suivantes :

- chacun des attributs textuels est converti individuellement en vecteur de termes; puisque les descriptions textuelles sont courtes, la sélection de termes ne repose pas sur la fréquence de mots-clés mais plutôt sur les syntagmes nominaux de type *Nom-Nom*, *Adj-Nom* ou *Nom-Prep-Nom*. Les syntagmes sont identifiés avec l'aide d'un lexique (Ward 1994).

- un poids, similaire au *tf*idf*, est par la suite attribué aux syntagmes des différents vecteurs de termes.

- la similarité entre chaque paire de vecteurs textuels est déterminée selon la métrique du cosinus. Cette mesure est combinée aux similarités des autres attributs structurés des cas (cartes et attributs descriptifs).

En résumé, le système DRAMA illustre bien comment faire l'intégration de quelques attributs textuels au sein d'un système CBR structurel. Également il offre la

[10] Le "concept mapping" est un formalisme de représentation qui décrit par un graphe bi-dimensionnel la structure cognitive de la conception (les concepts et leurs interrelations). Contrairement aux réseaux sémantiques, les cartes conceptuelles ne sont pas contraintes syntaxiquement et n'ont pas de sémantique.

particularité que la similarité entre les portions textuelles des cas est établie à partir des syntagmes nominaux. Bien que le système permette l'adaptation des diagrammes de design, les auteurs ne proposent pas de techniques pour adapter les portions textuelles de cas en fonction du nouveau design.

2.3.4 CBR-Answers – réseau pour la recherche de cas

Une approche pour améliorer la performance d'un système CBR est de structurer sa base de cas. Pour le système CBR-Answers, développé par Mario Lenz (Lenz & Burkhard 1997, Lenz & Glintschert 1999), on utilise une structure de réseau pour "compiler" la base. Durant la phase de recherche, les valeurs de similarité sont propagées dans les nœuds du réseau et permettent de déterminer la pertinence de chacun des cas.

Tel qu'illustré à la figure 12, le réseau de recherche de cas (*case retrieval net*) contient un ensemble de nœuds, les entités d'information (IE). Les IEs décrivent des éléments de documents (représentés par des rectangles arrondis) tels que des mots-clés, des termes complexes (*keyphrases*), des paires attributs-valeurs et des catégories du domaine. Ils décrivent également des identifiants de cas (représentés par des hexagones).

Les liens du réseau décrivent soit a) l'appartenance d'une entité d'information à un cas (liens continus), soit b) une relation de similarité entre deux entités (liens pointillés). Des poids qui indiquent le degré de similarité entre deux entités ou l'importance d'une entité par rapport à un cas sont attribués aux liens. Les liens représentent la structure d'inférence et guident la propagation des valeurs de similarité dans le réseau.

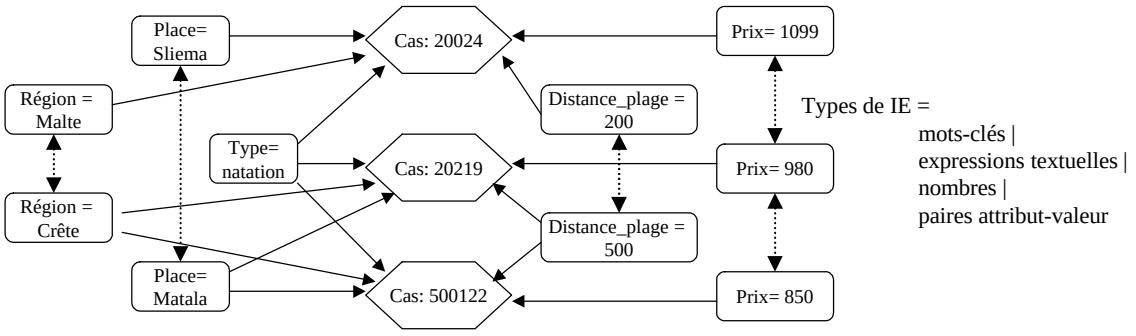


Figure 12 : Exemple de réseau de recherche de cas (adapté de Lenz *et al*. 1998)

Pour le traitement de cas décrit à partir de documents textuels, Lenz propose une procédure pour l'indexation des cas et la construction du réseau (Lenz *et al*. 1998). Cette structuration, qui repose sur une analyse lexicale des textes, est comme suit :

- chaque terme des documents est étiqueté selon sa catégorie lexicale et est normalisé en le remplaçant par sa racine morphologique. Les mots ayant une racine commune sont regroupés. Ces racines forment les traits des cas.

- les termes sont classés selon trois catégories (inutile, utile et potentiellement utile) et on retient ceux des deux dernières catégories. Les définitions proposées pour ces catégories sont :

  - inutile : des déterminants, des auxiliaires, des pronoms et des termes que l'on retrouve dans une liste de mots-outils.

  - utile : les adverbes, les adjectifs, les verbes et les noms.

  - potentiellement utile : les termes n'appartenant pas aux deux catégories précédentes.

- on complète la construction du réseau par une vérification manuelle des termes utiles et potentiellement utiles.

La recherche dans un réseau de recherche de cas débute par le découpage des termes de la requête en mots clé et leur association à des entités d'information du réseau. Les entités sont ensuite activées. Les valeurs de similarité sont alors propagées parmi les différents IEs du réseau pour déterminer les IEs similaires. Finalement les valeurs de pertinence sont propagées aux nœuds des cas pour établir un ordonnancement parmi les différents cas de la base.

La mesure de similarité globale entre un nouveau problème Q et un cas C est estimée en fonction de leur similarité pour chacun des IE (*sim*) à l'aide de la fonction suivante :

$$SIM(Q,C) = \sum_{e_i \in Q} \sum_{e_j \in C} sim(e_i, e_j)$$

Le système CBR-Answers a été utilisé pour développer quelques applications commerciales de service à la clientèle dont FallQ (pour un fournisseur de services en télécommunication) et Simatic (pour le groupe Automation & Drive de Siemens AG). Ces systèmes permettent la recherche de documentation technique (par ex. des spécifications de composantes, des problèmes connus) et la réponse aux questions fréquentes (documents de type FAQ).

Des expérimentations menées avec FallQ, qui contient 45 000 documents, indiquent un temps de réponse qui varie entre 0.01 et 0.20 secondes par requête. Ce résultat illustre la bonne performance des algorithmes de propagation dans les réseaux de recherche de cas. Une étude a été menée avec Simatic pour évaluer la contribution des différents niveaux d'entités d'information. Avec un corpus de 500 documents, l'utilisation de simples mots-clés offre la plus faible performance en terme de courbe précision-rappel et l'ajout successif de chaque couche d'entités (terme de thesaurus, de glossaire et thème du document) augmente significativement la performance du système. Cette étude illustre bien l'importance de la phase de structuration de documents lors de la création de la base de cas. Par contre, l'étude ne prend pas en compte l'ajout de paires attribut-valeurs, un élément important dans la structuration de cas semi-structurés.

2.3.5 **SMILE – Factorisation des cas par catégorisation de textes**

Ces travaux de Steffanie Brüninghaus (Brüninghaus & Ashley 1997) explorent des approches d'apprentissage automatique pour l'indexation des cas textuels. L'apprentissage permet de catégoriser des textes selon des attributs du domaine que les auteurs appellent "facteurs". Les textes sont des comptes-rendus d'actions en justice impliquant des fraudes de secrets industriels. Ces travaux ont été initiés par Aleven

(Aleven & Ashley 1996) qui a proposé un système tutoriel, basé sur le CBR structurel, pour enseigner aux étudiants de droit comment argumenter pour ce type de procès. Les "facteurs" représentent des situations qui jouent un rôle positif (favorable) ou négatif (défavorable) lors de l'argumentation de la cause. Les "facteurs" sont reliés selon une hiérarchie comprenant des niveaux spécifiques, des niveaux abstraits et des niveaux de thèmes généraux.

Dans les travaux d'Aleven, les cas étaient indexés manuellement. Les travaux sur SMILE visent à extraire automatiquement des facteurs à partir de ces textes légaux. On note trois séries de travaux :

- *la catégorisation de textes complets* (Brüninghaus 1997) : chaque texte est représenté comme un vecteur de fréquence de mots-clés. Des expérimentations ont été menées pour catégoriser un corpus de 147 cas selon 26 facteurs à l'aide de techniques d'apprentissage tel que *Naïve Bayes*, *Winnow*, *Rocchio* et *Exponentiated Gradient*. Pour la plupart des facteurs, les algorithmes ont identifié peu d'exemples positifs, amenant des faibles performances en terme de justesse (*accuracy*), de précision et de rappel.

- *la catégorisation de passages* (Brüninghaus 1999) : divers passages sont étiquetés manuellement pour indiquer la présence de facteurs dans le texte (voir figure 13). Ces passages servent de corpus d'entraînement pour apprendre comment catégoriser les portions de textes. Un thesaurus est également utilisé pour identifier les correspondances entre des termes de différents passages. Ceci permet de détecter la présence de facteurs exprimés avec des mots différents mais significativement équivalents. A partir d'un échantillon de 2200 passages (de longueur moyenne de 7.5 termes), l'algorithme ID3 a atteint jusqu'à 80% de précision et de rappel (minimum : 30% de précision et 50% de rappel). L'utilisation du thesaurus est par contre moins concluante. Pour quelques facteurs, elle apporte des améliorations de 40% tandis qu'elle apporte une dégradation de 10% à 20% pour d'autres facteurs.

- *la catégorisation à partir d'informations extraites* (Brüninghaus & Ashley 2001) : plus récemment, Brüninghaus a proposé d'utiliser le système d'extraction d'information AutoSlog (Riloff 1996) pour repérer trois types d'information : les entités nommées, les *case frames*, et les négations. Les informations extraites seraient alors utilisées par le processus d'apprentissage pour identifier la présence de facteurs dans les textes légaux. Ces travaux sont en cours et aucune expérimentation n'a encore été menée.

Case 37
<F15/> Plaintiff's packaging process involved various "tempering steps" that were not used by competitors or described in the literature. </F15> <F16/> Only a handlful of plaintiff's employees knew of the packaging operations, and they were all bound by secrecy agreements. </F!6>. <F6/> There was also testimony that packaging information was closely guarded in the trade.</F6>
<F1/>Plaintiff's president sent a letter to defendant which conveyed plaintiff's manufacturing formula. </F1>. <F21/> The letter also stated….

Figure 13 : Étiquetage de passages selon des facteurs (tirée de Brüninghaus 1999)

2.3.6 PRUDENCIA – structuration de cas de type *template mining*

PRUDENCIA est un système qui facilite la recherche documentaire en jurisprudence légale (Weber 1998, Weber *et al*. 1998). Il permet de rechercher des situations, décrites dans des documents textuels, qui sont similaires à une nouvelle cause juridique.

La principale contribution de ces travaux est de proposer une démarche pour convertir des textes légaux en cas structurés. Cette démarche s'appuie principalement sur la forte structuration des documents légaux utilisés dans ce projet. On retrouve dans chaque texte un certain nombre de sous-sections (par ex. "en-tête", "résumé", "corps", "conclusion"). Les sous-sections comportent une régularité puisque leur contenu est homogène (mêmes thèmes) et qu'on y retrouve des phrases identiques situées aux même endroits. Cette régularité facilite le processus d'extraction du contenu des textes.

Avant d'être exploités par un système CBR, les documents sont structurés à l'aide de formulaires comprenant neuf champs (type de pétition, numéro de cas, district, page,

date, fondation, thème, lois secondaires, catégorie, résultat, unanimité). Les champs des formulaires et leurs valeurs admissibles ont été sélectionnés par un expert du domaine. Le processus de structuration des cas reposent sur un certain nombre de méthodes qui alimentent les formulaires. On note les méthodes suivantes :

- Directe : des attributs sont explicites et leurs valeurs sont situées à des positions fixes dans le texte ;

- par mots-clés : pour chaque champ du formulaire, on recherche dans les sous-sections du texte des mots-clés contenus dans une liste d'expressions. Pour faciliter la recherche, la racine des mots et un dictionnaire de synonymes sont utilisés ;

- par patron : des expressions régulières permettent d'obtenir des numéros d'articles de loi (par ex. "for infringing articles 26 & 97 of Penal Code"). Les patrons sont définis manuellement ;

- par règle : des règles permettent de tenir compte de la dépendance entre différents champs du formulaire et de diriger ainsi la méthode vers la sous-section adéquate. Les règles sont définies manuellement ;

- par comparaison : le champ "thème" est déterminé suite à une comparaison avec d'autres cas structurés (similaire à une classification par les plus proches voisins).

Le système contient 3500 cas de jurisprudence. La recherche se fait par comparaison entre les champs des formulaires à l'aide de métriques binaires (0 ou 1) et graduées (sur une échelle [0,1]). Quelques expérimentations permettent d'évaluer le temps de recherche du système, mais aucune évaluation du processus de recherche en terme de précision et rappel n'est présentée.

2.4 Comparaison des travaux en CBR textuel

Les deux tableaux suivants résument les principaux points des travaux présentés à la section précédente. Le Tableau 2 décrit les particularités de la tâche accomplie par le système et le Tableau 3 présente les particularités techniques de l'approche CBR préconisée.

On note qu'une majorité de systèmes sont utilisés pour des tâches de type "recherche d'information" (IR) : DRAMA, FAQFinder, CBR-Answers, SPIRE et PRUDENCIA. Par contre, la plupart de ces applications se démarquent des approches typiques de recherche d'information par l'utilisation de connaissances du domaine dans la structuration de la base de cas et dans la prise en compte de la tâche à accomplir dans le processus de recherche (mesures de similarité du domaine).

Travaux	Tâche	Caractéristiques du domaine	Caractéristiques des cas
FAQFinder	*Recherche* de documents structurés selon un format question-réponse (FAQ).	Pas de domaine en particulier. Tout texte de type *frequently-asked questions* peut être considéré, indépendamment du sujet du groupe d'intérêt.	Les 919 fichiers initiaux contiennent 23260 FAQs. Les FAQs, d'une longueur moyenne de 245 mots, sont structurés en deux parties : une courte question et une réponse habituellement plus longue.
Drama	*Recherche* de dossiers de design et préservation de la connaissance des concepteurs.	Aéronautique, ce qui laisse présager un vocabulaire restreint et un groupe de concepts relativement limités.	Les 62 cas sont en partie structurés (diagrammes, attribut-valeur) et en partie textuels. Les attributs textuels sont courts et moins importants que les attributs structurés.
Spire	*Recherche* de passages pertinents dans un corpus de documents.	Gestion de faillite personnelle. L'approche ne dépend pas du domaine. Applicable à un domaine restreint seulement.	Aucune contrainte sur la nature des textes. Une base de cas décrit des documents du corpus (cas structurels) et l'autre base contient des extraits textuels (moyenne : 46 termes significatifs).
CBR-Answers	*Help-desk* pour la *recherche* de documents question-réponse et de documents corporatifs (gestion de connaissance).	Domaines d'automatisation de processus et télécommunications. Exploitation du vocabulaire et des concepts du domaine. Approche valide pour les domaines peu restreints.	Seuls les documents question-réponse sont structurés. La longueur des documents varie. La base de l'application FallQ contient 45 000 cas.
Smile	*Catégorisation* de textes légaux selon différents facteurs.	Domaine de la fraude relié aux secrets industriels. Basés sur une terminologie et des concepts propres aux domaines.	Des textes légaux relativement longs et complexes. Aucune structuration des textes à priori. La base contient 147 cas.
Prudencia	*Recherche* de textes légaux.	Domaine de la jurisprudence légale.	Des textes légaux relativement complexes (moyenne de 725 mots). Prise en compte la structure rhétorique des textes. La base contient 3500 cas.

Tableau 2 : Caractéristiques des domaines des systèmes CBR textuel

Une autre différence par rapport aux applications de recherche d'information est que ces applications reposent sur la recherche d'un seul cas similaire pour accomplir leur tâche. Il n'est donc pas souhaitable que ces applications retournent le plus de cas pertinents possibles. En fait, plusieurs bases de cas de ces systèmes ne contiennent que des cas pertinents dont le degré de similarité varie.

L'étendue des domaines d'application de ces systèmes varie. Les approches de SMILE, SPIRE et PRUDENCIA présument que des connaissances du domaine sont disponibles pour la sélection des attributs et la structuration des cas. En revanche, il y a peu de contraintes pour FAQFinder car on ne tente pas de modéliser les domaines abordés par les FAQs de USENET.

Travaux	**Indexation**	**Structuration**	**Recherche**
FAQFinder	Mots-clés, provenant des fichiers, ayant fait l'objet de *stemming* et filtrés par rapport à une liste de mots-outils.	Création automatique de vecteurs de mots-clés et attribution de poids (*tf***idf*).	Style IR avec cosinus. Utilise des métriques de similarité statistique(*tf***idf*) et de similarité sémantique (*edge-counting*). Reformulation de questions (paraphrasage) pour faciliter la comparaison.
Drama	Sélection de syntagmes nominaux tirés des textes à l'aide d'étiquetage lexical.	Création automatique de vecteurs de termes (syntagmes) et attribution de poids (*tf***idf*).	Style IR avec opérateur de cosinus.
Spire	Un groupe d'attributs (*features*) du domaine sélectionné par le concepteur.	Création manuelle de *frames* et d'extraits textuels.	Comparaison d'attributs pour le CBR et recherche IR par le système INQUERY.
CBR-Answers	Mots-clés de fichiers, syntagmes nominaux et termes du domaine ajoutés manuellement par le concepteur.	Création d'un réseau qui relie i) les attributs entre eux et ii) les cas aux attributs.	Propagation des valeurs de similarité dans un graphe dirigé (*case retrieval net*).
Smile	Un groupe de facteurs provenant d'une modélisation manuelle du domaine.	Catégorisation des documents, de passages ou d'informations extraites à partir de techniques d'apprentissage automatique.	Comparaison de la présence/absence de facteurs (décrit dans Aleven & Ashley 1996).
Prudencia	Des attributs fournis par un expert du domaine juridique.	Création semi-automatique de *frames* (*template mining*)	Comparaison d'attributs avec mesures binaires et graduées.

Tableau 3 : Caractéristiques techniques des systèmes CBR textuel

On note que le degré de structuration et la longueur des textes varient également. La plupart des textes utilisés dans ces applications sont peu structurés. Quelques textes n'offrent pas de découpage (par ex. SMILE, CBR-Answers), d'autres se limitant à

distinguer les portions « problème » et « solution » des cas (les FAQs de FAQFinder). Seuls les textes juridiques de PRUDENCIA offrent un grand nombre d'attributs explicitement décrits dans les textes. La petite taille du corpus et la complexité des textes représentent un défi pour SMILE, ce qui explique l'approche distincte qui y est utilisée.

Le Tableau 3 illustre qu'aucun des projets présentés ne propose de processus complexes dans le choix des traits de cas. Soit que le choix est fait manuellement (SMILE, SPIRE, PRUDENCIA), soit statistiquement par des méthodes du domaine de la recherche d'information (DRAMA et FAQFinder). La démarche proposée par CBR-Answers est mixte. Pour les approches statistiques, les textes font l'objet d'étiquetage lexical et d'extraction de racine.

Tel qu'illustré à la figure 14, il est intéressant de noter le niveau de structuration des cas par rapport au mécanisme de recherche de chacun des systèmes. Les systèmes FAQFinder et SPIRE misent principalement sur des mécanismes plus élaborés de recherche pour accomplir leur tâche. Les systèmes SMILE et PRUDENCIA axent plutôt leurs efforts sur un enrichissement des cas pour atteindre de bonnes performances. CBR-Answers est le seul système à œuvrer sur les deux plans. Les approches que nous proposons au chapitre 3 se situe plus près travaux de FAQFinder et SPIRE.

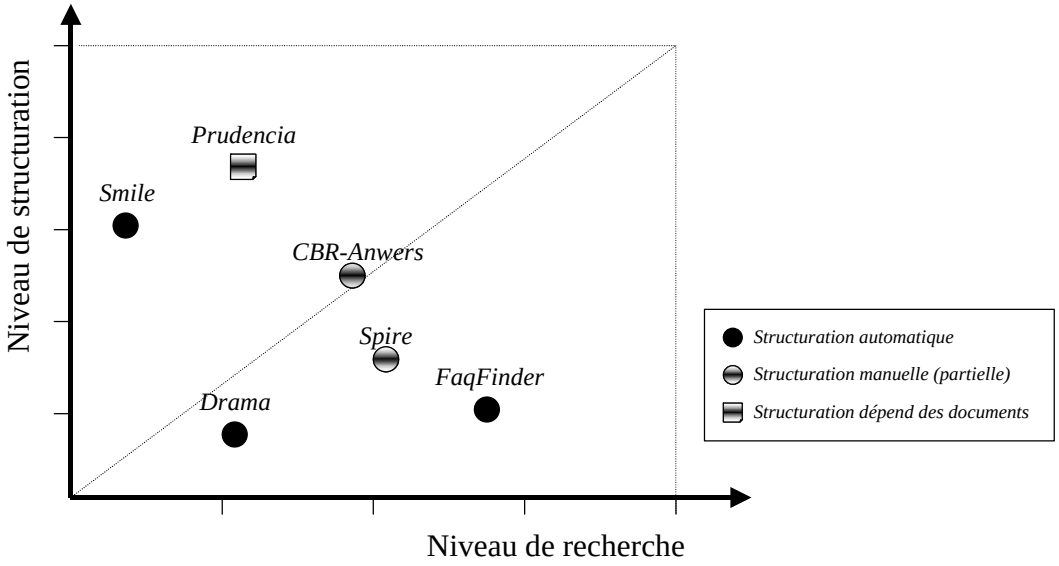


Figure 14 : Positionnement des approches en CBR textuels

Quelle est la meilleure approche ? Devrait-on miser sur une représentation conceptuelle de cas, sur des fonctions de similarité plus riches ou sur un formalisme de recherche plus élaboré ? Bien que l'on puisse identifier les avantages et désavantages de chaque système (voir Tableau 4), il ne se dégage malheureusement pas de réponse à cette question à partir des travaux répertoriés. Toutefois plusieurs facteurs jouent un rôle important dans le choix de l'approche.

Travaux	Pour	Contre
FAQFinder	Une combinaison fructueuse de similarités statistique et sémantique.	Le système éprouve des difficultés à identifier les questions sans réponse
Drama	La simplicité et l'intégration des attributs structurels et textuels.	Les syntagmes nominaux apportent-ils une contribution significative ?
Spire	Les extraits de texte permettent de bien définir le contexte de recherche.	L'approche se transpose mal dans un cadre de résolution de problème.
CBR-Answers	La rapidité de la phase de recherche.	Il y a des limitations sur la nature des métriques de similarité.
Smile	L'approche de structuration de cas par apprentissage est efficace.	L'utilisation de connaissances sémantiques semble inefficace.
Prudencia	On y exploite la forte structuration des textes du domaine applicatif.	L'approche est limitée aux domaines restreints et la méthodologie de structuration manque de généricité.

Tableau 4 : Avantages et désavantages des systèmes CBR textuel

Il semble que la complexité/longueur des textes et l'étendue du domaine soient les principaux facteurs à considérer. La diversité du domaine rend difficile l'acquisition de connaissances du domaine et favorise donc l'utilisation de techniques de recherche plus élaborées. Une structuration naturelle des textes facilite l'utilisation de métriques de similarité plus complexes. La concentration de la base de cas dans un domaine pointu avec un vocabulaire restreint oblige le concepteur à indexer et à structurer avec précision chacun des cas. Finalement le niveau d'autonomie du système doit être pris en compte. Un système CBR sans supervision humaine peu difficilement mener une représentation interne élaborée des nouveaux problèmes à résoudre.

Les travaux que nous avons présentés ne nous permettent pas de bien cerner l'influence de ces facteurs. Les applications construites à partir des textes longs portent sur des domaines restreints (SMILE, SPIRE, PRUDENCIA), ce qui favorise une structuration plus élaborée des cas. Et les textes des applications à domaine plus vaste

(CBR-Answers et FAQFinder) sont des FAQs plutôt courts. Or, de plus amples recherches devront être effectuées pour tenter de quantifier chacune de ces dimensions pour des corpus de textes mieux diversifiés.

Bien que déjà mentionné à quelques reprises, il est important de souligner qu'aucun système CBR textuel ne fait d'adaptation de texte. Il y a lieu de se demander pourquoi. Le peu de structuration des solutions dans les documents est habituellement invoqué comme la principale raison motivant l'absence d'adaptation. On pourrait également affirmer que la nature des tâches à accomplir est une autre limitation. Pour plusieurs applications de type "recherche d'information", il est préférable de repérer le cas qui satisfait la tâche à accomplir et de laisser l'usager extraire les informations qui répondent à ses besoins. D'autres applications ne se prêtent pas naturellement à l'adaptation ; par exemple, les applications de jurisprudence. Celles-ci visent à identifier les avis légaux qui peuvent être utilisés pour appuyer un plaidoyer et non pas à rédiger une nouvelle décision légale. Finalement des considérations techniques sont à considérer. L'adaptation de textes devrait reposer sur des techniques de linguistique informatique pour l'analyse syntaxique/sémantique et la génération des textes. Or il y a lieu de croire que la communauté CBR ne maîtrise pas actuellement les outils nécessaires pour aborder ces tâches.

2.5 Autres travaux connexes

2.5.1 *Authoring* de base de cas.

Dans le modèle CBR structurel, des bases de données du domaine peuvent servir de point de départ pour construire la base de cas. Toutefois en l'absence de telles ressources, le processus de construction (*authoring*) est manuel et repose sur des séances d'acquisition de connaissance avec des experts du domaine.

Dans le CBR textuel, nous pouvons tirer profit du fait que les textes contiennent une description des problèmes et des solutions. Le problème de structuration de cas consiste alors à expliciter le contenu de ces textes. Pour cette tâche, des techniques de

traitement de la langue naturelle permettent d'automatiser partiellement ce processus et ainsi réduire l'implication du concepteur du système (voir Tableau 4). Par exemple, le niveau "paires attributs-valeurs" présente une opportunité pour l'utilisation de techniques d'extraction adaptative d'information (voir section 3.8.3).

Plusieurs niveaux de structuration sont possibles (voir Tableau 5 pour une présentation par niveau croissant de structuration). Toutefois, à l'exception de CBR-Answers, les travaux en CBR textuel n'utilisent qu'un seul niveau de structuration : mots-clés seulement (FAQFinder), catégories seulement (SMILE), termes complexes seulement (DRAMA), passages seulement (SPIRE).

Paires attributs-valeurs	Description d'entités nommées et d'événements à partir de techniques d'extraction automatique d'information (voir section 3.8.3).
Catégories et concepts	Obtenus soit manuellement, soit par un processus d'apprentissage automatique ou soit par substitution de concepts plus abstraits tirés d'une taxonomie.
Termes composés et *keyphrases*	Termes du domaine applicatif qui peuvent être obtenus a) manuellement à partir de glossaires et lexiques, ou b) par un traitement automatique tels que la catégorisation de groupe de mots (Gutwin *et al*. 1999, Turney 2000) ou la construction automatique de lexique (Roark & Charniak 1998, Grefenstette 1992).
Mots-clés	Obtenus, comme dans les systèmes de recherche d'information, par découpage et lemmatisation. Sélection basée sur des mesures de fréquence (*tf*idf*) et des listes de mots outils.

Tableau 5 : Explicitation du contenu textuel par niveau de structuration –[11]

Toutefois, nous jugeons qu'une des principales lacunes des approches actuelles en CBR textuel est la représentation uniforme des cas. CBR-Answers permet l'utilisation

[11] Des représentations plus complexes, à l'aide de diagrammes tels que des scripts, des graphes conceptuels ou des cartes cognitives, peuvent également être considérées. Ces représentations peuvent permettre d'atteindre une précision élevée du système en exploitant une représentation élaborée des entités du domaine et de leur relations. Toutefois, ces solutions sont rarement préconisées car il est très difficile de structurer les cas selon ces schémas complexes de représentation. De plus, elles entraînent une dégradation de la performance de la recherche car il est difficile d'établir des similarités entre des diagrammmes.

des différents niveaux pour structurer les cas textuels. Une étude a été menée par Lenz (Lenz *et al*. 1998) pour comparer la performance du système en ajoutant successivement chacun des niveaux. Cette étude indique une amélioration de la précision du système avec l'ajout de chacune de ces composantes.

La littérature CBR offre quelques méthodologies pour la construction de bases de cas. Le projet INRECA propose une méthodologie pour la construction de systèmes CBR industriels (Bergmann *et al*. 1998). Cette approche, qui relève plutôt du génie logiciel, est lourde et décrite à un niveau très général. De plus, elle est dépendante du modèle CBR structurel et elle n'offre pas de ligne directrice pour des problèmes importants tels que la sélection des attributs. Plus récemment, des approches ont été proposées pour l'*authoring* de cas en général. Quelques unes reposent sur la mise en correspondance d'une ontologie du domaine d'application avec une description des fonctions d'un système CBR (Fuchs *et al*. 2001, Diaz-Agudo & Gonzalez-Calero 2001). Cette voie semble difficile à réaliser dans un cadre CBR textuel en raison de l'absence de modèles sous-jacents. D'autres approches ont été proposées pour étendre les travaux de maintenance de cas à la tâche d'*authoring* (McSherry 2001, Smyth & McKenna 1999). Ces travaux présument que les cas sont complètement structurés et n'abordent que la dimension "sélection de cas". Ils ne proposent donc pas de solutions pour structurer des textes non-structurés. Finalement, des techniques de visualisation ont été proposées pour faciliter la construction d'une base de cas (Mullins & Smyth 2001). Cette voie est prometteuse mais elle se limite actuellement aux cas structurels qui peuvent faire l'objet d'adaptation.

La construction d'une base de cas peut se faire selon plusieurs dimensions. Des travaux du domaine de la maintenance de base de cas abordent les dimensions du "nombre de cas" et du "poids des attributs". Des techniques sont proposées pour sélectionner les cas qui devraient faire partie de la base (Smyth & McKenna 1999, Yang & Wu 2000) et pour varier le poids de chacun des attributs des cas (Aha & Breslow 1997). Ces deux dimensions permettent d'ajuster la performance du système (par ex. le

temps de réponse du processus de recherche) et de maintenir la qualité des solutions (par ex. la couverture du système).

Ces techniques de maintenance sont insuffisantes pour le processus d'*authoring* de bases de cas textuels. La maintenance survient lorsque des cas existent, qu'ils sont complètement structurés et qu'ils ont déjà été exploités par le système. Par contre, l'*authoring* survient au début du cycle de vie d'un système CBR. Ceci suggère que la sélection des index et la structuration des cas (création des triplets attribut-valeur-poids) soit centrale au processus d'*authoring*.

2.5.2 Utilisation du CBR en traitement de la langue naturelle.

Une approche de réutilisation de documents a été proposée par Branting et Lester (Branting & Lester 1996) pour la rédaction de nouveaux documents. Afin de rendre les documents auto-explicatifs (*self-explanatory*), chaque cas est structuré selon trois niveaux : le but de rédaction (actes illocutoires), les actions de rédaction (la structure rhétorique) et les exemples de textes pour chaque action. Ce formalisme a été appliqué à la rédaction de textes légaux (par ex. un testament), des textes qui sont habituellement longs et complexes. Ces recherches sont les seules que nous avons répertoriées dans la littérature CBR qui abordent le problème de l'adaptation de textes. L'exploitation de la structure du document et de règles du domaine permettent de modifier le contenu des documents. Cette approche est complexe et difficile à mettre en oeuvre pour des documents courts, tels que ceux retrouvés dans notre application.

Le CBR a aussi été utilisé pour accomplir des tâches de traitement de la langue naturelle. Dans cette littérature, les appellations *instance-based learning* et *memory-based reasoning* sont plutôt adoptées pour désigner l'utilisation de base de cas (Marquez & Padro 2000). Claire Cardie (Cardie 1993, Cardie 1996) fut l'une des premières à aborder les problèmes d'analyse lexicale et sémantique à l'aide de techniques à base de cas. Elle propose, entre autres, une approche pour la sélection d'attributs qui s'appuie sur des arbres de décision.

Le système TIMBL (Daelemans *et al*. 2000), développé à l'Université de Tilburg, est également pertinent. Ce système effectue une compression de la base de cas sous forme de structure d'arbre qui est, par la suite, utilisée pour la classification de nouvelles instances. Le système a été utilisé pour l'analyse lexicale, pour le *chunking* (analyse syntaxique de surface) et pour la détection d'entités nommées.

2.5.3 Extraction d'information

Les techniques d'extraction d'information visent à repérer des informations pertinentes dans des documents pour instancier des canevas structurés tels des *frames* (pairs attribut-valeur). Un système d'extraction d'information est une combinaison d'analyseur lexical, de système à base de règles et/ou de techniques statistiques (Kosseim & Lapalme 1998). Traditionnellement les règles de ces systèmes sont construites manuellement. Des travaux récents explorent la possibilité d'utiliser des techniques d'apprentissage automatique pour l'acquisition de ces règles du domaine (Ciravegna 2000, Ciravegna 2001). Ces techniques se révèlent particulièrement efficaces pour des textes structurés (près de 100% de rappel et précision pour une dizaine d'exemples) ou semi-structurés (plus de 50% de rappel et 80% de précision pour une centaine de documents). Par contre, l'extraction d'information à partir de textes non-structurés (*free text*) se révèle une tâche plus ardue (moins de 50% de précision et rappel pour plusieurs centaines d'exemples, voir Soderland 1999).

Ce domaine est particulièrement intéressant pour le raisonnement à base de cas. Il offre des solutions pour la structuration de portions de cas textuels comportant une régularité. De plus, l'identification d'entités nommées permet de bien identifier les spécificités d'un cas. Finalement les techniques d'extraction adaptative pourraient être étendues à l'acquisition de connaissances permettant l'adaptation de passages textuels.

2.6 Discussion

Suite à ce survol du domaine, nous avons identifié quelques points qui guideront l'orientation de nos travaux. Premièrement, plusieurs de ces approches se prêtent mal à

notre tâche de réponse. Notre tâche repose principalement sur le contenu textuel des messages. Il s'avère donc que les autres champs (adresses courriels et sujet) jouent un rôle peu significatif. Une approche misant sur la catégorisation de textes (comme celle de SMILE) requiert un nombre prédéterminé de thèmes ; or cette exigence convient mal à un domaine qui évoluent comme celui du service aux investisseurs. De plus, une approche de structuration complète des cas (comme celle de Prudencia) exige une structure rhétorique et une homogénéité que nous ne retrouvons pas dans nos cas.

La plupart des approches dans la littérature ne font pas de distinction entre les problèmes et les solutions. Souvent, la notion de solution n'intervient pas dans la définition de la tâche à accomplir et ne semble pas présente dans la description des cas. Or, dans notre application, ces deux composantes sont distinctes et jouent toutes deux un rôle important.

Les travaux sur FAQFinder sont les plus semblables aux nôtres. La tâche du système est de fournir des réponses à des requêtes textuelles. Les cas comportent des descriptions distinctes de problèmes et de solution. Et la longueur des textes est relativement courte. On note toutefois les différences suivantes : a) il y a un déséquilibre entre la longueur de leurs questions et celle de leurs réponses, b) leurs questions portent sur un seul thème, et c) le contenu de leurs réponses est habituellement générique (par opposition aux spécificités que nous retrouvons dans nos messages).

Finalement, la littérature actuelle en CBR textuel ne nous apporte pas de pistes ou de solutions pour entreprendre la réutilisation de messages antécédents.

Chapitre 3 . Recherche de cas pertinents

*Aucune des approches CBR textuel vues au chapitre précédent ne tente d'exploiter les descriptions de solutions dans la phase de recherche. Or, la base de cas que nous utilisons pour notre application de réponse au courrier électronique présente des propriétés qui justifient cet aspect. Dans ce chapitre, nous étudions l'utilisation des relations entre les mots de problèmes et de solutions dans le calcul de similarité des cas et nous estimons les gains potentiels que procure l'insertion de ces relations dans la phase de recherche. Nous utilisons deux approches statistiques, les modèles de cooccurrences et les modèles d'alignement, pour modéliser les relations et nous comparons les résultats obtenus avec un schéma de type tf*idf. Nous concluons ce chapitre par une discussion des autres points ou approches qui mériteraient d'être approfondis dans des travaux futurs.*

La première étape du processus de réponse est de sélectionner un message antécédent pouvant être à la base d'une nouvelle réponse. Tel qu'illustré à la figure 15, le module CBR doit recommander à l'usager les paires de messages <requête, réponse> permettant le mieux d'aborder une nouvelle requête. La qualité du processus de réponse dépend grandement de cette étape car la sélection de messages peu pertinents force l'usager à repérer lui-même les messages qu'il veut réutiliser. Ce qui nécessite des manipulations supplémentaires et entraîne des délais dans la composition de la nouvelle réponse. Pour de courts messages, ces manipulations peuvent s'avérer trop coûteuses et rendre l'utilisation du module CBR moins attrayante. Un bon ordonnancement des cas pertinents est donc important afin d'assurer la production efficace d'une réponse.

D'un point de vue CBR, cette étape correspond à la recherche du (des) cas le(s) plus similaire(s). Plus précisément, le module CBR doit repérer un cas dont la solution est utile pour résoudre le nouveau problème (Figure 16). Notre application présente des particularités intéressantes car les cas sont "fortement textuels", i.e. que les contenus des problèmes et des solutions sont tous deux textuels. De plus, le contenu et la forme narrative des solutions sont tous deux importants pour l'approche CBR de résolution de problèmes.

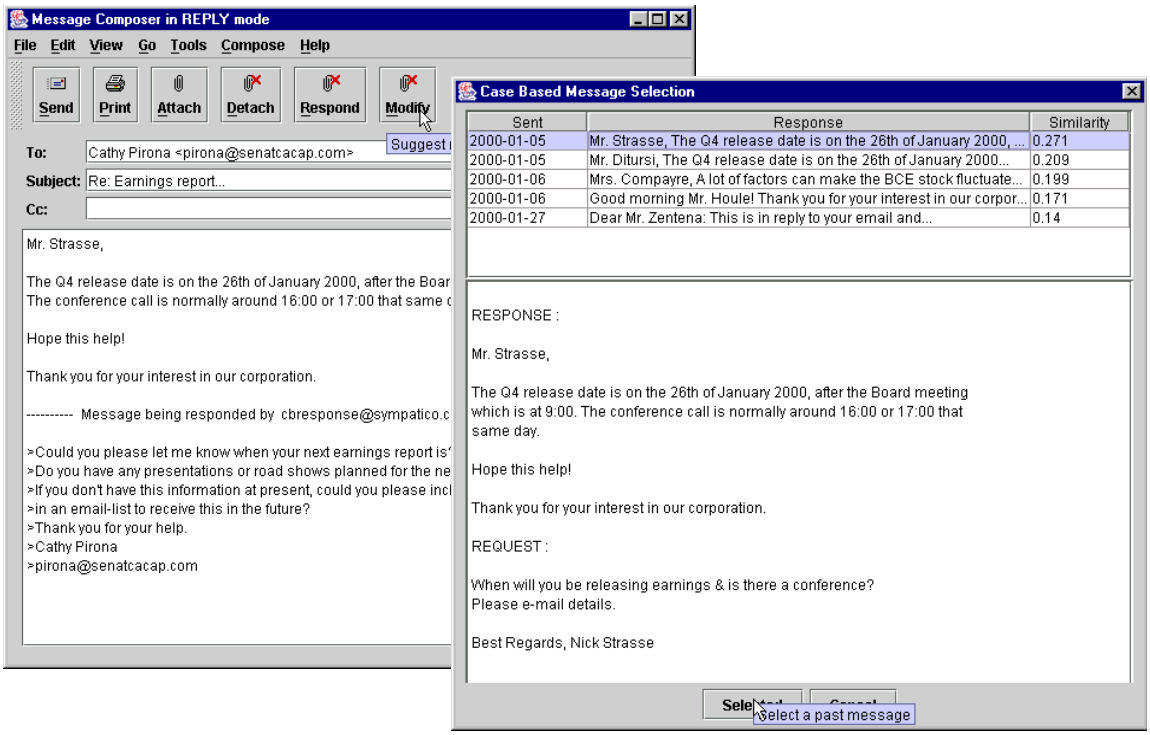


Figure 15 : Sélection de messages antécédents

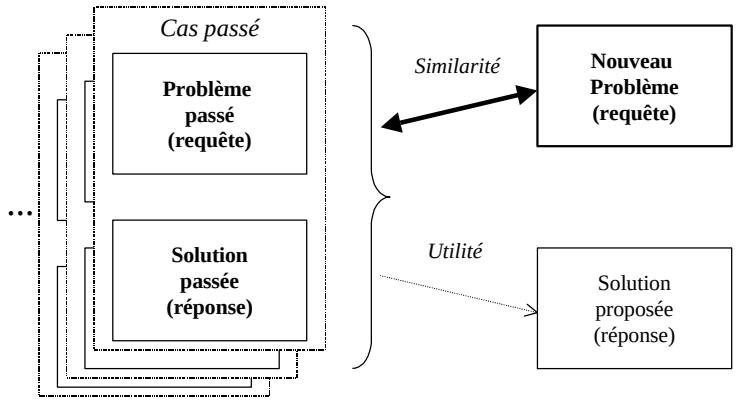


Figure 16 : Similarité des problèmes et utilité des solutions

Pour estimer le niveau de performance pouvant être atteint dans cette étape, nous avons mené quelques expérimentations avec un schéma simple d'estimation de similarité, un cosinus de vecteurs de termes. Ces expérimentations ont indiqué qu'un tel schéma de similarité entre requêtes donne une précision approximative de 57%. Cette performance ressemble aux résultats obtenus pour des expérimentations similaires avec des foires aux questions (Burke *et al*. 1997). Pour obtenir une indication de l'utilité des réponses, nous avons répété les mêmes expérimentations mais en comparant les solutions au lieu des

requêtes. La précision du système augmente à plus de 72%, soit une amélioration de 25%. Une description plus détaillée de ces expérimentations est présentée à la section 3.5.

Ce résultat indique que l'utilisation des solutions pourrait apporter une contribution à la phase de recherche. Dans ce chapitre, nous explorons cet aspect et nous apportons des éléments de réponses aux trois questions suivantes :

- étant donnée une nouvelle requête, comment établir des liens entre celle-ci et une solution ? Nous nous intéressons plus particulièrement aux relations lexicales entre problèmes et solutions ;
- quelles stratégies permettent d'incorporer ces relations dans le processus de recherche ?
- quels gains apportent ces stratégies et quelles sont les raisons qui expliquent ces gains ?

3.1 Exploitation des solutions en CBR textuel

Le CBR textuel est l'exploitation de cas décrits dans des documents textuels. La plupart des études dans ce domaine ont porté sur les descriptions textuelles de problèmes avec peu d'attention accordée aux solutions. L'ordonnancement de cas textuels déterminé par la pertinence avec un nouveau problème cible est normalement estimé par la similarité des descriptions de problème. Ce choix est justifié, en partie, par l'hypothèse CBR que la similarité des problèmes est garante de l'utilité des solutions.

Toutefois un cadre textuel impose une révision de cette hypothèse. La faible structuration des descriptions n'apporte aucune garantie que les similarités entre cas sont mieux capturées par les problèmes que par les solutions. Ceci se produit lorsque l'uniformité dans la rédaction des solutions est plus grande que celle des descriptions de problèmes. Par exemple, dans le domaine de la relations aux investisseurs, les solutions sont écrites par un nombre limité d'analystes financiers tandis que les problèmes sont soumis par divers investisseurs (corporatifs et individuels) ayant différents niveaux

d'expérience et de formation sur le marché financier. L'utilisation des solutions permettrait une comparaison plus homogène des cas. Cette uniformité favorise la prise en compte des réponses textuelles pour établir la similarité durant la phase de recherche.

De plus, plusieurs éléments de la relation entre un problème et une solution peuvent être exploités dans le cycle CBR. Par exemple, lorsque la formulation des solutions calque la description des problèmes, la solution textuelle est rédigée pour aborder les différentes portions d'une description de situation. Conséquemment, une correspondance peut être établie entre des paragraphes, phrases, groupes syntaxiques et termes de chacune des composantes d'un cas. Des bénéfices peuvent à nouveau être anticipés suite à l'exploitation de ces correspondances.

Finalement, il se peut que l'on désire prendre en compte des propriétés des descriptions de solutions. Dans le cadre de la réponse au courrier électronique, un usager pourrait juger préférable de sélectionner les solutions les plus simples à réutiliser. À problème équivalent, on tirerait profit du fait que les descriptions plus courtes sont plus simples à modifier. Ainsi la phase de recherche viserait à établir un compromis entre la similarité des problèmes et la complexité des solutions. Un tel formalisme a été défini pour le CBR structurel (Smyth & Keane 1998).

Peu de travaux en CBR tentent d'exploiter dans la phase de recherche des informations provenant de descriptions de solutions. Il semble que l'on présume que l'absence d'éléments de solution dans la description du nouveau problème rend impraticable l'utilisation de solutions antérieures. Nous proposons dans les prochaines sections des techniques permettant l'insertion d'éléments de solutions dans la recherche.

3.2 Insertion des solutions dans la phase de recherche

Afin d'incorporer la solution d'un cas dans la phase de recherche, une approche naïve serait de fusionner dans une même représentation interne les mots provenant du problème et de la solution du cas. On peut alors utiliser cette structure agglomérée pour estimer la similarité entre les cas. Toutefois, quelques limitations sont anticipées pour

cette approche. Premièrement, il n'est pas garanti que le même vocabulaire décrive les problèmes et les solutions. Ces composantes peuvent être rédigées par différents intervenants ayant diverses formations et pouvant présenter différents points de vue. Par exemple, un investisseur débutant sollicitant de l'aide pourrait difficilement référer à des indicateurs financiers appliqués par des analystes professionnels. Également, une description de situation peut ne pas aborder directement une solution. Par exemple, la réponse à une requête portant sur des aspects financiers détaillés peut être de consulter un site web ou de lire des documents. Ainsi, afin d'exploiter de telles caractéristiques de cas, nous avons besoin d'approches et de modèles pour représenter le transfert lexical entre les descriptions de problèmes et les composantes de solutions.

3.2.1 Approches proposées

Nous considérons deux approches pour insérer les solutions dans la phase de recherche : l'expansion de solutions et l'estimation de l'utilité des solutions (Figure 17).

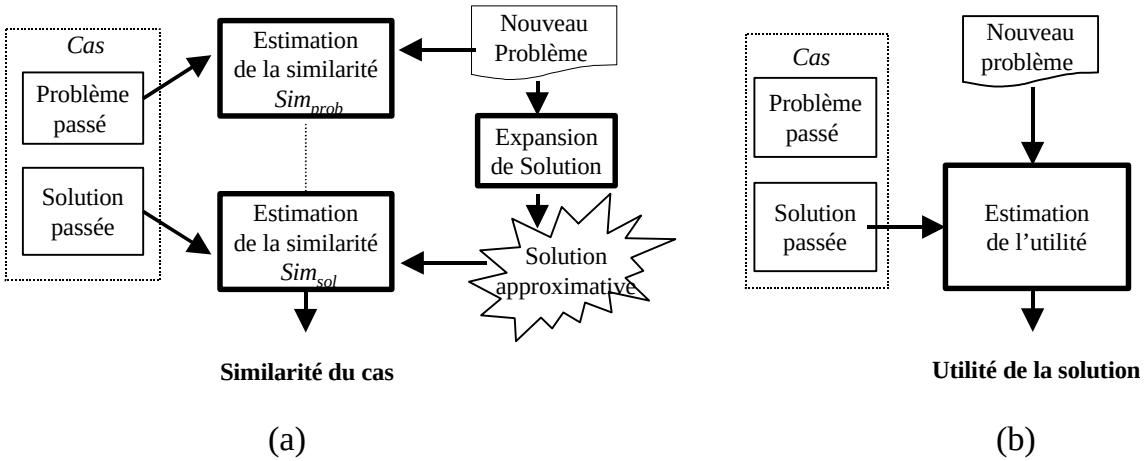


Figure 17 : Approches pour insérer les solutions dans la phase de recherche : (a) expansion de solution et (b) estimation de l'utilité d'une solution

La première approche s'inspire des techniques d'expansion de requêtes en recherche d'information et consiste à générer une structure utilisée pour mener la phase de recherche. Cette structure représente une approximation du contenu d'une solution pouvant correspondre au nouveau problème. Cette structure n'a pas à être exhaustive (c.-

à-d. elle ne contient pas tous les termes possibles) ni même structurée (ex. l'ordre des termes). Elle sert uniquement à décrire les attentes du système par rapport au contenu de la solution nécessaire pour résoudre un problème. Cette approximation peut alors être utilisée pour balayer la base de cas et établir un ordonnancement des cas de la base. La similarité de chaque cas repose en partie sur la comparaison entre sa solution et celle résultant de l'expansion. La similarité globale peut alors être une combinaison des similarités entre les problèmes et entre les solutions (dont l'une est une approximation).

La deuxième approche consiste à évaluer l'utilité de la solution d'un cas antécédent par rapport à un nouveau problème. Étant donné un nouveau problème, la base est balayée en appliquant la fonction d'utilité à chacun des cas et la solution obtenant la meilleure évaluation est retenue. L'avantage de cette approche est qu'elle ne requiert pas la génération d'une structure artificielle. Toutefois, il est nécessaire de trouver une approche qui permet de capturer la notion d'utilité.

3.2.2 Modélisation des relations entre les descriptions

Ces deux approches nécessitent des modèles pour générer une approximation de solution ou estimer son utilité. Nous abordons la construction de ces modèles d'un point de vue lexical. Nous tenons à éviter les approches qui reposent sur une description sémantique des cas car elle nécessite une structuration manuelle de leur contenu textuel. De plus, chaque nouvelle description de problème soumise au processus de résolution doit faire l'objet du même niveau de structuration, ce qui entraîne des manipulations de l'usager et de délais de traitement.

Dans la littérature CBR textuel, la modélisation des relations lexicales entre cas repose uniquement sur la notion de similarité. La plupart de ces efforts sont basées sur une représentation vectorielle de cas comportant des mots-clés (Burke *et al*. 1997), des n-grammes de caractères (Aha *et al*. 2001) et des termes complexes (Wilson & Bradshaw 2000). L'approche la plus répandue pour établir la similarité entre les descriptions de problème et les cas candidats est d'utiliser un cosinus des vecteurs de termes. Toutefois,

cette approche présente quelques limitations puisqu'elle exige la correspondance exacte entre les termes (ou les n-grammes).

Pour surmonter cette contrainte, quelques auteurs (Burke *et al*. 1997, Brüninghaus & Ashley 1999) font usage des relations sémantiques entre mots. Ils utilisent des ressources linguistiques (ex. thesaurus) pour établir la similarité sémantique de mots différents dont les significations sont reliées entre elles. Bien qu'on ait observé des améliorations, cette approche peut également causer quelques problèmes puisque la notion de similarité sémantique est plutôt difficile à établir. De plus, cette approche ne se transpose pas facilement à notre problème puisque, à notre connaissance, aucune ressource spécifique au domaine du service aux investisseurs n'est actuellement disponible. Des ressources plus génériques, comme WordNet, ne recouvrent pas tellement bien les termes contenus dans notre corpus de messages. Approximativement 38% des termes du corpus, tels que de la terminologie financière où des noms de compagnies, ne sont pas inclus dans ce thesaurus. De plus, la similarité sémantique entre les termes ne donne aucune indication sur l'utilité des mots que l'on retrouve dans les solutions.

Toutefois, puisque notre base de cas est relativement substantielle, nous pouvons obtenir à l'aide de méthodes expérimentales, une estimation de ces relations sous la forme d'associations entre des mots. Un avantage d'utiliser ces associations, sélectionnées soit par des tests statistiques ou par des mesures de probabilités, est leur représentativité du domaine de discours.

L'idée principale sous-jacente à cette approche est qu'un cas textuel représente la conversion lexicale d'une description de problème en une description correspondante de solution. La base de cas forme donc un corpus de texte parallèle (un bitexte) qui décrit un *mapping* du langage des problèmes (requêtes) au langage des solutions (réponses).

Notre but est d'évaluer les bénéfices d'aligner des descriptions de problèmes et de solutions durant le processus de recherche, et de comparer cette approche avec celle plus communément utilisée pour la recherche en CBR textuel.

Des méthodes permettent de découvrir des associations, représentées par des modèles statistiques, entre des paires de mots provenant des descriptions de problèmes et solutions. Nous proposons d'utiliser deux techniques tirées du traitement statistique des langues naturelles pour déterminer quelle approche donne les résultats les plus prometteurs pour améliorer la phase de recherche. Ces modèles sont les cooccurrences de mots et les alignements de traduction. Les cooccurrences de mots reposent sur l'hypothèse que la présence des mots du problème influence l'utilisation d'autres mots dans la solution. Les alignements de traduction impose une relation plus contraignante puisque chaque mot d'un problème est présumé être la traduction directe d'un seul mot de solution. Ces modèles sont décrits dans les sections suivantes.

Ainsi, on peut concevoir que l'utilité d'une solution est composée de la similarité entre les problèmes et du degré de correspondance entre le problème et la solution du cas antécédent (Figure 18). Ainsi le potentiel de réutilisation d'un cas dépend de la pertinence de sa description de problème et de l'adéquation de la solution correspondante.

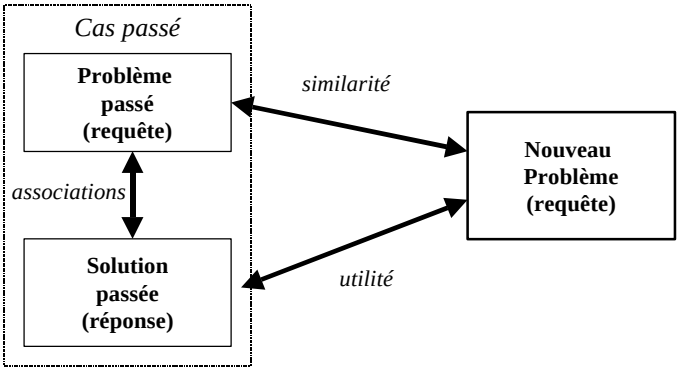
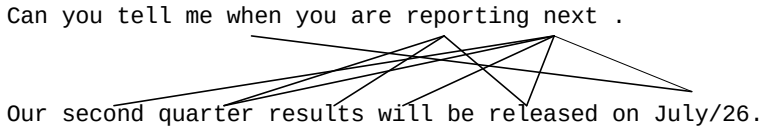


Figure 18 : Utilité = similarité + degré d'association

Dans notre cadre de travail, nous supposons que les cas textuels sont hétérogènes, i.e. qu'ils décrivent diverses situations et proposent des solutions peu similaires. Notre utilisation des associations lexicales pour discriminer entre des cas pourrait avoir un impact limité sur une base de cas homogène où l'on répéterait la majorité des mots dans la plupart des cas. Ces bases de cas devraient plutôt reposer sur l'utilisation de techniques misant sur la structuration de textes, comme l'extraction d'information.

3.3 Exploitation de cooccurrences de mots

La première approche que nous utilisons pour modéliser les associations entre les problèmes et les solutions est basée sur les cooccurrences de mots. Les cooccurrences indiquent que la présence d'un mot spécifique à un problème devrait augmenter la vraisemblance de retrouver un autre mot spécifique dans sa solution. Ainsi un mot de problème peut influencer l'occurrence de plusieurs mots de solutions, et inversement. Par exemple, dans la paire de phrase problème-solution suivante,



les cooccurrences capturent, à différents degrés d'intensité, que "*reporting*" est relié à la divulgation ("*released*") de résultats ("*results*") et que "*July/26*" est la date quand ("*when*") le prochain ("*next*") rapport sera disponible.

Tel qu'illustré à la Figure 19, nous cherchons à obtenir, pour chaque mot présent dans les requêtes, une liste de mots susceptibles d'être retrouvés dans les réponses correspondantes. Pour obtenir ces listes, nous traitons le contenu textuel des messages en découpant chacun des termes individuels (tokénisation), en identifiant leur racine morphologique (lemmatisation), et en leur attribuant une étiquette lexicale (*tagging*).

À partir des textes lemmatisés, nous comptons la fréquence de toutes les paires de mots (w_i, w_j) qui viennent respectivement des problèmes et de leurs solutions correspondantes. Une réduction du vocabulaire est effectuée pour ne retenir que les paires qui présentent un intérêt pour notre tâche de réponse et pour diminuer la taille de la matrice de compte mots-mots. Les termes du vocabulaire sont choisis en fonction de leur fréquence dans le corpus et de leur étiquette lexicale. La fréquence est un moyen d'éliminer des termes qui sont circonstanciels et qui n'impliquent pas de similarité entre les descriptions. Par exemple, l'occurrence d'un même prénom de personne dans deux messages différents apporte peu d'information sur leur similarité. Pour les catégories lexicales, nous nous intéressons plus particulièrement aux noms, verbes, adjectifs,

adverbes (temps, lieu), quantités et à leurs relations dans les requêtes-réponses. Les mots de ces catégories véhiculent l'essentiel du contenu des descriptions de cas et sont les plus susceptibles de se retrouver impliqués dans des associations de mots.

Par la suite, nous sélectionnons les paires de mots les plus significatives en nous basant sur une mesure d'information mutuelle (Manning & Schütze 1999). L'information mutuelle, exprimée par l'équation suivante, indique la quantité d'information qu'un mot apporte à un autre :

$$I(w_i, w_j) = \log \frac{P_{cooc}(w_i, w_j)}{P_{prb}(w_i) P_{sol}(w_j)}$$

où $P_{cooc}(w_i, w_j)$ est la proportion de cooccurrences qui comportent respectivement le terme de problème w_i et le terme de solution w_j (c.-à-d. le nombre d'occurrences conjointes de w_i et de w_j divisé par le nombre total de cooccurrences pour tous nos cas). $P_{prb}(w)$ et $P_{sol}(w)$ sont respectivement la probabilité de retrouver le terme w dans les problèmes et dans les solutions. La mesure d'information mutuelle est reconnue pour favoriser les paires dont les termes sont moins fréquents mais donne toutefois une bonne indication du degré de dépendance entre les termes.

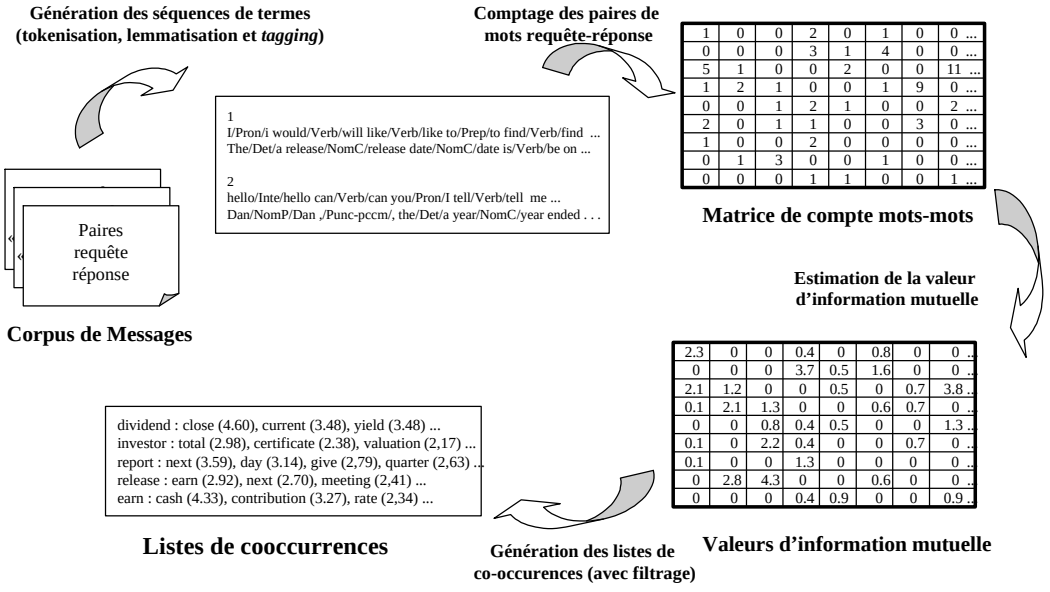


Figure 19 : Étapes de la génération des listes de cooccurrences

Les listes de cooccurrences obtenues contiennent un sous-ensemble des mots de solution associés à chacun des termes de problème. Pour chaque mot de problème w_i, les différents w_j sont ordonnancés en ordre décroissant d'information mutuelle. Les listes sont ensuite tronquées en utilisant des seuils basés sur la valeur d'information mutuelle et sur le nombre de mots associés à chacun des termes.

Pour insérer ces cooccurrences dans la phase de recherche, nous préconisons une approche d'expansion (voir Figure 20). Le problème cible est converti en solution approximative en cumulant des termes provenant des listes de cooccurrences. Pour chaque mot w_i du problème cible, on ajoute à la solution approximative les termes w_j de sa liste de cooccurrences. La solution approximative correspond aux termes des listes de cooccurrences pour chacun des mots présents dans le problème cible $Prob_{cible}$ c.-à-d.

$$Solution\_approximative = \left\{ w_j \middle| \exists w_i \in Prob_{cible} \text{ et } cooc(w_i, w_j) > 0 \right\}$$

où *cooc(i,j)* est une fonction binaire indiquant si le terme w_j fait partie de la liste du terme de requête w_i.

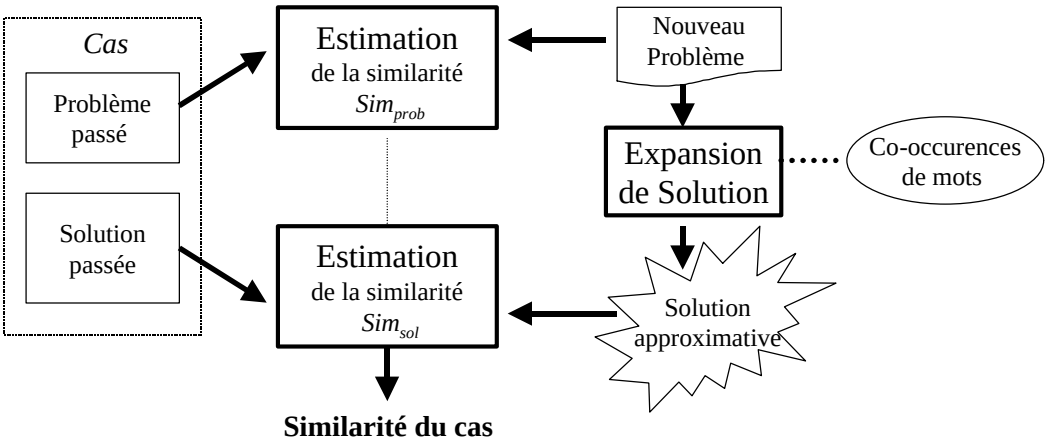


Figure 20 : Expansion de solution avec des listes de cooccurrences

Cette opération fournit un sac de mots, une représentation vectorielle donnant l'importance de chaque mot dans la solution approximative. Pour quantifier cette

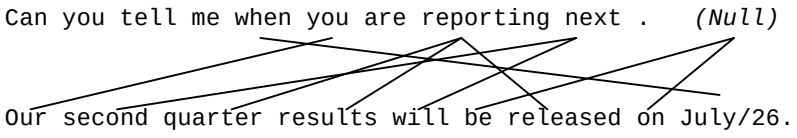
importance, nous associons à chaque mot w_j du vecteur un poids $poids_{approx}$ correspondant à :

$$poids_{approx}(w_j) = \sum_{w_i \in Prob_{cible}} cooc(w_i, w_j) \times poids_{problème}(w_i)$$

où $poids_{problème}(w_i)$ est le poids du terme w_j dans la requête. Lors de nos expérimentations, nous avons fait des essais avec différents jeux de poids (fréquence, *tf*idf*, normalisation). Nous avons obtenu de meilleurs résultats lorsque $poids_{problème}$ est un *tf*idf* normalisé et que $poids_{approx}$ est simplement normalisé. Nous présentons à la section 3.5.4 les résultats obtenus en appliquant ce schéma à notre base de cas.

3.4 Estimation avec un modèle de traduction

Le deuxième modèle que nous étudions, les modèles d'alignement, est utilisé en traduction automatique statistique. Le concept d'alignement entre deux phrases consiste à déterminer, pour chaque mot d'une phrase, le mot de l'autre phrase dont il est issu. Pour reprendre notre exemple précédent, une approche de traduction impose que chaque mot de la deuxième phrase (la phrase cible) soit généré à partir d'un seul mot de la première phrase (la phrase source). Ces modèles permettent aussi qu'un mot cible soit généré sans mot source (le mot *null*) pour des fins grammaticales.



Pour transposer cette idée au CBR textuel, nous pouvons imaginer qu'il existe une langue pour décrire les problèmes et une autre pour décrire les solutions. Ainsi un cas peut être visualisé comme la traduction d'un problème particulier dans la langue des solutions. Le modèle qui gouverne cette traduction, appris à partir de la base de cas, peut être utilisé pour ordonnancer les solutions précédentes selon leur pertinence à un

problème cible. Toutefois la génération de nouvelles descriptions de solutions à partir de ces modèles n'est pas envisagée, car ce serait pousser l'analogie beaucoup trop loin.

Pour calculer ces modèles, plusieurs approches ont été proposées en traitement des langues naturelles. Les modèles IBM (Brown *et al*. 1993) ont été développés pour l'apprentissage, à partir d'un corpus, de probabilités nécessaires pour établir l'alignement de textes parallèles. Ces modèles sont de complexité croissante et prennent en compte différents facteurs tel que la génération de mots multiples, la distorsion de la position des mots et le regroupement de mots. Pour mener nos expérimentations, nous utilisons le modèle IBM1 qui peut être formulé tel que suit dans le cadre CBR. La probabilité de trouver un problème *Prb* étant donné une solution *Sol* est :

$$p(Prb \mid Sol) = \sum_{a} p(Prb \mid a, \mathrm{Sol}) p(a \mid Sol)$$

Ceci correspond à la probabilité d'obtenir une séquence *Prb* parmi tous les alignements possibles α avec la séquence *Sol*. Après quelques manipulations et simplification, la probabilité conditionnelle de ce modèle est exprimée comme suit :

$$p(Prb \mid Sol) = \frac{e}{(l+1)^m} \prod_{j=1}^{m} \sum_{i=1}^{l} t(Prb_j \mid Sol_i) \qquad (1)$$

où les séquences *Prb* and *Sol* contiennent respectivement *m* et *l* mots. Cette expression est utilisée pour l'apprentissage du modèle de traduction. Le résultat du processus d'apprentissage est un tableau de transfert *t* qui donne les probabilités de générer un mot cible $Prob_j$ étant donné le mot Sol_i dans la description source. Le modèle de transfert peut être obtenu en appliquant un algorithme EM qui itérativement assigne et révise les probabilités des paramètres du modèle jusqu'à ce que la convergence soit atteinte.

Pour utiliser ces alignements dans la phase de recherche, nous préconisons une approche d'évaluation d'utilité. L'expression (1) est utilisée pour l'ordonnancement des solutions lors du balayage de la base de cas. Elle ne mesure pas la similarité entre deux textes mais donne plutôt la probabilité que l'un découle de l'autre. Ainsi, cette mesure

probabiliste donne le degré d'association entre un problème et une solution, une forme d'utilité d'une description par rapport à l'autre.

L'approche de générer les problèmes à partir des solutions et non l'inverse peut sembler contre-intuitive. Toutefois, mis à part quelques aspects techniques reliés au modèle du canal bruité sous-jacent à cette approche, il est à noter que les probabilités sont multiplicatives et que la comparaison de solutions de longueurs différentes favoriserait celles ayant moins de termes. Ainsi, en comparant les solutions pour leur contribution à une description de problème de longueur fixe, nous nous assurons qu'elles sont évaluées sur une base commune.

3.5 Expérimentations

3.5.1 Démarche préconisée

Nous présentons les résultats d'une comparaison du modèle de cooccurrences, du modèle de traduction et de la similarité *tf***idf* fréquemment utilisé en CBR textuel. Les résultats ont été obtenus par une évaluation de type *leave-one-out*, pour laquelle un cas est retiré de la base avant l'apprentissage des modèles et utilisé pour évaluer la phase de recherche.

Pour mener nos expérimentations, nous utilisons un sous-ensemble de notre corpus de messages du domaine du service aux investisseurs, 102 paires de messages regroupés sous le thème "*financial information*". Chaque paire de messages représente un échange entre un investisseur et un analyste financier de BCE. Les messages couvrent une variété de sujets dont des requêtes portant sur des documents, des résultats financiers, le comportement boursier et des événements corporatifs. La longueur des messages individuels varie de quelques mots à plus de 200 avec une moyenne d'environ 87 mots. Les réponses, fournies par quelques 5 à 10 analystes, sont plus uniformes dans leur format et leur structure que les requêtes provenant de différents investisseurs. La plupart des messages sont bien rédigées et utilisent un vocabulaire adéquat.

Pour obtenir une meilleure appréciation des résultats, nous avons subdivisé ce corpus de test en quatre sous-groupes :

- Divulgation de résultats financiers et appels conférence : ce sont des requêtes de routine. Les messages sont uniformes et utilisent un vocabulaire restreint.

- Requêtes portant sur des aspects financiers : les messages sont plus diversifiés et peuvent contenir des requêtes détaillées et des explications variant d'un message à l'autre. Quelques réponses à des messages spéculatifs sont génériques et les abordent indirectement (*ex.* "*consult our web site*").

- Listes de distribution : des requêtes d'investisseurs pour se joindre à des listes de distribution. Toutefois l'adhésion à ces listes n'est pas toujours confirmée dans les réponses.

- Autres messages : ils portent sur des sujets non similaires, avec peu d'antécédents pouvant être utilisés comme base pour formuler une nouvelle réponse.

Pour comparer les approches, nous utilisons les trois critères présentés au Tableau 6.

Mesure	Définition	Importance pour l'application
Rang moyen	la position du premier cas pertinent dans la liste des plus proches voisins	Le nombre de messages qu'un usager doit lire, si la liste est consultée chronologiquement, pour repérer une base de réponse.
Première position	le pourcentage d'essais pour lesquels le plus proche voisin est pertinent	L'aptitude du système à sélectionner une réponse pertinente sans l'aide de l'usager (i.e. en choisissant par défaut le cas le plus similaire).
Précision	la proportion de cas pertinents dans les k plus proches voisins (k=5).	Indique la proportion de cas recommandés pouvant être utilisée par l'usager pour formuler une réponse.

Tableau 6 : Critères d'évaluation des cas sélectionnés

3.5.2 Structuration des cas

Pour cette étude, nous souhaitons évaluer la capacité du module CBR à sélectionner les cas en se basant seulement sur leur contenu textuel. Ainsi nous supprimons les en-têtes des messages (ex. date, sujet, récipiendaire…) et les parties du corps du message qui ne sont pas textuelles. Les textes sont tokénisés, les termes sont associés à une étiquette lexicale et leur racine morphologique est obtenue par lemmatisation. Des logiciels de notre laboratoire (*Tok*, *Lmtag* et *Lemmatize*) ont été utilisés pour faire ces traitements.

Également, pour favoriser la comparaison des messages sur une base commune, nous avons réduit la spécificité des messages en remplaçant les dates, numéros de téléphone, URL et adresses courriel par une étiquette sémantique (*ex.* DATE). Ceci nous permet également d'évaluer le rôle que joue ces informations dans les associations requêtes-réponses.

Pour la représentation interne des cas, nous conservons une représentation vectorielle et la séquence des termes lemmatisés pour les descriptions de problème et de solution. Pour réduire le vocabulaire de la base de cas, nous filtrons les termes en nous basant sur leur fréquence dans le corpus et sur l'utilité de leur catégorie lexicale. Pour les résultats présentés dans ce chapitre, nous avons conservé les termes ayant une fréquence supérieure à 3 et dont la catégorie lexicale est un nom, un verbe, un adjectif ou un adverbe.

3.5.3 Similarité avec *tf*idf*

Une approche fréquemment préconisée pour mesurer la similarité textuelle est de comparer les représentations vectorielles d'un nouveau problème cible *Prb* avec la description de problème d'un cas *C*. A chaque terme des vecteurs est affecté un poids *w(t)* déterminé à partir de la fréquence dans la description (*tf*) et sa distribution relative dans la base de cas (*idf*) exprimé comme suit :

$$idf(t) = \log\left(\frac{|CB|}{|CB:\ \text{où}\ \ tf(t) > 0|}\right)$$

où |*CB*| représente la taille de la base de cas.

La similarité locale est restreinte aux termes identiques et la similarité globale est déterminée par un cosinus normalisé des deux vecteurs :

$$sim(Prb, C) = \frac{\sum_t w_{Prb}(t) \times w_C(t)}{\sqrt{\sum_t w_{Prb}(t)^2 \times \sum_t w_C(t)^2}}$$

Avec ce schéma de recherche, nous obtenons les résultats suivants pour notre corpus (Tableau 7) :

Groupe	**Rang moyen**	**Premier**	**Précision**
Tous	1.952	58.7%	57.3%
A	1.080	92.0%	80.0%
B	2.385	51.7%	51.0%
C	1.550	77.7%	62.2%
D	3.000	33.3%	30.3%

Tableau 7 : Résultats avec *tf*idf*

Ces résultats indiquent que la précision globale est approximativement 57% et que le plus proche voisin est pertinent dans près de 59% des essais. Un cas pertinent est en première position la plupart du temps pour les groupes A et C. Ces descriptions de cas sont des messages de routine ayant un vocabulaire restreint. Toutefois, des performances inférieures sont observées pour les groupes B et D, qui recouvrent un registre plus large de sujets.

Pour évaluer les bénéfices potentiels d'utiliser des associations de mots dans le processus de recherche, nous avons répété cette expérimentation en utilisant les solutions au lieu des problèmes. Une précision globale de 71.9% a été atteinte. Le plus proche voisin était presque toujours pertinent pour les groupes A et C. Le rang moyen a été

significativement réduit pour les groupes B (1.429) et D (2.083). Ceci fournit donc des motivations pour incorporer les solutions dans la recherche de cas.

3.5.4 Cooccurrences de mots

Les résultats obtenus en générant une solution approximative à partir des listes de cooccurrences non-filtrées sont présentés dans le tableau suivant (Tableau 8) :

Groupe	Rang moyen	Premier	Précision
Tous	2.048	68.3%	60.5%
A	1.640	68.0%	69.8%
B	1.333	83.3%	80.0%
C	1.833	83.3%	66.7%
D	4.000	25.0%	35.0%

Tableau 8 : Résultats avec des listes de cooccurrences

Les résultats indiquent que ce schéma d'expansion améliore légèrement la précision globale[12] (60.5% vs. 57.3%) de la phase de recherche et préserve le rang de la première solution pertinente dans la liste de similarité (2.048 vs. 1.952). Bien que la précision globale soit supérieure au *tf***idf* et que la pertinence du plus proche voisin soit améliorée pour les groupes B et C, on observe une détérioration significative de la performance pour le groupe A.

Nous pouvons expliquer ce résultat par trois observations. Premièrement, le modèle de cooccurrence peut détecter si une solution est commune à deux problèmes différents (par exemple, un message générique utilisé pour répondre à différentes requêtes spéculatives). Ce comportement explique en partie les améliorations pour le groupe B.

[12] La précision est estimée comme le pourcentage de réponses pertinentes contenu dans les k plus proches voisins (dans nos expérimentations k=5). Les résultats présentés pour le schéma d'expansion sont basés seulement sur la similarité entre les solutions (i.e. les réponses).

Deuxièmement, les cas ayant des descriptions de problèmes plus longues tendent à obtenir de meilleures évaluations. Plus de termes dans les descriptions mènent à des approximations plus étoffées et ainsi recouvrent plus de situations possibles. Même après normalisation, on retrouve ces solutions à un rang plus élevé. Dans notre base de cas de test, les descriptions les plus longues sont peu reliées aux autres cas. Ainsi leur présence dans plusieurs des listes de plus proches voisins explique en partie la dégradation observée pour le groupe A.

Notre troisième observation est que bien que le modèle de cooccurrence introduise des mots valables dans les solutions approximatives, il introduit également du bruit. Quelques exemples de listes de cooccurrences reliées au groupe A (sujet - "*release of earnings report*") sont présentés dans le tableau suivant (Tableau 9). Les listes pour "*release*" et "*report*" contiennent des associations qui sont représentatives des discussions de ce groupe de message (*ex.* "*schedule*", "*conference*", références temporelles). Toutefois, ils introduisent des termes moins pertinents comme "*far*", "*also*" and "*detail*". Les résultats obtenus pour "*Earnings*" révèlent également quelques limitations. Ce terme est largement répandu dans la base de cas et est associé à plusieurs différentes paires de cooccurrences. La liste résultante contient quelques associations qui pourraient contribuer à la recherche de cas du groupe B mais pas à ceux du groupe A.

Mot de problème	**Liste correspondante de cooccurrences (mots de solution)**
Release	*earn, BCE_Emergis, CGI, EMAIL_ADDRESS, next, meeting, schedule, release, TIME, conference* **…**
Earnings	*EPS, reflect, study, read, such, accounting, note, analysis, next, holding, prior, item, give, also* **…**
Report	*next, day, give, quarter, far, TIME, DATE, detail, release, afternoon, after, reply, also, date, corporation, number* **…**

Tableau 9 : Exemples de listes de cooccurrences

L'approche d'expansion de solution par cooccurrence est paramétrique car les listes peuvent être tronquées à l'aide de valeurs de seuil sur la fréquence de cooccurrence et la valeur d'information mutuelle. Les résultats présentés précédemment proviennent de listes sans troncature. Tel qu'illustré dans les trois diagrammes suivants (Figure 21), le choix de ces seuils peut rendre cette approche bien plus avantageuse que celle de *tf*idf*.

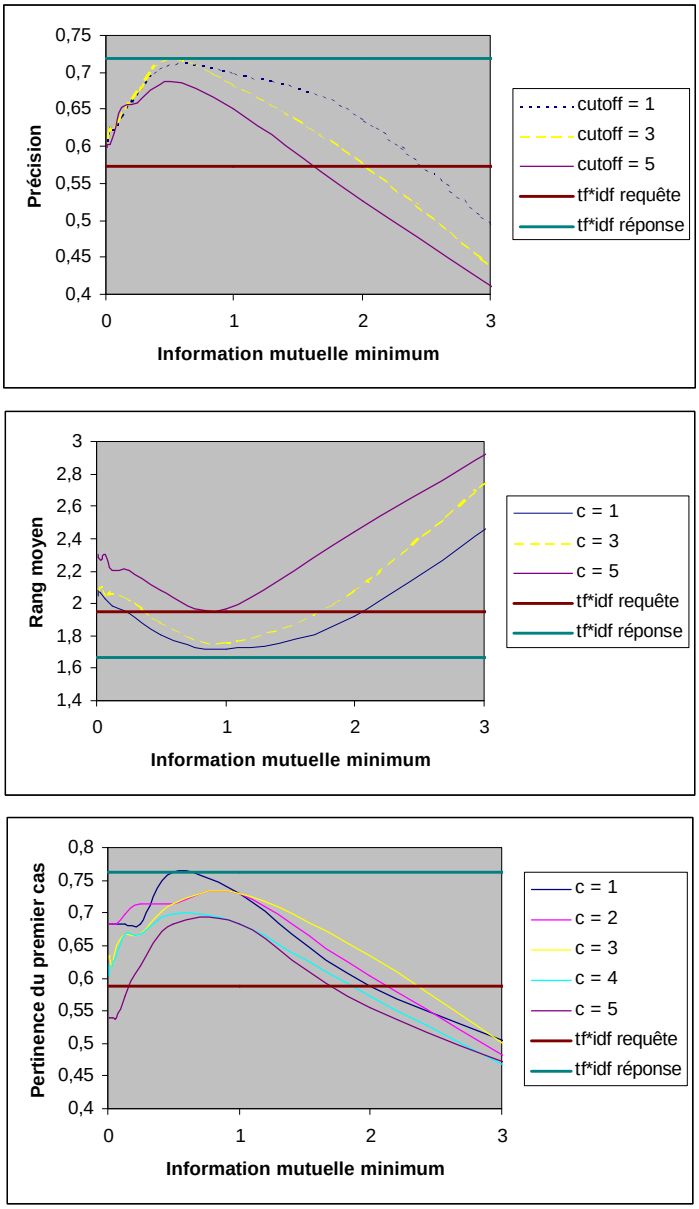


Figure 21 : Courbes des indicateurs en fonction du seuil d'information mutuelle

Un choix judicieux de paramètres permet d'atteindre des valeurs de précision et de rang moyen qui se rapprochent du niveau optimum, le *tf*idf* obtenu à partir de la similarité des requêtes. On observe que les valeurs optimales sont obtenues pour un seuil d'information mutuelle variant entre 0.5 et 1.0. Ainsi on élimine les paires de mots qui sont peu dépendantes entre elles et qui amènent un certain bruit dans les approximations de solutions. On note également que le filtrage de cooccurrences dégrade la performance

du système. Ainsi il vaut mieux conserver les paires de mots peu fréquentes car elles apportent de l'information qui aide à repérer les cas pertinents.

L'expansion par cooccurrence s'avère donc une approche intéressante, surtout lorsqu'un filtrage adéquat est appliqué. Elle permet également de combiner la similarité des solutions avec celles des problèmes, ce qui peut amener d'autres améliorations aux performances du système. Toutefois, il est difficile d'interpréter la nature des associations entre les mots de problèmes et de solutions. Ceci peut être dû à la taille restreinte de notre base de cas.

3.5.5 Modèle de traduction

Pour ces expérimentations, nous avons obtenu une table de transfert IBM1 en utilisant le logiciel GIZA++ (Och & Ney 2000) sur notre base de cas de test. En utilisant ce modèle, l'évaluation de l'utilité de solutions nous donne les résultats suivants :

Groupe	Rang moyen	Premier	Précision
Tous	1.721	63.9%	56.9%
A	1.320	76.0%	74.4%
B	1.464	75.0%	58.6%
C	2.000	66.7%	60.0%
D	4.000	33.3%	25.0%

Tableau 10 : Résultats avec un modèle de traduction

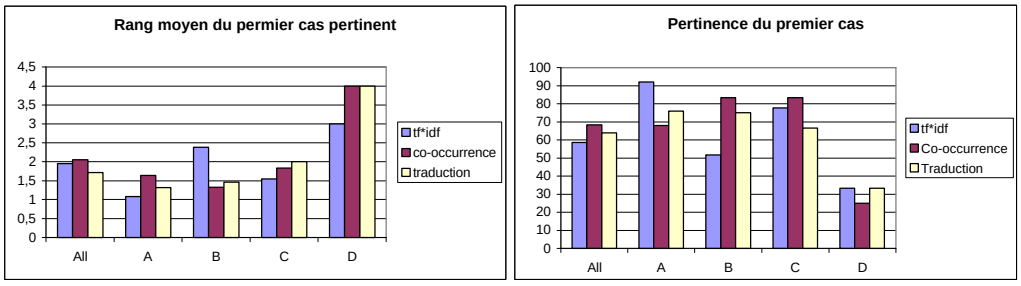


Figure 22 : Comparaison des trois mesures de similarité

Nous observons une amélioration significative dans la sélection globale du premier voisin pertinent par rapport au *tf*idf* (voir Figure 22). Ceci est expliqué en partie par les résultats obtenus sur les groupes A et B. En reprenant notre exemple précédent,

nous observons que les nouvelles listes obtenues par ce modèle sont limitées à quelques mots, la plupart d'entre eux étant très pertinents (voir Tableau 11). Ce modèle introduit moins de bruit que l'approche par cooccurrence sans filtrage. Pour illustrer la qualité des résultats des deux premiers groupes, il est intéressant de noter la justesse de quelques-unes des associations obtenues par le modèle de traduction :

Mot de problème	Traduit des mots de solutions suivants
Release	*Release, call, that, conference*
Earnings	*Result, date, earnings, NULL, conference,next*
Report	*Do, quarter, date, give*

Mot de problème	Traduit des mots de solutions suivants
Distribution	*List*
Dial	*PHONE_NUMBER, participate*
Conference	*Conference, release, usually, dial*

Tableau 11 : Exemples de listes de traduction

On s'attend à ce que des améliorations soient obtenues suite à l'entraînement du modèle sur une plus grande base de cas. Néanmoins, le modèle présenté offre un bon compromis entre les résultats obtenus par l'approche de similarité par *tf*idf* et un usage adéquat des solutions. Son principal désavantage est qu'il est difficile de combiner les valeurs de probabilité avec la similarité des problèmes.

3.6 Autres approches possibles

Il existe d'autres approches qui nécessitent peu d'intervention humaine et qui mériteraient d'être considérées lors des travaux ultérieurs. Nous passons en revue quelques-unes de ces approches.

3.6.1 Expansion des problèmes (requêtes) par cooccurrences

Est-ce que l'extension des descriptions de problèmes par l'ajout de termes apporterait les mêmes améliorations de performance que l'expansion de solution ? Cette approche fréquemment préconisée en recherche d'information tente de faire le pont entre

des requêtes contenant peu de termes et des documents de bien plus grande taille. Cette problématique diffère de la nôtre car nos descriptions de problèmes et de solutions sont de longueur comparable et l'uniformité des solutions est plus grande.

Pour obtenir une appréciation de cette approche, nous avons mené une expérimentation avec le schéma suivant pour obtenir des cooccurrences entre mots de requêtes :

- le comptage dans une matrice des termes en cooccurrence dans les requêtes dont la distance est moindre qu'une certaine taille de fenêtre de mots ;
- la comparaison des vecteurs de mots par une métrique de similarité. L'intuition est que le comportement des termes est décrit par les termes avoisinants et qu'un voisinage semblable indique une similarité des termes. Nous avons essayé différentes métriques (cosinus, Tanimoto, Dice, Overlap…) pour comparer les vecteurs ;
- la création des listes de cooccurrences, avec un filtrage possible sur la valeur de similarité et le nombre de cooccurrences.

Les meilleurs résultats obtenus en appliquant ce schéma avec une métrique de Tanimoto[13] sont les suivants :

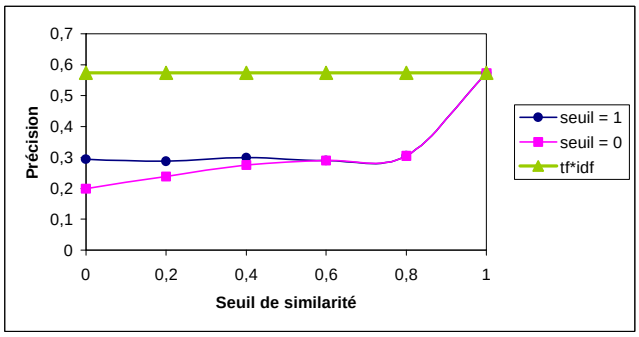


Figure 23 : Résultats obtenus avec des cooccurrences de problèmes

[13] La métrique de Tanimoto s'exprime comme suit $\frac{|X \cap Y|}{|X \cup Y|}$.

On constate que l'ajout de cooccurrences aux descriptions de problèmes dégrade la précision du système (les courbes rose et bleu de la figure 23) par rapport à celle obtenue par l'approche de *tf***idf* (la droite verte). En variant le seuil de similarité, on observe que la performance demeure inférieure au *tf***idf* jusqu'à ce que l'on réduise significativement le nombre de termes ajoutés aux requêtes, c.-à-d. lorsque le seuil de similarité entre termes tend vers 1.0. De plus amples expérimentations seraient nécessaires pour approfondir cette voie mais ce résultat rejoint les conclusions obtenues en recherche d'information indiquant que les approches naïves d'expansion de requêtes apportent peu de bénéfices à la performance des systèmes de recherche (Voorhees 1994).

3.6.2 Filtrage des cooccurrences

Nous avons soulevé le problème que les listes de cooccurrences contiennent plusieurs associations qui sont difficiles à expliquer par rapport au domaine et qui ne semblent pas contribuer à la phase de recherche. Nous avons tenté de remédier à ce problème par un filtrage basé sur la fréquence et la valeur d'information mutuelle. On pourrait toutefois envisager une étape supplémentaire dans le processus de génération où les listes sont ajustées par l'une des techniques suivantes :

- élagage manuel : le concepteur du système (ou un expert du domaine) passe en revue les listes proposées par le système et élague les termes jugés inadéquats. La difficulté réside dans la nature des associations jugées utiles pour l'application. Relèvent-elles de la sémantique du domaine applicatif ou de phénomènes linguistiques qui dépendent des descriptions textuelles ?

- rétroaction sur la pertinence : à partir de listes de cas jugés pertinents par un évaluateur externe, il est possible de modifier l'importance des cooccurrences dans les listes. Ces modifications consistent à augmenter ou diminuer le poids d'un terme en fonction de sa présence (ou son absence) dans les cas pertinents. Des techniques de type *relevance feedback* ou *credit assignment* permettraient de mettre en oeuvre cette approche. La technique de Rocchio (*pseudo*

relevance feedback) permet également d'ajuster ces poids sans jugement de pertinence.

- *clustering* de termes : on pourrait envisager des construire les listes de cooccurrences non pas à partir des termes individuels mais plutôt à partir de regroupement de mots. L'avantage de cette technique est que l'on peut attribuer une signification aux groupes de mots et ainsi mieux capturer la nature des associations entre les problèmes et les solutions.

Cette voie peut difficilement être envisagée pour les modèles de traduction puisqu'elle entraîne une perte de la masse de probabilité.

3.6.3 Ressources linguistiques pour similarité sémantique

L'insertion de relations lexicales dans le cycle CBR a été étudiée par (Burke *et al*. 1997) et (Brüninghaus and Ashley 1999) via l'utilisation d'un thesaurus. Pour nos travaux, nous n'avons pas retenu cette approche après avoir constaté que les ressources disponibles ne fournissaient pas une bonne couverture de notre vocabulaire. L'identification de relations sémantiques dans les listes de cooccurrences à partir de ces ressources linguistiques présente un intérêt. Nos résultats indiquent que peu de relations de synonymie et d'hypernymie que l'on retrouve dans un thesaurus sont présentes dans les listes d'associations que nous avons obtenues. Ainsi nous sommes portés à croire que ce type de connaissance serait peu utile dans la description des relations entre des descriptions de problème et de solutions. Toutefois, d'autres ressources telles que des glossaires du domaine du service aux investisseurs pourraient combler une partie de ces lacunes.

3.6.4 Représentation structurée des cas

Tel que décrit au chapitre 2, le concepteur d'un système CBR textuel se voit confronté au dilemme suivant : étendre l'approche de recherche ou favoriser une structuration plus élaborée des cas. Nous avons retenu la première option puisque qu'elle

permet une réutilisation efficace des messages antécédents lors de la construction de la base de cas. De plus, comme les problèmes cibles soumis au processus de recherche doivent être structurés au même niveau que les cas de la base, cette approche exige une intervention de l'usager qui peut être coûteuse et difficile à mettre en oeuvre.

Toutefois, d'autres approches de représentation des cas peuvent être envisagées. Une structuration par catégorie sémantique (Bruninghaus & Ashley 1997) pourrait être ajoutée à la représentation lexicale préconisée dans nos travaux. Des travaux en classification automatiques menés par (Dubois 2002) sur notre corpus de relations aux investisseurs révèlent que la construction de classificateurs thématiques se révèle une tâche ardue en raison du déséquilibre du notre corpus (une répartition non-uniforme des messages sur les différents thèmes). Une représentation structurée des intentions présentes dans les messages pourrait également favoriser une meilleure comparaison des requêtes et rendre superflu un transfert lexical vers les réponses. Des travaux sont actuellement en cours (Bélanger 2003) pour tenter de capturer les intentions des requêtes de notre corpus à partir de patrons syntaxiques et sémantiques. Finalement des progrès au niveau des approches d'extraction d'information sur des textes non structurés pourraient créer une percée significative pour la construction de bases de cas. Mais il est important de rappeler qu'aucune de ces techniques ne permettrait de traiter des cas portant sur des problèmes n'ayant pas été considérés lors de la conception du système. Ce qui représente un désavantage pour les domaines dont les changements dynamiques sont fréquents.

3.7 Conclusion

Nos expérimentations sur les deux modèles et leur comparaison avec une approche de type *tf***idf* indiquent que l'insertion de relations lexicales dans la phase de recherche apporte des améliorations significatives en terme de précision et d'ordonnancement des cas. Cette approche est bien adaptée aux propriétés de notre base de cas, c.-à-d. la forte homogénéité des descriptions de solutions. Toutefois, elle présente des lacunes dont le bruit lexical généré par les listes de cooccurrences. Nous avons proposé d'autres voies pour poursuivre ces travaux. Une étude comparative de ces

approches sur des textes présentant différentes caractéristiques (longueur, dépendance au domaine, homogénéité des descriptions) permettraient de mieux cerner les limitations et les avantages de chacune d'entre elles.

Dans ce chapitre, nous avons estimé l'efficacité de nos deux approches de recherche de cas à partir d'une évaluation externe de la pertinence des messages. Ces évaluations de pertinence reposent sur le jugement d'un humain et sont utilisées lorsque la conception du module CBR est complétée. Au chapitre suivant, nous étudions une démarche pour déterminer, lors de la construction du module CBR, comment structurer notre base de cas et sélectionner une approches de recherche.

Chapitre 4 . Démarche de construction du module CBR

L'utilisation des réponses passées dans la phase de recherche du module CBR peut apporter des gains en termes de précision et d'ordonnancement de cas. Ces gains dépendent des paramètres sélectionnés lors de la construction[14] de la base de cas et des métriques de similarité utilisées dans la phase de recherche. Dans ce chapitre, nous proposons une démarche méthodologique qui s'appuie sur des propriétés de la base de cas pour guider le processus de construction. Nous présentons des indicateurs de qualité et évaluons l'impact des paramètres du système CBR sur les valeurs des indicateurs. Pour certaines étapes du processus de construction, nous tentons de déterminer si ces métriques corroborent les évaluations obtenues pour le processus de recherche, et nous proposons des critères qui permettent de déterminer si l'utilisation de solutions passées dans la phase de recherche, au lieu des problèmes passés, est avantageuse.

La conception d'un système CBR repose principalement sur la structuration de la base de cas et sur les stratégies de recherche développées pour cette base. L'étude de la phase de recherche CBR au chapitre précédent nous a permis de déterminer que l'utilisation des solutions dans le processus de recherche est parfois avantageuse pour repérer les cas pertinents. Les techniques utilisées apportent des améliorations en termes de précision et rang moyen, mais leur performance dépend des caractéristiques du corpus. Lors de nos travaux, nous avons observé que nos choix de structuration de cas (par exemple, la conversion en poids *tf*idf*) et de stratégies de recherche (par exemple, la sélection d'un seuil d'information mutuelle) avaient un impact significatif sur les performances du module CBR. L'adoption d'une bonne méthodologie de construction du système permettrait de guider ces choix et d'anticiper les performances attendues du système opérationnel. De plus, devant la multitude de stratégies de recherche possibles, il serait intéressant de déterminer a priori la stratégie de recherche la plus appropriée en fonction de la structuration choisie pour concevoir une base de cas. Dans ce chapitre, nous présentons quelques résultats de nature exploratoire portant sur ces questions.

---

[14] Nous utilisons le terme *construction* au sens du terme anglais *authoring*, qui désigne en CBR la création des connaissances (*knowledge containers*) d'un système. Notre étude porte sur les choix reliés au vocabulaire, à la base de cas et aux métriques de similarités.

Une méthodologie de construction doit miser sur la quantification des facteurs qui influencent le comportement du système. Dans la littérature CBR, les indicateurs de performance sont surtout utilisés pour la maintenance des bases de cas. Néanmoins, nous proposons d'utiliser certains de ces indicateurs pour la construction de la base de cas afin de choisir les paramètres et la stratégie de recherche les plus appropriés. Ces mesures nous permettent ainsi de comparer les performances a priori des mécanismes de recherche.

En premier lieu, nous décrivons la démarche suivie pour construire notre module de raisonnement à base de cas (section 4.1). Par la suite, nous passons en revue un certain nombre des mesures utilisées en raisonnement à base de cas (section 4.2). Nous proposons un cadre de travail pour guider le processus de construction de la base de cas basé sur des indicateurs que nous avons sélectionnés et développés (section 4.3). Le but ultime d'un tel cadre de travail est d'aider le concepteur d'un système à choisir l'une des approches étudiées au chapitre précédent. Finalement, nous vérifions si les propriétés de la base de cas permettent d'anticiper les résultats qui seront obtenus par le système CBR textuel opérationnel. Nous tentons de déterminer si la base de cas est adéquatement structurée, et si cette structuration favorise l'une ou l'autre des stratégies de recherche. Nous présentons les résultats expérimentaux obtenus pour notre base de cas (section 4.4) et nous discutons d'aspects connexes à ces travaux (section 4.5).

4.1 Construction des connaissances du module

La conception des connaissances du module CBR comporte les étapes illustrées à la figure 24. Afin de supporter ce processus, nous explorons les utilisations possibles d'indicateurs de performance pour adresser les points suivants :

- L'estimation de l'impact de la structuration des cas sur les caractéristiques de la base de cas ; plus particulièrement la sélection des termes du vocabulaire, la normalisation des vecteurs de termes et la modification des poids ;

- l'évaluation de notre hypothèse voulant que les cas dont les solutions sont plus homogènes ont avantage à utiliser des associations de termes ; et

- la sélection de paramètres de recherche qui permettent une exploitation judicieuse de la base de cas ;

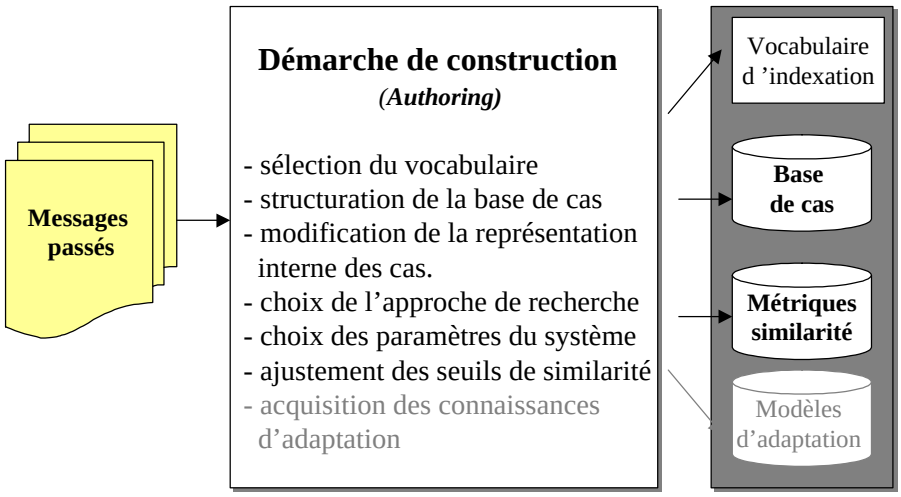


Figure 24 : Démarche de construction de la base de cas

Contrairement aux travaux en maintenance de base de cas, notre objectif dans l'utilisation des indicateurs n'est pas de sélectionner les cas qui mènent à une base de cas compacte et minimale, mais plutôt de choisir la structuration de la base de cas, la stratégie de recherche et les paramètres du système qui rehausseront la qualité des solutions.

Les indicateurs de performance peuvent être exploités à chacune des étapes du processus de construction d'une base de cas, et plus précisément aux étapes suivantes :

a) Sélection du vocabulaire : (Lenz *et al.* 1998) émet l'hypothèse que l'utilité de chacun des termes du vocabulaire peut être déterminée en fonction de sa fréquence dans le corpus et de sa catégorie lexicale. Ainsi, le choix des catégories ou du seuil de fréquence auront un impact sur le reste du processus de construction. Nous vérifions cette hypothèse pour notre base de cas. Toutefois, comme il est difficile d'évaluer l'impact de ces choix lors de la construction du vocabulaire, l'évaluation est reportée aux étapes ultérieures de la construction.

b) Construction des cas : les cas, pour alimenter la phase de recherche, contiennent des représentations vectorielles et séquentielles de termes. Notre évaluation porte principalement sur la représentation vectorielle et l'impact, sur la phase de recherche, de sa normalisation et de sa conversion en poids *tf*idf*.

c) Choix de la stratégie de recherche : au chapitre précédent, nous avons émis l'hypothèse que les stratégies de recherche utilisant les solutions devraient être préconisées lorsque les solutions sont plus uniformes que les problèmes. Nous validons cette hypothèse par les tests suivants :

   - on vérifie que la densité des solutions est supérieure à celle des problèmes (uniformité).

   - on vérifie que les voisinages des problèmes et des solutions se recoupent plus fréquemment lorsque nous utilisons des méthodes de recherche par association.

d) Construction des connaissances de similarité pour les modules de recherche : les connaissances pour les stratégies de recherche que nous préconisons sont les listes d'associations. Dans le cas de l'approche par alignement de traduction, aucun paramètre particulier n'est nécessaire (hormis pour le lissage). On peut donc construire cette connaissance sans avoir recours aux indicateurs proposés. Toutefois, pour l'approche par cooccurrence, il peut être avantageux de tronquer des listes en fonction des valeurs d'information mutuelle afin d'améliorer les performances du système. Par ailleurs, on déterminera ce seuil d'une façon expérimentale en comparant les indicateurs de voisinage définis dans les sections suivantes.

e) Choix des seuils de similarité et de rejet : pour certaines applications, il serait intéressant de fournir des seuils de similarité permettant d'estimer la pertinence des plus proches voisins. Idéalement, deux seuils (supérieur et inférieur) qui définissent trois régions devraient être sélectionnés :

- les cas nécessairement pertinents : la similarité des ces cas dépasse le seuil supérieur. Ce seuil détermine le niveau de confiance du système à sélectionner des cas en toute autonomie ;

- les cas non pertinents : la similarité de ces cas est en deçà du seuil inférieur ; ce seuil permet d'éliminer les recommandations inutiles ;

- les cas sont potentiellement pertinents : il s'agit de déterminer la pertinence des cas dont la similarité se situe entre les deux seuils et qui nécessitent une évaluation externe par l'usager.

4.2 Indicateurs proposés dans la littérature CBR

Les principales métriques que nous avons répertoriées dans la littérature (Watson 1997), (Reinartz *et al.* 2001), (Racine et Yang 1997), (Leake et Wilson 1999b), (Smyth et McKenna 1998) pour caractériser une base de cas sont présentées au tableau 12.

Indicateur	Description
Nombre de cas	le nombre de cas est une indication de la performance du système, surtout lors de la phase de recherche. Toutefois un plus grand nombre de cas entraîne des temps de recherche plus longs, sans garantir qu'un plus grand nombre de problèmes seraient adéquatement résolus.
Densité des cas	cette mesure reflète la proximité moyenne des cas. La densité inter-cas permet d'estimer la compétence de la base de cas à résoudre des problèmes. Plus la base est dense, plus la concentration de cas dans une région est grande et moindre est la contribution de chacun des cas individuels pour résoudre des problèmes dans cette région. Il est recommandé d'utiliser une base dont la densité est plutôt faible, ayant une distribution plus uniforme des cas. Cette mesure peut être évaluée à partir de la similarité entre les cas, et peut exiger l'utilisation du module de recherche du système CBR. Cette mesure peut être utilisée dans un cadre textuel.
Distribution des cas	Cette distribution résume la répartition des cas sur l'ensemble des problèmes possibles. Une distribution plus uniforme des cas favorise la résolution de problèmes variés. Une distribution irrégulière indique que certains problèmes sont susceptibles de ne pas être résolus. Il est possible de définir cette mesure lorsque les champs des différents attributs sont bien déterminés. On peut alors identifier les "trous" dans la base. Toutefois, cet indicateur est difficile à définir dans un cadre textuel.
Rectitude (*correctness*)	un cas est correct si sa solution permet de résoudre le problème qui lui est associé. Habituellement, cette propriété est supposée vraie dans la plupart des systèmes. Nous présumons que ceci est le cas pour notre application de réponse

	au courrier électronique.
Unicité et redondance	un cas est unique si aucun autre cas de la base ne contient simultanément la même description de problème et de solution. La redondance entre deux cas peut être définie soit comme a) la correspondance exacte entre les cas, b) l'un des cas étant un sous-ensemble de l'autre, ou c) une très grande similarité entre les cas. La notion de redondance est particulièrement utile en maintenance de base de cas où l'on cherche à élaguer les cas qui contribuent peu aux performance du système.
Minimalisme	un cas est minimal si tout autre cas ayant la même solution n'est pas un sous-ensemble de celui-ci. La relation inverse est celle de *subsumption* (le recouvrement d'un cas par un autre). Idéalement, on souhaiterait conserver dans notre base les cas minimaux. Toutefois pour notre application CBR textuel, il est intéressant de conserver dans la base différentes formes narratives d'une même solution, offrant ainsi à l'usager une plus grande variété de réponses possibles.
Consistance	des cas ayant des mêmes descriptions de problème ne devraient pas avoir de solution différente. On veut ainsi éviter que de petites variations dans les descriptions de problèmes entraînent des changements majeurs dans les solutions. On retrouve des définitions dans la littérature basées sur a) le taux de recouvrement entre les descriptions et b) des règles définissant des combinaisons de valeurs ou d'attributs jugés conflictuels par le concepteur du système. En CBR textuel, des variantes similaires de problèmes peuvent entraînent des solutions différentes. Par exemple, la négation d'une proposition peut mener à une réponse totalement différente.
Cohérence	des cas sont cohérents si, pour une même solution, la description des problèmes varie peu. La notion de cohérence est ambiguë car on s'attend à ce qu'une solution soit utilisable dans plusieurs contextes et que les problèmes correspondants soient diversifiés. Cet indicateur est difficile à mettre en pratique.
Couverture et atteignabilité	ces mesures, définies pour des systèmes CBR dotés de capacité d'adaptation, désignent l'étendue des problèmes qui peuvent être résolus par un système. On peut les définir à partir de l'ensemble des cas similaires permettant de résoudre un problème cible. Ces mesures combinent la capacité de déterminer les plus proches voisins d'un cas cible et d'évaluer si ces voisins peuvent être modifiés pour reconstruire la solution cible. En raison de l'absence de formalisme d'adaptation en CBR textuel, il n'existe pas de définition pour ces indicateurs.
Régularité	cette mesure a été proposée pour décrire la relation entre les descriptions de problèmes et de solutions afin de s'assurer que des problèmes similaires ont des solutions similaires. Cette mesure est analogue à la consistance. Elle est toutefois définie non pas par le taux de recouvrement entre les cas, mais plutôt par les propriétés du voisinage des cas. Des mesures proposées à la section suivante reprennent cette idée de caractériser les cas par leur voisinage.

Tableau 12 : Indicateurs de performance dans la littérature CBR

Du point de vue de leur applicabilité au CBR textuel, la définition de ces mesures présente quelques lacunes. Plusieurs de ces indicateurs requièrent que les cas soient structurels, qu'ils soient homogènes (définis à partir d'un même nombre limité

d’attributs) et qu'ils possèdent pour la plupart des valeurs pour chacun des attributs. Hors, une base de cas textuels est hétérogène par définition, la représentation interne des cas contient peu de termes (par rapport au vocabulaire du système), et le taux de recouvrement entre les différents cas est faible en raison des descriptions peu répétitives.

L’utilité des indicateurs tel que le minimalisme ou la redondance est restreinte. Un cas est redondant s’il est défini à partir des mêmes attributs qu'un autre et il est minimal si aucun autre cas n’est défini à partir d’un sous-ensemble de ses attributs. Les cas textuels comptent plusieurs centaines d’attributs et peu d’entre eux sont structurés à partir d’une même combinaison d’attributs. Ainsi, d’un point de vue lexical, la notion de redondance entre cas textuels n’existe pas vraiment et il est peu fréquent que les problèmes et les solutions aient exactement la même formulation.

Finalement, les mesures reliées à l’adaptation de cas (couverture et atteignabilité) sont impraticables dans l’état actuel du domaine. Comme on ne retrouve pas de travaux sur l’adaptation en CBR textuel, ces mesures peuvent difficilement être définies pour des cas comportant des textes.

4.3 Indicateurs pour la construction de notre base de cas

Dans cette section, nous décrivons les indicateurs que nous avons retenus pour mener notre démarche de construction et les aspects utilisés pour définir ces indicateurs.

4.3.1 Aspects à mesurer par les indicateurs

La construction d’un module CBR repose sur les cas initiaux mis à la disposition du concepteur. Or la définition d’indicateurs pour guider ce processus de construction dépend des aspects que l’on désire mesurer sur notre base de cas. Pour notre application, nous proposons d’évaluer les aspects présentés au Tableau 13.

D’un point de vue CBR textuel, ces aspects offrent trois niveaux d’analyse : la comparaison des textes avant structuration, la similarité des textes tels que structurés dans la base de cas et l’évaluation de l’association entre les problèmes et solution d’un cas en

fonction des autres cas de la base. Nous définissons à la section suivante des indicateurs qui exploitent ces trois aspects.

Aspect à mesurer	Description	Motivation
La présence d'attributs dans les cas	Comme un cas textuel contient peu de termes, la présence de certains mots donne une indication du domaine couvert par ce cas. On peut comparer la présence de mots entre deux cas par des mesures binaires telles que celles de Dice, d'Overlap ou de Tanimoto (Manning & Schütze 1999).	On vise à déterminer si les descriptions initiales de cas abordent les mêmes thèmes, et ce indépendamment du mécanisme de similarité utilisé par le système ou de l'importance relative attribuée à chacun des attributs.
La similarité relative entre les cas	Cet aspect donne une indication de la proximité des cas après la structuration de la base de cas. Lorsque calculé au niveau de la base de cas, on obtient la répartition des cas dans l'espace de résolution de problème. Ces indicateurs prennent en compte les valeurs et les poids de chacun des attributs. Ils peuvent être estimés à partir des valeurs de similarité calculées par le module CBR.	Par cet aspect, on vise à évaluer la qualité de la structuration des cas, i.e. l'identification des attributs pertinents d'un cas et l'ajustement de leur contribution aux calculs de similarité du système.
Le voisinage des cas	Il s'agit d'estimer quelle portion de la base de cas partage une relation avec un cas cible. Les relations sont définies à partir de la similarité des problèmes et des solutions. Par exemple, on peut définir la redondance comme la proportion de cas ayant conjointement des problèmes et des solutions similaires à un cas cible. La zone de voisinage d'un cas est établie à partir de seuils sur la similarité des problèmes et des solutions.	Nous voulons évaluer l'aptitude du mécanisme de recherche à repérer une ou plusieurs solutions pertinentes. Pour notre application, le repérage de cas pertinents repose sur le calcul des similarités entre les problèmes/solutions et sur la capacité d'associer un problème à une solution.

Tableau 13 : Aspects à mesurer pour la construction d'un module CBR textuel

4.3.2 Les indicateurs retenus pour notre étude

Afin de mener notre étude sur la construction de la base de cas, nous avons retenu des indicateurs pour appuyer les tâches de structuration des cas (c.-à-d. déterminer les termes du vocabulaire et leur importance dans la description des cas), de sélection d'une stratégie de recherche et de sélection des paramètres pour optimiser la performance de cette stratégie. Plusieurs des mesures présentées à la section 4.2 ne s'appliquent pas ou contribuent peu à ces tâches. Par exemple, nous construisons le module CBR à partir d'un nombre fixe de cas dont on présume la rectitude. Tel que mentionné précédemment, nous n'avons pas de formalisme pour déterminer dans un cadre textuel les indicateurs de distribution de cas, de couverture et d'atteignabilité. Les indicateurs d'inconsistance et de minimalisme servent habituellement à guider l'élagage de cas dans la base, une tâche que nous ne considérons pas à ce stade-ci dans nos travaux. . Et bien que l'indicateur de cohérence soit intéressant d'un point de vue théorique (il indique la diversité des problèmes associés à une solution), cette information semble avoir peu d'incidence sur l'efficacité du raisonnement. Nous avons donc retenu 3 mesures pour mener nos travaux :

- Le taux de recouvrement : cette mesure rejoint la notion de redondance et est définie à partir de la présence d'attributs dans les cas.

- La densité de cas : nous utilisons la définition proposée par (Smyth et McKenna 1998), qui s'appuie sur la similarité relative entre les cas

- La cohésion d'un cas : tout comme l'indicateur de régularité, la cohésion est définie à partir du voisinage d'un cas. Toutefois la définition et l'exploitation de cet indicateur vise à exploiter la relation entre les problèmes et les solutions d'un cas.

Nous définissons ces indicateurs dans les paragraphes suivants.

a) Les taux de recouvrement d'un cas - *recouvrement(C,CB)* :

Cet indicateur de la présence d'attributs nous donne un aperçu de la distribution des termes dans les cas et de l'homogénéité des descriptions. On peut l'exprimer à l'aide du nombre moyen d'attributs communs à deux cas. En utilisant la métrique d'*overlap*, on obtient la formulation suivante :

$$recouvrement_{desc}(C,CB) = \sum_{C' \in CB} \frac{overlap_{desc}(C,C')}{(|CB|-1)}$$

où

$$overlap_{desc}(C,C') = \frac{\left|A_C^{desc} \cap A_{C'}^{desc}\right|}{\min\left(\left|A_C^{desc}\right|,\left|A_{C'}^{desc}\right|\right)}$$

et A_C^{desc} sont les attributs de la description soit du problème (A_C^{prob}) ou de la solution (A_C^{sol}) du cas C.

Nous utilisons cette mesure pour estimer le degré d'uniformité des descriptions de cas et nous comparons les valeurs obtenues pour les problèmes et les solutions. Un faible niveau de recouvrement indique l'unicité d'une description de cas, tandis qu'un fort degré de recouvrement indique qu'il y a une redondance avec certains cas de la base.

b) La densité des cas - *densité(C, CB)* :

Nous reprenons cet indicateur de similarité tel que défini dans la littérature (Smyth & McKenna 1998). La densité peut-être simplement mesurée par la moyenne des similarités entre les cas :

$$Densité_{probl}(CB) = \frac{\sum_c \sum_{c'} sim_{prob}(c,c')}{|CB| \times (|CB|-1)} \quad \text{et} \quad Densité_{sol}(CB) = \frac{\sum_c \sum_{c'} sim_{sol}(c,c')}{|CB| \times (|CB|-1)}$$

Tout comme le taux de recouvrement, nous utilisons cet indicateur pour estimer la distribution des termes dans la base de cas et l'uniformité relative des problèmes et des solutions. Par ailleurs, la densité permet de mesurer l'impact de l'uniformité des descriptions sur le mécanisme de recherche.

c) La cohésion d'un cas - *cohésion(C,CB)* :

L'indicateur de cohésion que nous proposons permet de mesurer l'étroitesse de la relation entre les problèmes et les solutions. Pour élaborer cet indicateur, nous nous appuyons sur l'idée qu'un cas présente une forte cohésion si, dans un voisinage donné, d'autres cas présentent des relations problème-solution similaires.

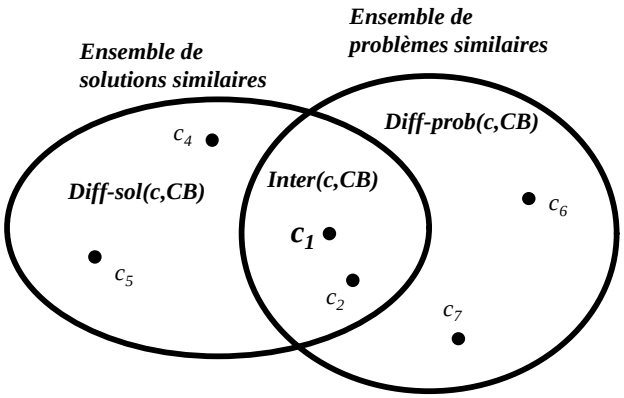


Figure 25 : Ensembles de cas ayant des problèmes et/ou des solutions similaires

Pour cerner plus précisément cette intuition, considérons deux ensembles liés à un cas c_1 : l'ensemble des cas ayant des problèmes similaires ($E_{problème}$) et l'ensemble des cas ayant des solutions similaires ($E_{solution}$) (Figure 25). Ces ensembles sont définis à l'aide des seuils de similarité $\ddot{a}_{prob}$ et $\ddot{a}_{sol}$ comme suit :

$$E_{problème}(c_1,CB) = \{ c \in CB:\ sim_{prob}(c_1^{prob}, c^{prob}) > \ddot{a}_{prob} \}$$

$$E_{solution}(c_1,CB) = \{ c \in CB:\ sim_{sol}(c_1^{sol}, c^{sol}) > \ddot{a}_{sol} \}$$

où sim_{prob} *et* sim_{sol} sont respectivement les similarités des problèmes et des solutions de deux cas. Dans notre cadre textuel, les mesures de similarité sont celles que nous avons étudiées au chapitre 3.

A partir de ces ensembles, on détermine trois groupes dont les cas ont :

- des solutions et des problèmes similaires à c_1

$$Inter(c_1,CB) = E_{problème}(c_1,CB) \cap E_{solution}(c_1,CB)$$

- uniquement des problèmes similaires à c_1

$$Diff\_prob(c_1,CB) = E_{problème}(c_1,CB) - Inter(c_1,CB)$$

- uniquement des solutions similaires à c_1

$$Diff\_sol(c_1,CB) = E_{solution}(c_1,CB) - Inter\ (c_1,CB)$$

Ces trois ensembles permettent de mesurer la qualité de la relation entre les problèmes et les solutions. L'union correspond au nombre de cas contenus dans ces trois ensembles. Le degré de cohésion peut donc être défini comme suit :

$$\text{degré\_cohésion}(c_1) = Inter(c_1,CB) / Union(c_1,CB)$$

Un cas est cohésif s'il se comporte comme son voisinage. Cette mesure indique si la mise en correspondance de la solution et du problème est conforme aux autres cas similaires de la base. Nous proposons cette définition pour estimer si le problème et la solution d'un cas semblent bien alignés ou plutôt dissonants. Un cas dont le degré de cohésion est faible est soit unique, soit peu redondant, soit inconsistant ou soit incohérent. Nous proposons à la section 4.5.1 des définitions de ces quatre propriétés à partir des ensembles de voisinage de cas.

4.4 Résultats de l'évaluation

Nous présentons dans cette section les résultats obtenus lors de la construction de notre système. Nous avons évalué les indicateurs à l'aide de notre base de cas et quantifié l'impact des différents points identifiés à la section 4.1 pour mener notre démarche.

4.4.1 Filtrage du vocabulaire par les catégories lexicales

Tel que suggéré par (Lenz *et al.* 1998), l'utilité des termes constituant le vocabulaire du système peut être définie en fonction de leur fréquence et de leur catégorie lexicale. Quelle est la contribution de chacun ? Dans cette section, nous évaluons des bases de cas construites à partir de termes provenant de différents groupes de catégories lexicales. Pour cette étape, nous utilisons les indicateurs de recouvrement et de densité. Les premiers résultats, présentés au tableau 14, le sont pour une base de cas dont la représentation repose uniquement sur les noms du corpus. Par la suite, nous ajoutons successivement d'autres catégories lexicales dont nous déterminons la contribution cumulative. L'ordre a été choisi selon notre perception de l'utilité de chacune de ces catégories. Le recouvrement global est la moyenne du recouvrement des problèmes et des solutions.

Catégories Lexicales	Taille du vocabulaire	Recouvrement global	Densité	
			Problème	solution
Noms	215	0,377	0,079	0,102
+ verbes	306	0,407	0,080	0,099
+ adjectifs	354	0,388	0,074	0,095
+ quantités	380	0,386	0,073	0,097
+ adverbes	408	0,379	0,072	0,095
+ pronoms	420	0,409	0,078	0,100
+ prépositions	444	0,426	0,082	0,106

Tableau 14 : Évaluation selon les catégories lexicales

On remarque que les variations du recouvrement sont étroitement liées à la densité des cas, et que la densité des solutions est nettement supérieure à celle des problèmes, ce qui suggère une uniformité plus grande des solutions. Nous observons le même phénomène pour les résultats de recouvrement présentés à la section suivante.

On note de légères variations des résultats selon la catégorie lexicale. On remarque que le poids de la représentation repose principalement sur les noms et que les autres catégories apportent peu de contribution. Une évaluation menée sur chacune des catégories lexicales individuelles indique que celles-ci offrent des valeurs nettement inférieures aux noms.

L'ajout de mots grammaticaux tels que les pronoms et prépositions amènent une augmentation de la densité de cas. Ceci est probablement dû au faible nombre de mots appartenant à ces catégories et à leur occurrence fréquente dans les descriptions de cas. L'ajout de catégories lexicales contenant un nombre de mots nettement plus grand, comme les adverbes ou les adjectifs, entraîne une diminution des valeurs de nos deux indicateurs.

Afin d'établir un lien entre ces indicateurs et les performances de recherche du système, nous présentons aux figures 26 et 27 les résultats obtenus pour une recherche de type *tf*idf*.

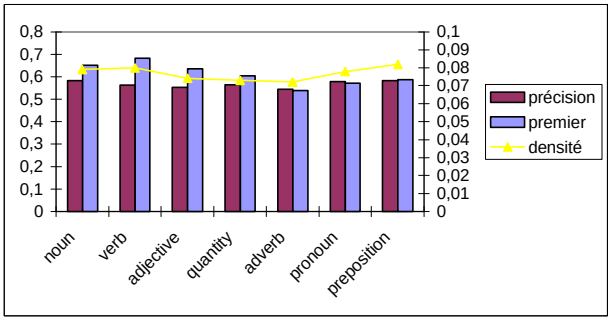


Figure 26 : Précision et densité des problèmes

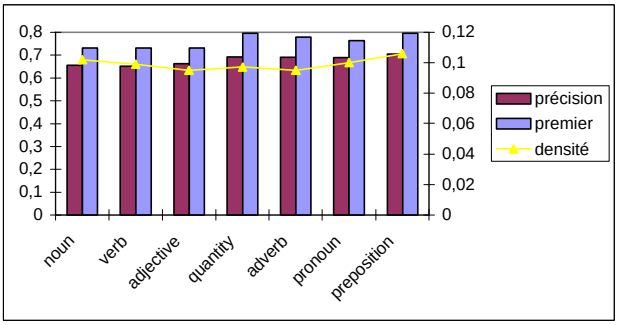


Figure 27 : Précision et densité des solutions

On note quelques variations sur les performances de recherche. Pour les problèmes, la courbe de pertinence du plus proche voisin (la courbe "premier") suit la courbe de densité. Les verbes contribuent légèrement à la performance du système tandis que les autres catégories apportent soit peu d'amélioration soit une dégradation de la précision. Pour les solutions, on remarque que les quantités jouent un rôle relativement

important. Ceci s'explique par le fait que les réponses contiennent des descriptions spécifiques portant sur des aspects quantitatifs (par exemple, dans les discussions sur des facteurs financiers). Les autres catégories apportent peu de contribution à la précision du système.

On peut donc affirmer que, hormis les noms, le choix des catégories lexicales a un impact limité sur la performance du système. Pour des considérations de modélisation des associations entre les mots de problèmes et les mots de solutions, nous avons quand même décidé des garder ces sept catégories lexicales dans nos expérimentations.

4.4.2 Filtrage du vocabulaire par un seuil de fréquence de mots

Dans cette section, nous évaluons si, tel que proposé par Lenz *et al.*, l'utilité des mots peut être déterminée par leur fréquence dans la base de cas. Avant de mesurer l'impact de la fréquence des mots, nous comparons le taux de recouvrement des problèmes et des solutions pour l'ensemble de la base de cas.

Lorsque tous les termes sont retenus, et ce indépendamment de leur fréquence, le taux moyen de recouvrement est de 25,7% pour les problèmes et de 36,9% pour les solutions. On rappelle que l'indicateur de recouvrement indique le nombre de termes en commun entre les cas et que pour cette mesure, la fréquence et le poids des termes ne sont pas pris en compte. Ainsi les problèmes ont en commun en moyenne 1 terme sur 4 et les solutions ont en moyenne plus de 1 terme sur 3 (Figure 28).

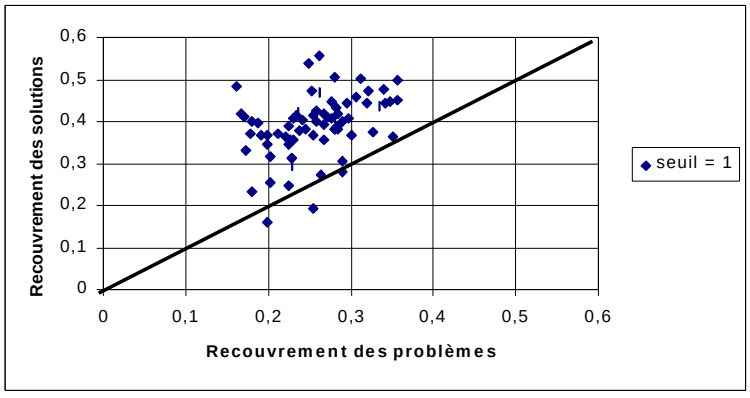


Figure 28 : Distribution du recouvrement des problèmes et des solutions

Il est intéressant de noter que la majorité des cas se situe au-dessus de la diagonale indiquant un recouvrement équivalent pour les problèmes et les solutions. Ainsi, dans la plupart des cas de notre base, les descriptions des solutions s'entrecoupent plus que les problèmes. Elles comportent plus de termes en commun et leur description est donc plus uniforme.

En rétrécissant le vocabulaire selon la fréquence des termes dans le corpus, on peut évaluer la variation du recouvrement des différents cas de la base (Figure 29).

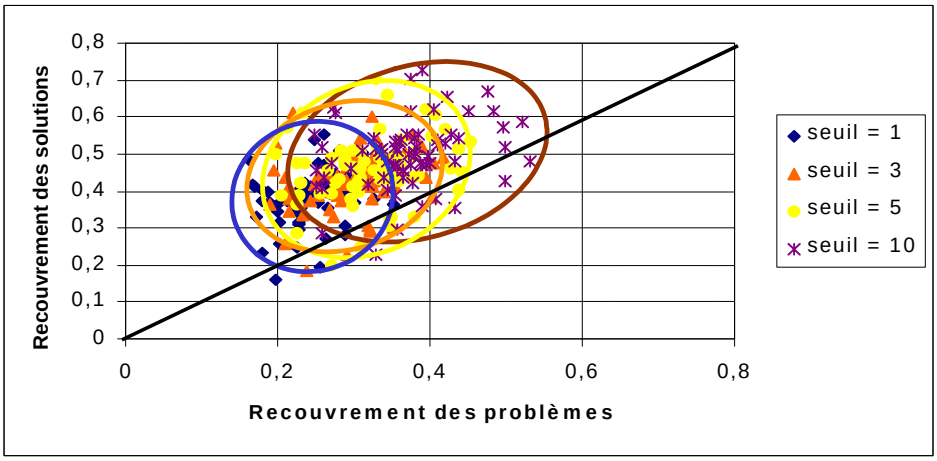


Figure 29 : Effet du filtrage en fréquence sur le recouvrement

Malgré la diminution du vocabulaire, la majorité des points demeure au-dessus de la diagonale, ce qui indique encore une fois une plus grande uniformité des solutions. On note une dispersion des cas (surface des ovales) ainsi qu'un élargissement du recouvrement des cas (déplacement des ovales vers le haut et la droite) au fur et à mesure que le seuil de fréquence augmente.

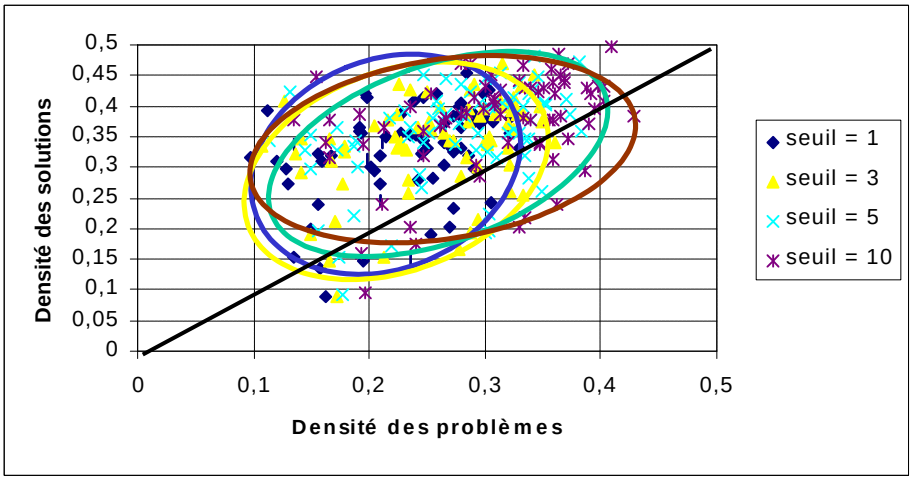


Figure 30 : Effet de la normalisation

Afin de relier ces résultats aux performances de recherche, nous comparons ces résultats aux valeurs de densités obtenues en tenant compte des poids des termes dans les cas. Nous utilisons pour la phase de recherche des représentations internes normalisées. La densité après normalisation des vecteurs de termes nous donne une valeur moyenne de 23,4% pour les problèmes et de 32,5% pour les solutions. On obtient la distribution de densité présentée à la Figure 30 (avec une mesure de cosinus normalisée) en fonction de la fréquence dans le corpus.

La dispersion de la densité est plutôt faible et suggère ainsi peu de changements dans la compétence de la base de cas. On peut donc anticiper que le filtrage en fréquence lorsque la représentation des cas est normalisée apportera peu de bénéfices à la performance de la phase de recherche. Son utilisation permettrait plutôt de faciliter l'interprétation des valeurs de similarité (situées entre 0 et 1).

4.4.3 Structuration *tf*idf* des poids de la base de cas

Une partie de l'évaluation de la structuration de la base de cas a été présentée par les résultats de la section précédente. Il nous reste toutefois à estimer l'effet de la conversion des poids des termes en valeurs *tf*idf*. Pour une évaluation à l'aide d'une mesure de cosinus, on obtient une densité de 7,9% pour les problèmes et de 10,3% pour

les solutions. Ces valeurs, inférieures à celles obtenues après normalisation, indiquent que le *tf*idf* dilate la base de cas et offre une plus grande répartition des cas dans l'espace de résolution des problèmes. Tel qu'observé dans nos expérimentations, cette réduction de densité se traduit par une amélioration de la performance de la recherche.

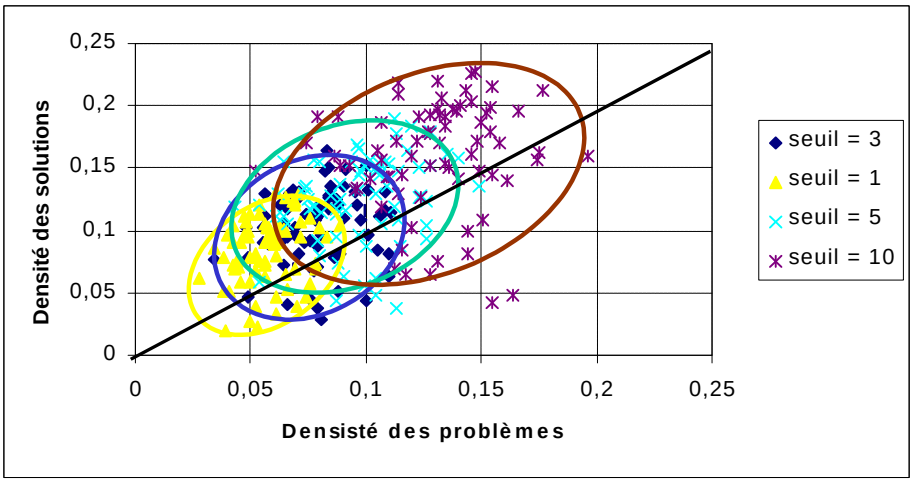


Figure 31 : Effet de la conversion en *tf*idf*

Afin de reprendre notre analyse du point précédent, nous présentons la distribution de la densité des cas en fonction du seuil des fréquences à la Figure 31.

La conversion des poids en *tf*idf* entraîne une progression des densités beaucoup plus importante lorsque le seuil de fréquence est augmenté. Ceci suggère que plusieurs des termes de faible fréquence ont des poids *tf*idf* relativement élevés.

4.4.4 Sélection d'une stratégie de recherche

Une stratégie qui permet à un cas de résoudre le plus de problèmes similaires peut être sélectionnée à l'aide d'un indicateur de voisinage des cas. Pour comparer les stratégies, nous avons effectué l'évaluation suivante : étant donné un ensemble de solutions cibles, quel est l'ensemble minimal de problèmes qui coïncident avec ces solutions ? Autrement dit, on cherche à déterminer les seuils de problèmes et de solutions qui résultent en un maximum de cohésion des cas. Nous avons donc suivi la procédure suivante :

a) pour un cas cible donné, on trouve un ensemble de solutions dans la base de cas dont la similarité est supérieure au seuil de solution ;

b) étant donné cet ensemble de solutions, on fait varier le seuil de similarité des problèmes pour identifier l'ensemble de problèmes qui permet un maximum de cohésion ;

c) en répétant l'expérience pour différents cas cibles et différents seuils de solution, nous obtenons des courbes de cohésion maximale.

d) nous comparons les courbes de cohésion maximum pour chacune des stratégies afin de voir si une tendance se dégage.

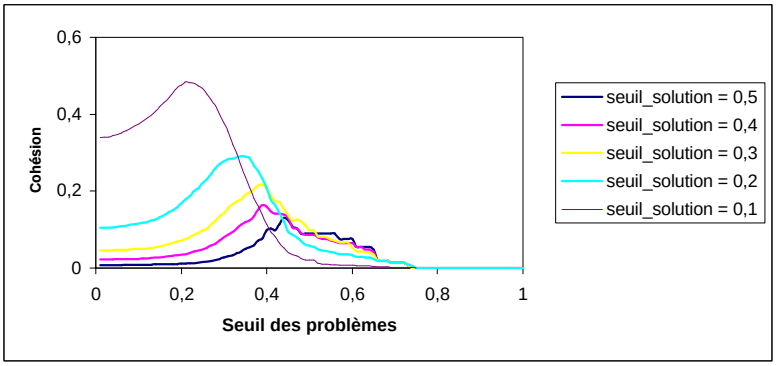


Figure 32 : Cohésion de l'approche de cooccurrences

En guise d'exemple, les courbes de cohésion obtenues pour différents seuils de solutions et de problèmes avec la stratégie de recherche avec cooccurrences (sans tronquer les listes d'associations) sont présentées à la Figure 32. À la figure 33, nous présentons les courbes de cohésion maximale obtenues pour les stratégies *tf*idf* et de cooccurrences. On note que des valeurs supérieures sont obtenues pour tous les seuils de solutions. Ce résultat supporte les conclusions obtenues au chapitre 3 à savoir que l'approche par cooccurrence se révèle un meilleur choix que l'approche *tf*idf* pour la base de cas de notre application de réponse au courrier électronique

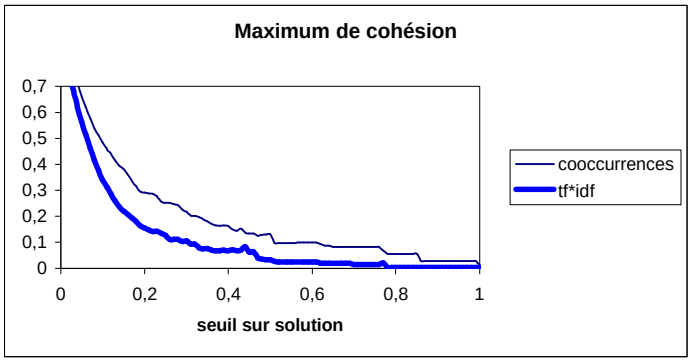


Figure 33 : Cohésion maximum des approches *tf*idf* et de cooccurrences

4.4.5 Sélection du seuil d'information mutuelle

En reprenant le même type d'analyse que celle menée à la section précédente, et en faisant varier les valeurs de seuils d'information mutuelle, nous avons obtenu les courbes de cohésion maximum présentées à la figure 34.

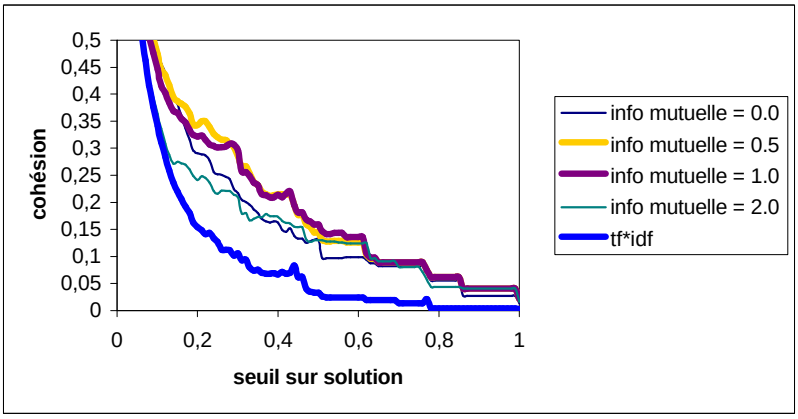


Figure 34 : Cohésion maximale en fonction de la valeur d'information mutuelle

On remarque que la stratégie de cooccurrences donne des valeurs supérieures à la stratégie de *tf*idf*, et ce, indépendamment du seuil de similarité des solutions. On note également que les valeurs maximales de cohésion sont obtenues pour des seuils d'information mutuelle se situant entre 0,5 et 1,0 (courbes jaune et mauve). Ces deux observations corroborent à nouveau les résultats obtenus lors de l'évaluation des modules de recherche au chapitre 3.

4.5 Discussion

Nous présentons dans cette section une discussion sur certains points connexes à notre étude sur l'évaluation du processus de construction. Les thèmes abordés sont la définition d'autres indicateurs de voisinage, la sélection du seuil de similarité, l'interprétation des ensembles de voisinage et le choix d'une méthode de structuration.

4.5.1 Autres indicateurs de voisinage

En plus de la cohésion, d'autres indicateurs que l'on retrouve dans la littérature CBR peuvent être définis à partir du cadre de voisinage que nous avons utilisé à la section 4.3.2. Voici quelques définitions qui pourraient être utiles pour des travaux futurs :

a) le degré d'*unicité* :

$$\text{degré\_unicité}(c1) = 1 - (Union(c_1, CB) / |CB|)$$

Un cas, de par sa description de problème ou de solution, s'apparente peu aux autres cas de la base. L'unicité permet de mesurer le degré d'originalité des descriptions de problèmes et de solutions.

b) le degré de *redondance* :

$$\text{degré\_redondance}(c1) = Inter(c_1, CB) / |CB|$$

Le cas, de par sa description de problème et ou de solution, peut être résolu par d'autres cas de la base. Lorsque les seuils de similarité sont élevés, cette mesure indique le degré de redondance du cas. Nous avons retenu l'intersection et non l'union pour définir cette mesure. Un cas dont on retrouve la solution ou le problème dans des cas différents n'est pas nécessairement redondant car il apporte une nouvelle manière de résoudre un problème.

c) le degré d'*inconsistance* :

$$\text{degré\_inconsistance}(c1) = Diff\_prob(c_1, CB) / E_{problème}(c_1, CB)$$

Cet indicateur reflète le fait que des problèmes similaires puissent être résolus différemment. L'ensemble des problèmes similaires est estimé par $E_{problème}(c_1,CB)$ et la proportion de solutions qui diffèrent correspond à $Diff\_prob(c_1,CB)$. Un nombre moindre de solutions inconsistantes donnent une plus grande pertinence au système.

d) le degré d'*incohérence* :

$$\text{degré\_incohérence}(c1) = Diff\_sol(c_1, CB) / E_{solution}(c_1, CB)$$

Cet indicateur reflète le fait que des solutions similaires puissent être envisagées pour des problèmes différents. Cette définition correspond à la proportion de cas présents dans *Diff_sol*.

4.5.2 Choix des seuils de pertinence des recommandations

Nous avions identifié à la section 4.1 comme objet d'étude la sélection de seuils de similarité pour statuer sur la pertinence d'un cas. Ces seuils sont utiles lorsque l'on souhaite présenter à l'usager, pour différentes situations, un nombre variable de solutions pertinentes. Il est à noter qu'actuellement, nous limitons actuellement nos suggestions dans notre système aux cinq cas les plus similaires.

Tel qu'illustré à la figure 35 , nous avons observé que la cohésion suit une courbe en cloche. Le comportement de cette métrique est étroitement lié à l'indicateur d'inconsistance. Pour des voisinages restreints (région A), la plupart des problèmes et des solutions coïncident (ce qui correspond aux cas de routine). Dans cette région l'inconsistance demeure faible. Par la suite (région B), l'élargissement du voisinage (c.-à-d. la réduction du seuil) amène des cas qui sont parfois cohésifs et parfois inconsistants. Lorsque les seuils deviennent faibles, l'ajout de cas ne fait qu'ajouter à l'inconsistance du voisinage (région C).

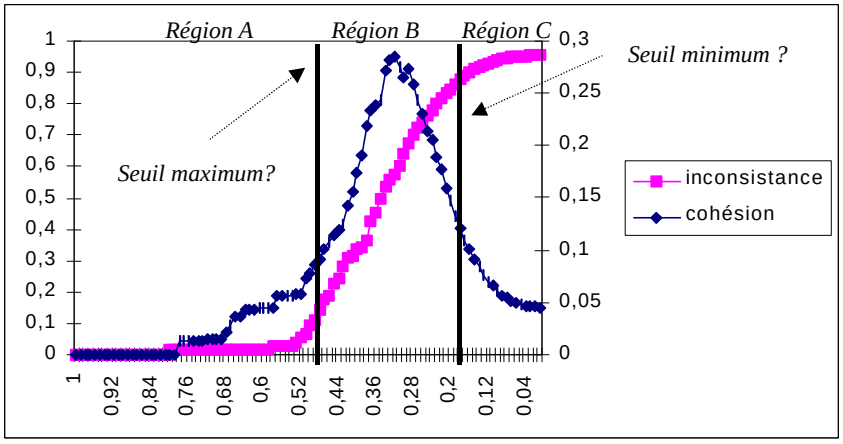


Figure 35 : Définition de seuils à partir de courbes de cohésion et d'inconsistance

Il nous semble qu'à partir de ces observations, des travaux futurs pourraient proposer des approches pour sélectionner des seuils minimal et maximal en fonction de la relation entre ces courbes d'indicateur. Ces approches nécessiteraient les choix des niveaux acceptables (inacceptables) d'inconsistance et des seuils de similarité des solutions.

4.5.3 Interprétation des indicateurs de voisinage

D'un point de vue plutôt qualitatif, on pourrait imaginer quatre situations (Figure 36) qui nous permettraient de relier les définitions d'indicateurs de voisinage à notre problème de réponse au courrier électronique :

- l'ensemble *Inter* contient la majorité des cas : ceci indique que la base contient des requêtes similaires à notre cas cible c_1 et que ces requêtes nous amènent vers des réponses analogues à celles de c_1. Le cas c_1 appartient donc à une catégorie de requête relativement bien définie et exprimée à l'aide d'un groupe restreint de mots. Au chapitre 3, les messages du sous-groupe A sur les dates de divulgation de résultats financiers en sont un exemple. Si la similarité de c_1 avec les autres messages est forte, on pourrait alors le retirer de la base de cas sans grande incidence sur le comportement du système.

- l'ensemble *Diff_sol* contient la majorité des cas : autrement dit, des solutions similaires à celle d'un cas cible sont utilisées pour différents problèmes. Cette situation représente plusieurs des cas dans notre base. Dans le cadre de notre application de réponse au courrier électronique, ceci se traduit par deux possibilités :

  - les réponses sont génériques et sont appliquées à des questions différentes ; des tests de généricité des messages tels que proposés par (Kosseim et Poibaud, 2001) pourraient aider à valider cette hypothèse.

  - Les réponses sont relativement uniformes et font suite à des questions différentes d'un point de vue lexical mais sémantiquement similaires. Puisque les questions proviennent de différents investisseurs, il est fort à parier que les textes paraphrasés présentent une faible homogénéité.

- l'ensemble *Diff_prob* contient la majorité des cas : cette situation indique que des questions similaires donnent lieu à des réponses exprimées différemment. Tel que mentionné précédemment, les réponses sont rédigées par un groupe restreint d'analystes et leurs textes sont assez uniformes. Il est donc légitime de présumer que des réponses exprimées différemment correspondent à des cas différents.

- Les ensembles *Diff_prob* et *Diff_sol* contiennent chacun un nombre considérable de cas : cette situation est difficile à interpréter. A prime abord, on serait porté à ne pas traiter immédiatement ces cas et à espérer qu'une meilleure structuration des autres cas permettra de réduire la taille de l'un ou l'autre de ces ensembles.

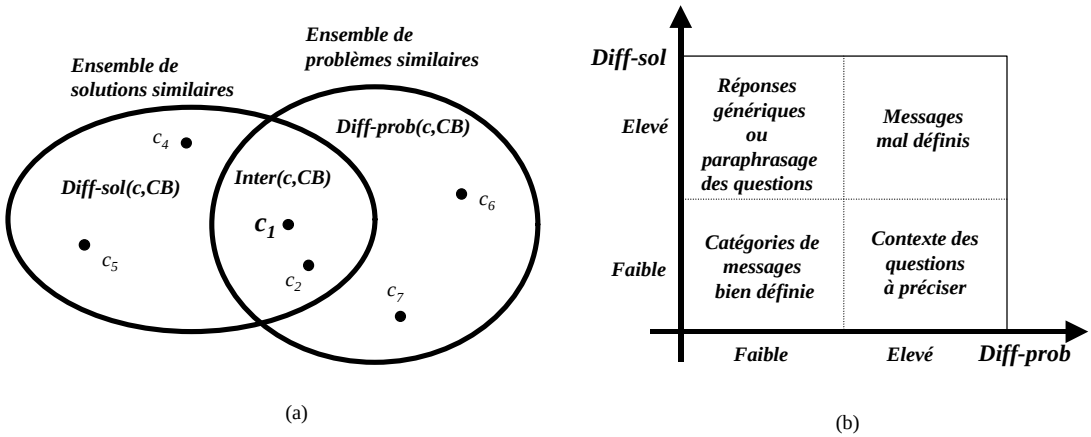


Figure 36 : Les ensembles de voisinage (a) et leur interprétation (b)

4.5.4 Choix de la structuration des cas en fonction du voisinage

Notre étude a été menée pour une base de cas dont les solutions sont plus homogènes que les problèmes. Il est donc pertinent de se demander si cette approche serait valide pour d'autres bases de cas. La réponse dépend de la nature de la base de cas considérée. Si nous reprenons notre découpage précédent (Figure 37), les cas de la région où la plupart des solutions et des problèmes concordent (*diff_prob* et *diff_sol* sont faibles) peuvent être adéquatement résolus par une approche de type *tf*idf*. Pour les bases de cas où l'ensemble *diff_sol* est élevé (indiquant l'homogénéité des solutions), les approches que nous avons proposées sont avantageuses. Pour les autres situations, il faudrait plutôt favoriser une meilleure structuration des problèmes soit par des approches cognitives (basée sur la modélisation du domaine) soit par des techniques pour mieux repérer le contenu informatif des descriptions de problèmes (par ex. l'extraction d'information).

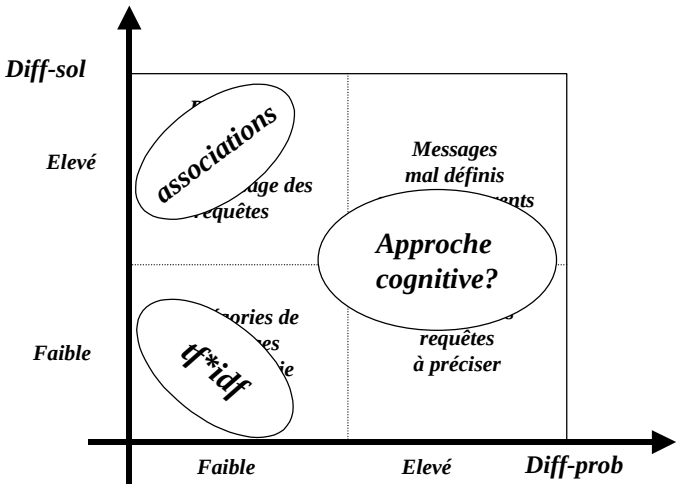


Figure 37 : Approche à préconiser en fonction des indicateurs de voisinage

4.6 Conclusion

Dans cette section, nous avons proposé une démarche pour guider le processus de construction du module CBR. Pour mener ces travaux, nous avons retenu trois métriques. Les métriques de recouvrement et de densité permettent de mesurer l'homogénéité des cas afin d'évaluer lesquels des problèmes ou des solutions permettent de mieux capturer la similarité entre les cas. De plus, nous avons défini une métrique de cohésion pour comparer les techniques de recherche de cas étudiées au chapitre 3.

Bien que la démarche que nous proposons dans ce chapitre soit utile pour notre application de réponse au courriel, son applicabilité à la construction d'autres systèmes CBR textuel est limitée car elle repose principalement sur l'hypothèse que les solutions sont plus homogènes que les problèmes. De plus, elle offre peu de support pour accomplir des tâches telles que la détection de cas inconsistants ou l'élagage de cas redondants. Des travaux additionnels seront nécessaires afin d'aborder ces aspects et évaluer l'adéquation de la métrique de cohésion pour des cas dont les solutions sont moins homogènes.

Chapitre 5 . Réutilisation d'un message antécédent

L'adaptation de cas, bien qu'abondamment étudiée pour des cas structurels, est une voie qui demeure inexplorée pour des solutions textuelles. Dans ce chapitre, nous présentons une approche de réutilisation textuelle dans le cadre de notre application de réponse au courrier électronique, c.-à-d. la réutilisation de courriels antécédents pour formuler des réponses à de nouvelles requêtes. Nous abordons plus particulièrement la tâche qui consiste à déterminer les portions réutilisables des cas passés. Par notre approche, nous visons à structurer les descriptions de solutions afin d'y identifier des portions variables, optionnelles et figées. Cette formulation de réutilisation de cas textuels s'inscrit dans les approches d'adaptation de type transformationnel et structurel du CBR. D'un point de vue applicatif, notre approche équivaut à créer dynamiquement un canevas de réponse à partir d'un message antécédent. Des techniques de regroupement, de condensation et d'extraction d'entités nommées sont préconisées pour mettre en oeuvre les étapes de cette approche. Une évaluation comparative et des résultats sont présentés ainsi qu'une discussion des travaux futurs.

5.1 Introduction

Contrairement aux approches CBR structurel qui offrent de nombreuses stratégies pour l'adaptation de cas structurés, la réutilisation des solutions textuelles demeure peu développée. Cet état du domaine s'explique par la nature des travaux en CBR textuel où dominent les tâches de recherche (*retrieval*) et les applications telles que la jurisprudence, les foires aux questions (*FAQ*), les leçons apprises, etc. lesquelles exigent peu de modifications au niveau des descriptions de solutions.

Néanmoins, certaines tâches orientées vers la rédaction de descriptions de nouvelles solutions pourraient bénéficier de la capacité de modifier le contenu des anciennes solutions textuelles. Un exemple d'une telle tâche est notre application de réponse au courrier électronique. Une réponse est une séquence d'énoncés qui satisfont le contenu d'une requête. Contrairement aux systèmes qui utilisent en guise de réponse des documents statiques et génériques (tels que les *FAQ*), la réponse aux courriels exige une personnalisation du contenu des messages et la modification d'informations spécifiques afin de satisfaire le contexte décrit dans une nouvelle requête. Des outils

d'aide à la rédaction de réponses pourraient s'avérer avantageux pour l'ajustement et la modification du contenu des courriels de réponse.

Dans ce chapitre, nous proposons une approche de réutilisation de solutions antérieures qui consiste en deux parties : la sélection des portions de texte méritant d'être récupérées dans la nouvelle solution et l'amélioration de la description de ces portions. Les cas que nous traitons, les courriels de requêtes et de réponses, sont des textes courts ayant une certaine répétitivité et comportant peu de structuration. Afin d'envisager la réutilisation de ces textes, on doit d'abord déterminer l'unité de texte à traiter, sa pertinence et sa spécificité. La plupart des approches CBR structurel de réutilisation adressent principalement la modification des solutions, dans des cas où les traits à modifier sont déjà prédéterminés. Dans un cadre textuel tel que celui de la réponse au courrier électronique, la faible structuration des solutions (c.-à-d. les réponses) rend difficile la mise en oeuvre de tels schémas. Ainsi, avant de modifier le contenu des messages, il est nécessaire de déterminer quelles portions des réponses méritent d'être réutilisées et modifiées. Nous concentrons nos efforts de recherche principalement sur l'identification dynamique des portions réutilisables. La mise en oeuvre des étapes de cette approche repose sur des techniques de traitement automatique des langues (TAL). Nous utilisons la base de cas de notre système pour acquérir les structures nécessaires à l'identification des portions réutilisables.

Dans les prochaines sections, nous revenons brièvement sur notre approche de réutilisation et nous la comparons aux approches de type "canevas de réponse". Par la suite, nous décrivons plus en détail le schéma de réutilisation que nous avons adopté ainsi que la mise en oeuvre des diverses étapes de ce schéma. Finalement, nous présentons des résultats expérimentaux d'évaluation et nous proposons quelques voies de recherche pour poursuivre ces travaux.

5.2 Réutilisation de réponses antécédentes

Après avoir sélectionné un cas (c.-à-d. une paire « requête, réponse ») par l'une des méthodes présentées au chapitre 3, le module CBR offre un support à la réutilisation en proposant à l'usager une version de la réponse annotée comme suit :

- les régions rouges indiquent les portions de texte jugées optionnelles par le système et qui pourraient être élaguées par le rédacteur.
- les régions vertes indiquent que ces informations spécifiques pourraient être modifiées pour tenir compte du contexte de la nouvelle requête.

Un exemple d'annotation est présenté à la Figure 38. La décision finale quant à la modification ou au retrait de passages textuels revient toutefois au rédacteur.

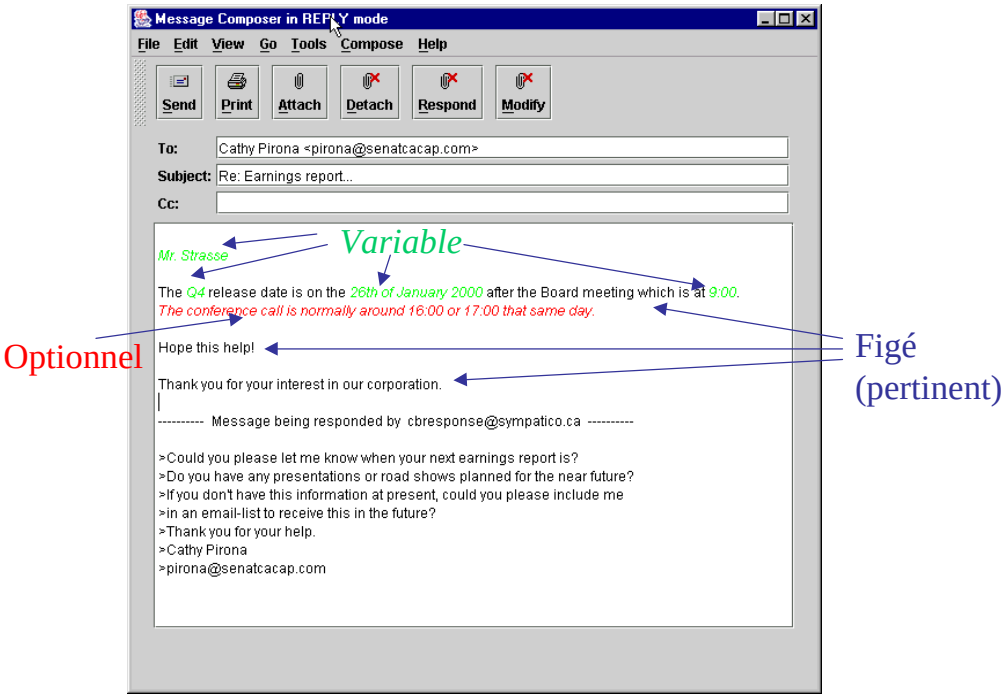


Figure 38 : Recommandation de réutilisation d'une réponse

La principale fonction de ce schéma de réutilisation n'est pas de modifier automatiquement un texte structuré mais plutôt de guider l'usager en identifiant les portions susceptibles d'être modifiées. On cherche lors de la réutilisation des réponses à

identifier les problèmes potentiellement intéressants et à les mettre en évidence afin que l'usager puisse prendre les actions nécessaires pour améliorer la qualité de la solution. D'un point de vue adaptation CBR, le système prend en charge la phase de réutilisation et laisse une partie du travail de révision à l'usager.

5.3 Gestion dynamique de canevas de réponse

En générant un texte contenant des annotations de réutilisation, on rejoint les approches de canevas de réponse ou de patrons à trous fréquemment utilisés en génération de texte. Toutefois, notre approche présente quelques particularités qui la rendent plus attrayante. Le choix de la position des trous, c.-à-d. l'annotation du contenu de la réponse, est dynamique car il dépend du potentiel de réutilisation du texte par rapport au contexte de la requête. Ainsi, l'utilisation multiple d'un message peut résulter en plusieurs canevas différents lorsque les choix d'annotations varient. De plus, le nombre de canevas n'est pas déterminé a priori lors de la conception du système. Tout nouveau cas devient un nouveau canevas qui pourra être réutilisé pour les épreuves subséquentes de réponse. Les canevas n'ont pas à être créés manuellement, ce qui facilite la construction initiale du système. Et l'ajout de cas à la base de cas enrichit le nombre de situations pouvant être traitées sans nécessiter des modifications substantielles du système.

Un autre point important est que la base de cas peut être mise à contribution pour déterminer l'annotation des portions de solutions. Les cas contiennent des descriptions qui sont récurrentes et qui sont des exemples permettant d'établir des relations entre le contexte des requêtes et les portions de réponse. Ces relations sont mises à profit pour évaluer l'adéquation d'un contexte pour la réutilisation de certaines portions de textes. Cette analyse peut être faite à partir du corpus de message, évitant ainsi l'acquisition d'un modèle de connaissance explicite.

L'insertion des annotations dans le texte d'une réponse équivaut à en faire une généralisation. De la séquence initiale d'énoncés, certains sont déclarés « optionnel » (ce qui donne différentes options de sous-séquences) et certaines portions de ce qui est figé

sont converties en texte « variable ». Comme la pertinence de certains énoncés dépend peu du contexte de la requête, nous avons observé qu'un niveau adéquat de généralisation est parfois difficile à déterminer. Tel qu'illustré à la partie (a) de la Figure 39, une généralisation légère permet de clairement mettre en relief les passages qui méritent d'être retravaillés par l'usager. Dans la partie (b) de cet exemple, les formes de courtoisie sont identifiées comme optionnelles et toutes les entités nommées sont déclarées comme variables. Ces choix entraînent un niveau de généralisation élevé qui rend plus difficile la réutilisation de cette portion du texte. Cet exemple illustre le besoin de faire preuve de parcimonie dans l'annotation des textes et de mesurer l'impact des stratégies agressives qui annotent fortement le texte.

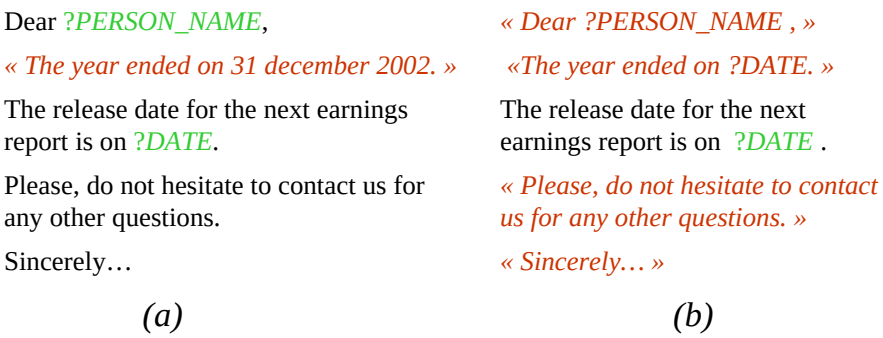


Figure 39 : Généralisation d'une réponse antécédente : (a) une généralisation de quelques passages ; et (b) une généralisation des formes de courtoisie

5.4 Étapes du schéma de réutilisation

En présence d'une nouvelle requête, certains énoncés d'une réponse antécédente ne satisferont plus le nouveau contexte. Bien qu'une approche d'adaptation qui reformulerait complètement les réponses textuelles ne puisse être envisagée avec les techniques actuelles CBR et de traitement des langues naturelles, quelques techniques peuvent néanmoins apporter une contribution pour préserver la pertinence des énoncés en fonction du nouveau contexte et s'assurer que le message est adéquatement circonstancié.

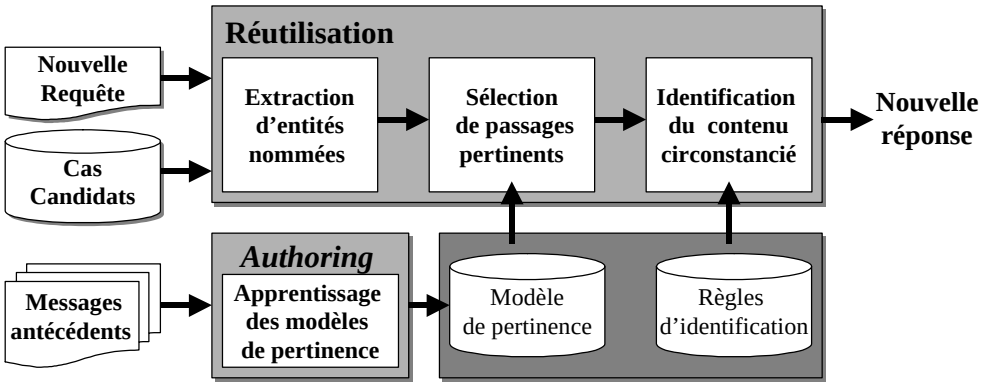


Figure 40 : Étapes du processus de réutilisation de solutions textuelles

Étant donnés un nouveau message et une solution antécédente sélectionnée durant la phase de recherche, nous implantons la réutilisation de cas textuel selon les étapes suivantes (illustré à la Figure 40) :

- *Sélection des passages optionnels* : cette étape correspond à déterminer les portions textuelles qui ne sont plus applicables au contexte de la nouvelle requête. Dans un premier temps, nous découpons les réponses antécédentes en passages, plus spécifiquement en phrases individuelles. Par la suite, nous évaluons les différents passages et nous retenons le sous-ensemble jugé le plus pertinent. La pertinence est établie en fonction du contenu la nouvelle requête. Cette évaluation permet de figer les portions pertinentes (c.-à-d. les recommander à l'usager) et de rendre optionnel le reste de la description de solution (c.-à-d. inciter l'usager à réviser ces passages). Nous comparons deux approches pour extraire ce sous-ensemble de phrases. Une description détaillée de ces approches est présentée à la section 5.4.

- *Identification du contenu circonstancié* : parmi les phrases pertinentes, quelles sont les portions de textes qui sont sujettes à modification ? Pour cette étape, nous identifions les différentes portions susceptibles d'être inexactes par rapport au contexte de la nouvelle requête. Ces portions sont circonstanciées parce qu'elles font référence à des individus, des lieux, des adresses et d'autres informations spécifiques qui peuvent varier selon le contexte ou les références temporelles. Ces informations résident principalement dans les entités nommées du texte et nous

disposons de techniques permettant d'en faire l'extraction. La pertinence de ces entités peut par la suite être déterminée selon le contexte de la nouvelle requête. L'utilisation de techniques d'extraction d'information (Cowie & Lenhert 1996) pour cette étape est présentée à la section 5.5.

- *Élagage et substitution :* le retrait des portions non pertinentes et la substitution des portions à être circonstanciées sont prises en charge à cette étape. Bien que ces opérations repose principalement sur l'usager, nous discutons de ces aspects aux sections 5.4 et 5.5.

Cette approche présente certains avantages. En réutilisant des formulations éprouvées de réponses, aucun traitement syntaxique n'est nécessaire et on favorise le contrôle de l'uniformité, la qualité, et la fluidité du contenu. En héritant du contenu d'un message antécédent, on évite les efforts de planification de contenu. En limitant les portions variables à des informations factuelles reposant sur des entités nommées, on évite les problèmes de génération de surface (morphologie, accord en genre et nombre, ponctuation…). Les réponses contiennent quelques phrases (habituellement moins de 10), mais il est plus rapide de repérer automatiquement les portions pertinentes (quelques millisecondes) que de localiser soi-même ces messages et de faire un copier-coller manuel (plusieurs secondes).

5.5 Sélection des portions optionnelles

La sélection des portions optionnelles permet de réorganiser le contenu de la réponse antécédente en recommandant les portions qui sont jugées potentiellement superflues. En déclarant des phrases optionnelles (ou figées), on s'assure que le texte de la réponse recouvre adéquatement le contenu de la nouvelle requête.

Le choix de granularité des portions de texte à élaguer peut varier en fonction de la tâche à accomplir : mots individuels, groupes syntaxiques, sous-séquences de mots, etc.… Pour notre application, la pertinence des énoncés de la solution repose sur la phrase, assurant ainsi la cohésion et l'intelligibilité du texte suite à l'élagage de certains

passages par l'usager. On suppose qu'un énoncé correspond à une phrase et que cet énoncé porte sur un thème. Par ailleurs, ce choix n'est pas critique à l'application des techniques que nous préconisons.

Pour trouver les phrases de la réponse qui recouvrent le mieux une requête, on accomplit les trois tâches suivantes :

a) le découpage : nous segmentons les réponses antécédentes en phrases individuelles; le logiciel que nous utilisons pour l'étiquetage lexical des textes (*lmtag*) détermine les débuts de phrase et de paragraphe.

b) l'évaluation de la pertinence : on estime pour chaque phrase individuelle la pertinence par rapport au contenu de la requête;

c) la sélection : on choisit les phrases qui semblent les plus prometteuses et on les présente à l'usager comme figées. Les autres phrases sont jugées optionnelles par le module.

Pour identifier les phrases figées/optionnelles, nous devons établir une correspondance entre les énoncés d'une solution et le contenu d'un problème. Tel que nous l'avons étudié au chapitre 3, des relations entre les descriptions textuelles permettent de vérifier qu'une solution trouve sa correspondance dans la requête. Toutefois, ces relations sont parfois faibles ou inexistantes pour les phrases accessoires du texte comme les salutations, les formes de courtoisie et les informations générales. Bien que ces phrases ne soient pas essentielles, elles jouent un rôle important dans la forme du message et, idéalement, elles devraient être préservées si elles se prêtent au contexte de la requête.

Nous étudions et comparons deux stratégies pour les étapes d'évaluation et de sélection :

- nous évaluons chaque phrase individuellement et nous retenons celles qui obtiennent un support suffisant du contenu de la requête. Pour évaluer une phrase

provenant d'une solution antécédente, on identifie les cas qui confirment ou infirment la correspondance entre une phrase cible et une requête. La similarité entre les différents cas de la base permet de déterminer si la phrase doit être retenue. Nous présentons cette approche à la section 5.5.1.

- la seconde stratégie est de retenir le sous-ensemble de phrases qui recouvre le mieux le contenu de la requête. Ce traitement de la pertinence au niveau d'un groupe de phrases équivaut à une condensation de texte, où on vise à trouver l'ensemble qui répond le mieux à la nouvelle requête. Ce type de condensation est fréquemment dénommé dans la littérature *query-biased*, *query-based*, *query-relevant* ou *user-focused*. Nous présentons cette stratégie à la section 5.5.2.

Notre but est de conserver les phrases de la réponse qui obtiennent un support suffisant de la requête. Pour déterminer ce support, la base de cas est utilisée pour modéliser les connaissances nécessaires à la construction des deux stratégies. La base contient différents exemples de cas qui permettent établir une correspondance entre les descriptions de problèmes et de solutions.

5.5.1 Approche par regroupement

Notre première approche est de déterminer pour chacune des phrases individuelles si elle devrait être retenue dans la solution proposée à l'usager. Pour guider cette décision, nous utilisons les cas de notre module CBR. Pour chaque phrase $phrase_j$ de la solution $Sol_{recommandée}$ (la réponse que l'on réutilise), nous identifions dans la base les cas $Cas_{support}$ qui comportent un ou plusieurs énoncés similaires à la phrase $phrase_j$ et les cas Cas_{rejet} qui n'en contiennent pas. Notre problème revient à déterminer, étant donnés un nouveau problème (une requête) et un certain nombre d'exemples de correspondance <*problème,solution*>, si la solution recommandée à l'usager devrait contenir la phrase cible $phrase_j$, c.-à-d. :

Etant donné que

$Cas_{support}$ { $Problème_1 \rightarrow Contient\ (Sol_1, phrase_j)$ … $Problème_i \rightarrow Contient\ (Sol_i, phrase_j)$

Cas_{rejet} { $Problème_{i+1} \rightarrow \text{`Ø}\ Contient\ (Sol_{i+1}, phrase_j)$ … $Problème_n \rightarrow Ø\ Contient\ (Sol_n, phrase_j)$

Est-ce que

$Nouveau\_Problème \overset{?}{\rightarrow} Contient\ (Sol_{recommandée}, phrase_j)$

Tel qu'illustré aux Figures 41 et 42, nous divisons notre base de cas en deux groupes qui permettent de déterminer les contenus de requêtes qui justifient l'utilisation d'une phrase spécifique dans la réponse. Par la suite, nous générons des distributions de requête qui caractérisent les ensembles $Cas_{support}$ et Cas_{rejet} . Suite à la comparaison des distributions et la nouvelle requête, nous déterminons l'appartenance de la phrase cible à la solution recommandée.

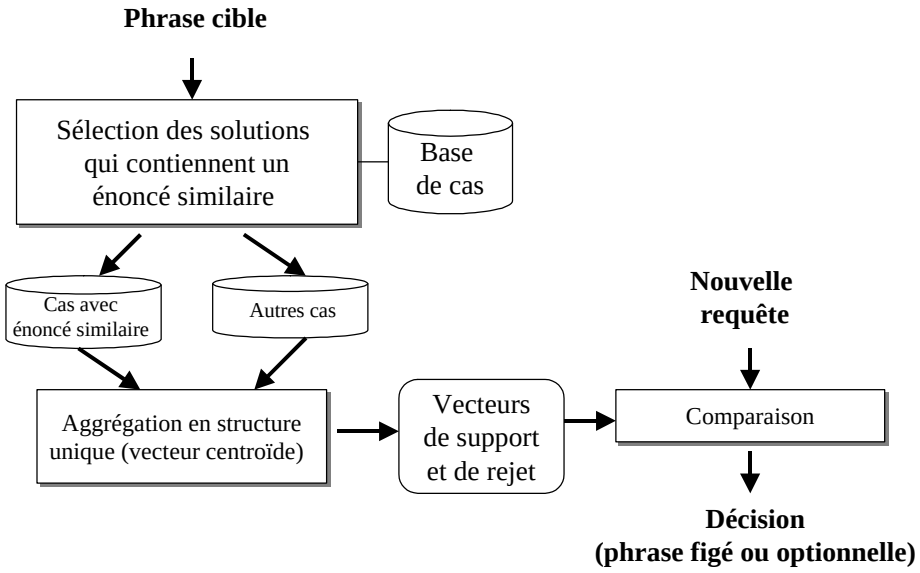


Figure 41 : Évaluation afin de décider si une phrase est optionnelle

L'appartenance d'une phrase cible à une solution (c.-à-d. le prédicat *Contient*) est estimée selon la similarité entre la phrase cible et chacune des phrases d'une solution.

Nous avons essayé quelques fonctions pour évaluer la similarité entre les phrases de solution et nous avons retenu pour nos expérimentations une fonction construite à partir du coefficient d'*Overlap* , c.-à-d.

$$similarité_{Overlap}\left(phrase_{cible}, solution\right) = \max_{phrase_i \in solution}\left(\frac{\left|phrase_{cible} \cap phrase_i\right|}{\min\left(\left|phrase_{cible}\right|, \left|phrase_i\right|\right)}\right)$$

Cette métrique permet d'estimer la proportion de mots en commun entre les solutions. Une métrique de similarité statistique donne de bons résultats pour nos descriptions de solution puisqu'elles présentent une bonne cohésion. Tel qu'observé dans nos expérimentations sur la recherche de cas pertinents, les réponses sont réutilisées par les analystes et la description d'énoncés similaires varie peu d'un cas à l'autre. Toutefois, d'autres métriques de nature sémantique ou l'utilisation des ressources du domaine pourraient s'avérer nécessaires pour des réponses ayant des formulations plus variées.

La fonction $similarité_{Overlap}$ est utilisée pour faire la partition de la base de cas. Tous les cas dont la valeur de similarité est supérieure à un seuil se retrouve dans $C_{support}$ tandis que les autres se retrouvent dans le groupe C_{rejet} . Quelques essais nous permettent de déterminer un seuil empirique sur la valeur de similarité entre phrases. Puisqu'ils ne dépendent pas de la nouvelle requête, ces ensembles de cas peuvent être déterminés lors de la construction du système CBR.

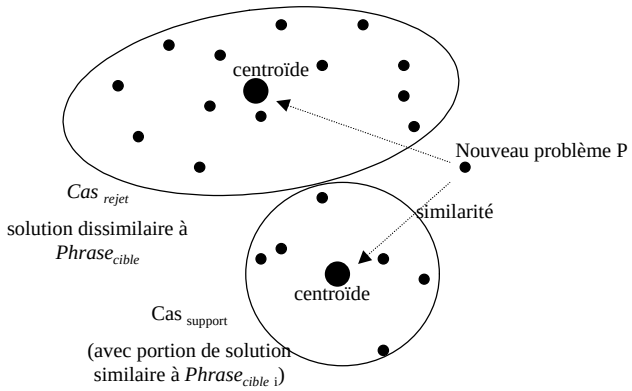


Figure 42 : Partition de la base de cas en groupes de support et de rejet

Les deux groupes de cas obtenus par la partition de la base caractérisent les descriptions de problème qui favorisent ou défavorisent l'utilisation d'une phrase similaire à $phrase_{cible}$. En estimant la proximité du nouveau problème *P* avec les cas de ces deux groupes, nous pouvons extrapoler s'il y a une correspondance entre *P* et $phrase_{cible}$. Pour estimer cette proximité, chaque groupe est représenté par une structure de centroïde qui agglomère la représentation vectorielle de ses problèmes. Les termes des problèmes ont fait l'objet d'une lemmatisation et ont été filtrés selon le vocabulaire du module CBR. Nous calculons le centroïde de chaque groupe à partir des vecteurs de fréquences des termes de requêtes. Nous évitons les poids de type *tf*idf* qui mesurent le pouvoir discriminant entre les cas alors que nous souhaitons plutôt discriminer entre les groupes de cas.

La similarité entre la nouvelle requête *P* et le centroïde de chacun des groupes est déterminée par un cosinus de leurs vecteurs. On retient la phrase si la similarité de la nouvelle requête avec $Cas_{support}$ est plus grande, c.-à-d.

$$similarité_{cosinus}(P, centroïde_{support}) > similarité_{cosinus}(P, centroïde_{rejet})$$

Si cette inégalité n'est pas respectée ou si la base ne comporte aucun cas ayant un énoncé similaire, alors la phrase est jugée optionnelle.

5.5.2 Approche par condensation

La deuxième approche que nous évaluons ne mise pas sur l'évaluation individuelle de chacune des phrases mais plutôt sur la qualité globale d'un sous-ensemble de phrases provenant de la réponse réutilisée. La présence de passages non pertinents est principalement due à l'occurrence de thèmes multiples dans les requêtes et dans les réponses. L'identification de ces passages correspond à la production d'un sous-ensemble des réponses précédentes qui recouvre le mieux le contexte de la nouvelle requête. En traitement des langues naturelles, ceci correspond à un processus de résumé contraint par une requête, plus spécifiquement à la condensation d'un texte basé sur les termes de la requête. Dans cette variante, une requête indique ce qui intéresse l'usager

(ce qu'il cherche) et les portions de textes qui se retrouvent dans le résumé doivent concorder avec l'expression de la requête.

Tel qu'illustré à la figure 43, la solution résultante S_c peut être produite par le retrait, de la solution originale *S*, de phrases qui peuvent être associées ("alignées") au nouveau problème *Q*.

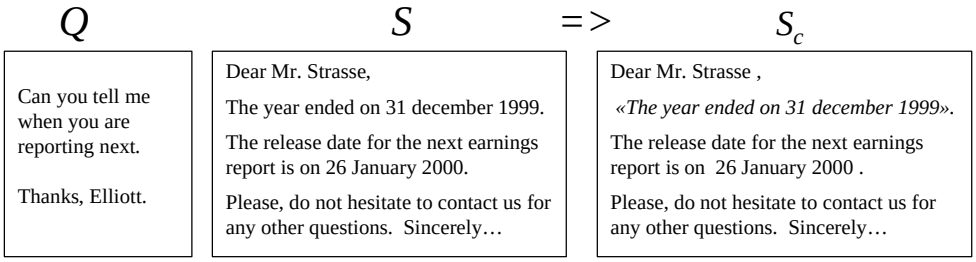


Figure 43 : Identification de passages pertinents par un processus de condensation

Tel que proposé par (Mittal & Berger 2000), le processus d'appariement tente de déterminer le sous-ensemble de S qui recouvre le mieux le problème *Q*. En terme de probabilité, nous tentons de trouver une solution condensée *S'*, qui maximise l'estimation de probabilité suivante :

$$S_c = f(Q,S) = \arg\max_{S'} P(S'|S,Q) \qquad (1)$$

L'expression $P(S'|S,Q)$ donne une estimation (en terme de probabilité) qu'une réponse *S'*, extraite d'une réponse antécédente *S*, puisse satisfaire les énoncés de la requête *Q*. A partir de la règle de *Bayes*, on peut définir une approximation de cette expression comme suit :

$$\begin{aligned} S_c &= \arg\max_{R'} P(S'|S,Q) \\ &= \arg\max_{R'} P(Q|S',S)P(S'|S) \\ &\sim \arg\max_{R'} P(Q|S')P(S'|S) \end{aligned}$$

Ainsi, cette formulation suggère que le texte qui est recommandé à l'usager (c.-à-d. figé) est un compromis entre et la préservation de l'intégrité de la solution antécédente et un sous-ensemble de cette solution qui colle le mieux au nouveau problème.

L'expression *P(S'|S)* peut être modélisée comme une pige aléatoire de mots provenant de la solution originale *S*. Quelques distributions (par e*x.* multinomiale, hypergéométrique) permettent d'évaluer la probabilité du condensé résultant. Dans nos travaux, nous modélisons la distribution *P(S'|S)* par une distribution multinomiale

$$P(S'|S) = \frac{\prod_{i \in S} tf_i! \times \left(tf_i / N \right)^{c(i \in S')}}{N!}$$

où tf_i est la fréquence du terme *i* dans la solution *S*, *N* est le nombre de terme dans *S* et *c* est le nombre d'occurrence du terme *i* dans le condensé *S'*. Comme les solutions sont relativement courtes et que la plupart des termes n'apparaissent qu'une fois dans le texte, nous pouvons faire une approximation de cette distribution comme suit :

$$P(S'|S) \approx \frac{1}{|S|! \times |S|^{|S'|}}$$

L'expression *P(Q|S')* correspond à la probabilité qu'un nouveau problème *Q* soit à l'origine d'une solution *S'*. Nous modélisons cette probabilité par un modèle IBM1 tel que préconisé au chapitre 3. Nous exploitons la base de cas du module pour apprendre les distributions du modèle à l'aide d'un algorithme EM. Certains paramètres du modèle sont obtenus lors de l'entraînement. Pour attribuer une probabilité aux nombreuses valeurs manquantes, nous utilisons un lissage de type *backoff*, c.-à-d.

$$p(q_i \mid s_j) = \begin{cases} t(q_i \mid s_j) & \text{la valeur de probabilité de transfert} \\ a_i \, p_{CB}(q_i) & \end{cases}$$

où *t* est le modèle de transfert obtenu par apprentissage, p_{CB} est la distribution des termes dans la base de cas (dans notre corpus de messages courriels) et a_i est une

constante de normalisation (*discounting*) qui ramène la masse de probabilité à 1. Nous menons nos expérimentations avec un vocabulaire fermé (celui sélectionné pour construire la base de cas). Il faudrait toutefois ajouter une distribution supplémentaire (par exemple une distribution uniforme) à cette fonction de lissage pour les applications qui prennent en compte l'ajout de nouveaux termes.

Pour sélectionner le condensé le plus prometteur, une recherche exhaustive est possible pour de courte solution mais s'avère très coûteuse en temps lorsque le nombre de phrases est considérable. Nous avons donc implanté aussi une recherche de type vorace qui part d'une solution complète et qui tente de retirer itérativement la phrase qui diminue le plus la valeur de probabilité. Si pour une itération, aucune phrase n'apporte de diminution de la probabilité exprimée en (1), alors la recherche s'interrompt et on suggère à l'usager les phrases qui n'ont pas été retirées.

5.6 Identification du contenu circonstancié

Pour notre application, il se révèle que l'exactitude des énoncés repose principalement sur des informations factuelles contenues dans des groupes nominaux. Nous avons observé dans notre corpus que la plupart des modifications à apporter pour réutiliser une réponse antécédente repose sur des informations spécifiques comme des numéros de téléphone, des noms de compagnies, des dates… Ces informations sont des entités nommées qui peuvent être obtenues par extraction d'information (IE). Ainsi une bonne extraction de ces entités avec une modélisation de leur rôle permet de cerner l'essentiel des informations à modifier.

La partie de notre application sur le traitement des portions variables est dépendante du domaine et nous l'avons construite manuellement à partir d'une analyse de notre corpus. Les trois étapes pour mener ce traitement sont les suivantes : l'extraction des entités nommées, l'attribution des rôles et les décisions de substitution.

a) L'extraction des entités nommées

Pour notre application, nous avons identifié les catégories d'entités suivantes qui présentent un potentiel de réutilisation :

- les dates : des années (par ex. 1999, *coming year*), des mois (par ex. *April*), des jours (par ex. *5 June 2000*, Jan/01, *Tuesday May 12*, *tomorrow*), des périodes de temps (par ex. *last ten years*, *past 6 months*, *first quarter*) et des marques temporelles (par ex. 16:00, 9:00AM, 2:00EST).

- les personnes : des combinaisons de noms, de prénoms, d'initiales (par ex. *P. J. Smith*) et de titres (par ex. *Mr.* ou *Madam*).

- les organisations : des noms propres qui ne désignent pas des individus. Plusieurs contiennent des mots-clés comme *Capital*, *Corporation*, *Associates*, *Bank*, *Trust*, *Inc*, *Department*,

- les adresses : on retrouve des URL, des adresses courriels et des numéros civiques.

- des lieux : des noms ou des acronymes de pays, d'états, de provinces, de régions et de villes (ex. *Canada*, *Boston*, *Thames Valley*, *USA*, *UK*, *NY*).

- des quantités : des monnaies (*currency*), des pourcentages, des entiers, des réels et des fractions.

- des numéros de téléphone : des expressions ont été définies pour les formats de différents pays.

La plupart de ces entités nommées sont obtenues avec le logiciel Gate (Cunningham et al. 2002). Nous avons également utilisé quelques expressions régulières (que nous avons implantées avec la librairie de la version 4 de Java) pour capturer les numéros de téléphone et d'autres informations spécifiques telles que les sections du site web de BCE (par ex. *Dividend information*) et des noms de document de la compagnie.

Par ce procédé, nous avons obtenu des textes étiquetés selon les catégories prédéterminées identifiées ci-haut.

b) L'attribution des rôles

Les catégories de base permettent une première estimation des portions de texte sujettes à modification. Toutefois, leur rôle dans le domaine du service aux investisseurs doit être mieux défini afin de prendre une décision quant à leur réutilisation. Pour définir le rôle d'une entité, nous tenons compte de sa catégorie et de son type (par ex. une date de type *time*), de la chaîne de caractères qu'elle contient et du contexte défini par les mots voisins dans la phrase. Par exemple, le rôle "*conference_time*" s'applique à une date de type *time* qui est précédée d'un des mots *conference* ou *call*. Le rôle "*subsidiary*" est une entité *Organisation* de type *Company* qui contient l'un des noms d'une compagnie filiale de BCE.

Pour la définition du rôle de l'émetteur d'une requête, nous pouvons aussi utiliser l'adresse courriel de retour. La position de la phrase dans la réponse est également prise en compte car les réponses commencent souvent par une forme de salutation du genre "*Good morning Mr Smith.*".

Une description des principaux rôles que nous avons définis pour notre application est présentée au Tableau 15.

Date	*publication date, release date, sending_date, fiscal_year, end fiscal year, financial quarter, meeting time, conference call time*
Lieu	country, home, work
Personne	*sender, responder*, investor
Quantité	*financial quarter, earnings, eps, dividend, rps, rate, year, P/E ratio, time duration, specific year, growth, share price*
Compagnie	main (i.e. BCE), *subsidiary*, employer, financial institution, tse, news_paper
Section du site Web et nom de document	Pas de spécialisation du rôle
Adresse	personnal address, company adress, *company_URL*, others_URL
No de téléphone	personnal number, *conference call number*, company number

Tableau 15 : Description des rôles d'entités nommées

c) La substitution des valeurs d'entités :

A cette dernière étape, nous déterminons si, pour certains rôles, une valeur de substitution peut être suggérée à l'usager. Le domaine du service aux investisseurs présente peu de prédictibilité quant à la substitution des valeurs d'entités nommées. Toutefois, il est possible pour certaines informations d'utiliser des relations entre les entités de la requête et de la nouvelle réponse. Nous avons considéré les 3 possibilités suivantes :

- un rôle ne fait jamais l'objet d'une modification : quelques rôles ne sont pas modifiables car leur valeur est soit figée pour le domaine ou soit non disponible pour le système et doit être déterminée par l'usager. Par exemple, BCE est la corporation principale pour tous les messages. Les références à des lieux sont difficilement remplaçables à partir des informations disponibles dans les messages. De plus, plusieurs rôles ne surviennent que dans les requêtes (par ex. les nom de journaux, une adresse ou un URL personnel, le nom d'un employeur).

- la valeur du rôle peut être extraite de la requête : quelques rôles peuvent être instanciés à partir des entités correspondantes présentes dans le contenu de la nouvelle requête. Par exemple, les messages peuvent être personnalisés si la requête contient le nom de l'investisseur qui a émis le message. Également, des dates indiquant l'année fiscale, le trimestre financier ou l'envoi du message sont parfois disponibles dans la requête. Finalement, les noms de filiales de BCE mentionnées dans les réponses proviennent souvent des requêtes. Ces rôles sont en *rouge* dans le tableau 15.

- la valeur du rôle peut être modifiée si elle est emmagasinée au niveau du système (*lookup*) : pour ces rôles, il n'est pas possible de localiser une valeur dans les messages et il faut définir un certain nombre de variables au niveau du système et stocker les recommandations qui pourront être suggérées à l'usager. La plupart des facteurs financiers, des dates, de références

temporelles, des noms de documents et de site web sont de ce type. Pour certains de ces rôles, une seule valeur est suggérée (par exemple, un numéro d'appel conférence) alors que pour d'autres, nous offrons une liste de valeurs possibles (par exemple les noms de documents). Dans ce dernier cas, l'usager sélectionne la valeur la plus appropriée. Ces rôles sont en *vert* dans le tableau 15.

Ainsi en restreignant la sélection de valeurs de substitution en fonction du rôle, l'efficacité de notre approche repose principalement sur la capacité du système à repérer les entités nommées (étape a) et à leur attribuer un rôle adéquat (étape b). Nous évaluons à la section suivante ces deux fonctions à partir de phrases tirées de notre corpus de messages.

5.7 Évaluation et résultats expérimentaux

Lors de nos expérimentations, nous avons retiré les entêtes et les signatures des messages. Toutefois, les formes de courtoisie comme les salutations et les commentaires généraux ont été conservés. Pour mener ces expérimentations, nous avons repris le corpus utilisé au chapitre 3, c.-à-d. les cas du répertoire *financial information*. Ce choix nous permet de déterminer, pour chacune des requêtes du corpus, des réponses antécédentes sélectionnées par le module de recherche. Cela nous permet de concentrer notre évaluation sur la réutilisation de messages recommandés à l'usager par le système.

5.7.1 Résultats sur les portions optionnelles

Nous débutons notre évaluation par quelques expérimentations avec les deux stratégies présentées à la section 5.5. Premièrement, nous avons évalué la sélection de portions optionnelles des réponses en fonction des requêtes qui leurs sont associées dans le corpus. Ceci correspond à une évaluation de type *leave-one-in* en utilisant chacun de nos cas comme problème cible. On obtient un *accuracy* (justesse) de 89% avec la stratégie par regroupement de cas et 77% avec l'approche par condensation. Comme les cas étaient présents dans la base lors de la sélection, on présume que ces résultats sont des

bornes supérieures que le système peut atteindre avec ces deux stratégies. On note une différence principale entre les deux stratégies. L'approche par condensation tend à laisser tomber les formes de courtoisie et les énoncés plus généraux, contrairement à la stratégie de regroupement qui tend plutôt à les préserver. Ainsi l'approche par condensation semble plus conservatrice que celle par regroupement.

Afin d'obtenir une estimation plus représentative de l'utilisation de notre module CBR, nous avons par la suite sélectionné un échantillon de 50 paires (requête, réponse) pertinentes obtenues par notre système lors de la phase de recherche[15]. Pour chacune de ces paires, nous avons déterminé à partir des différents messages de notre corpus le sous-ensemble de phrases devant être sélectionné. Lors de cette évaluation, nous avons parfois éprouvé des difficultés à déterminer si les phrases accessoires (par ex. les formes de courtoise) devraient être incluses ou non dans la réutilisation de messages. Pour tenir compte de cette difficulté, nous avons mené deux expériences distinctes en considérant les phrases accessoires comme facultatives ou non. Les résultats obtenus pour ces deux expériences sont présentés aux tableaux 16 et 17.

Stratégie	Précision	Rappel	*Accuracy*
Regroupement	84.1%	68.0%	70.8%
Condensation	78.4%	39.6%	50.2%

Tableau 16 : Sélection des portions pertinentes avec les phrases accessoires

Stratégie	Précision	Rappel	*Accuracy*
Regroupement	77.7%	76.0%	71.8%
Condensation	78.5%	53.1%	62.2%

Tableau 17 : Sélection des portions pertinentes sans les phrases accessoires

L'approche par regroupement reconnaît une forte proportion des phrases qui ont un lien avec la requête. La similarité des résultats aux tableaux 16 et 17 indique qu'elle

[15] Les pairs ont été obtenues avec l'approche de recherche utilsant des listes de cooccurrences présentées à la section 3.3.

préserve également la plupart des phrases accessoires. Quelques phrases sont rejetées alors que d'autres sont tous simplement négligées parce qu'elles sont largement répandues dans les différents cas de notre base.

L'approche de condensation présente un tout autre comportement. Plusieurs des mots contenus dans les phrases peu fréquentes ne sont associés à aucun mots des requêtes. De plus, presque toutes les phrases accessoires sont rejetées puisque qu'aucun lien statistique n'est établi entre ces phrases et leurs requêtes. Ainsi on note une augmentation du rappel lorsque ces phrases ne sont pas prises en compte (tableau 17).

L'approche par regroupement est paramétrique, car elle dépend du seuil de similarité entre les phrases de solutions pour la partition de la base de cas. En faisant varier ce seuil, nous obtenons les courbes de performance suivantes (Figure 44) :

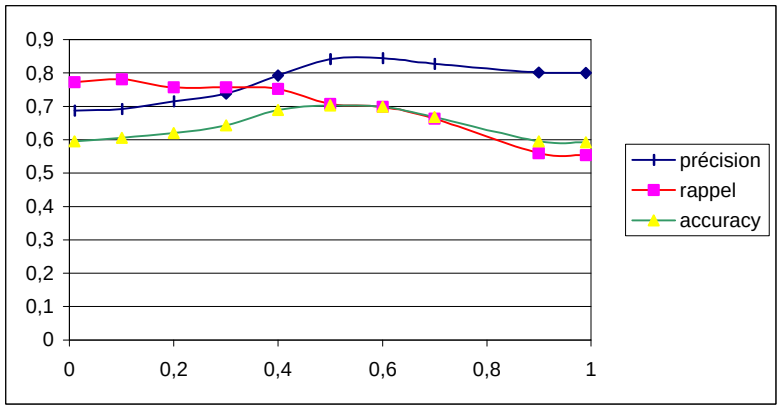


Figure 44 : Courbe de précision rappel en fonction du seuil de pertinence

On note qu'un seuil qui varie entre 0.4 et 0.6 donnent une bonne précision sur les sélections. Des seuils élevés donnent également des résultats intéressants, ce qui reflète une réutilisation fréquente des énoncés dans les réponses.

5.7.2 Résultats sur les portions variables

Pour cette partie de l'expérimentation, nous avons retenu 130 phrases de notre corpus qui contiennent plus de 250 entités nommées et auxquelles nous avons appliqué

les deux étapes d'extraction des entités et d'attribution des rôles. Lors de la compilation des résultats, nous avons attribué un point lorsque l'étiquetage du texte contient l'entité et ½ points lorsque l'étiquetage recouvre partiellement la description de l'entité (par exemple *« Bell » Canada*). Les résultats obtenus lors de nos expérimentations sont présentés au tableau 18.

Entité	*Extraction d'entités*		*Attribution de Rôle*
	Précision	*Rappel*	*Accuracy*
Date	91.7%	85.6%	82.9%
Temps	100%	100%	61.1%
Lieu	71.4%	93.5%	66.6%
Personne	100.0%	80.0%	81.8%
Quantité	92.2%	95.6%	68.7%
Organisation[16]	97.2%	83.3%	94.4%
Adresse[17]	100.0%	100.0%	100.0%
No de téléphone	95.4	90.9%	81.8%

Tableau 18 : Résultats pour l'extraction des entités et l'attribution de leur rôle

On remarque que l'extraction de plusieurs catégories d'entités donnent de bons résultats (les colonnes précision et rappel du tableau 18). Par exemple, les quelques erreurs de date portent sur les descriptions de trimestres financiers (par ex. *Q4*, *4th quarter*) et Bell Canada est annoté par Gate comme un nom d'organisation suivi d'un lieu. Nous pouvons éliminer plusieurs de ces erreurs avec l'ajout de quelques patrons d'extractions et de termes au lexique. Les noms de filiales de la famille BCE ont été ajoutés suite à l'expérimentation pour augmenter le rappel de la catégorie *Organisation*.

Par la suite, nous avons construit une base de règles à partir d'un sous-ensemble de notre corpus original et nous avons attribué des rôles aux entités nommées des 130 phrases de notre corpus de test. Les entités étaient étiquetées selon leur catégorie réelle. Les résultats indiquent que la justesse des affectations (colonne *accuracy* du tableau 18) se situe globalement à environ 76.7% d'efficacité. Nous estimons que ce résultat est bon

[16] Le terme "BCE" compte pour plus de la moitié des noms d'organisation. Nous avons obtenus pour les autres noms d'organisation une précision de 95.5% et un rappel de 68.3%.

[17] La plupart des adresses sont des URLs qui contiennent une référence au site web de BCE (www.bce.ca).

considérant la simplicité de nos règles. Pour certaines entités, il est parfois difficile de déterminer leur rôle à partir du contenu d'une seule phrase. Par exemple, des coréférences rendent parfois difficile l'interprétation de certaines phrases (par ex. "It will be at 17:00."). Toutefois, la plupart des erreurs sont liées à des descriptions de rôles que nous n'avions pas anticipées lors de la conception de notre base de règles.

5.8 Travaux pertinents en adaptation CBR

Afin de situer notre approche par rapport aux techniques d'adaptation en CBR structurel, mentionnons que l'on retrouve quatre grandes approches pour l'adaptation de cas :

- les approches *substitutionnelles* qui modifient des paramètres d'une solution,
- les approches *transformationnelles* qui modifient la structure d'une solution,
- les approches *génératives* qui recréent une toute nouvelle solution à partir d'un problème, et
- les approches *compositionnelles* qui fusionnent le contenu de plusieurs solutions en une seule.

Ainsi nos travaux constituent une amorce d'adaptation et offrent une composante substitutionnelle et une composante transformationnelle pour la réutilisation de solutions antécédentes. L'identification du contenu circonstancié permet de modifier certains passages, ce qui équivaut à la modification paramétrique de solution. De plus, l'identification des passages optionnels mène à l'abandon de certains des énoncés, ce qui résulte en une modification de la structure de la solution. Comme notre schéma de réutilisation repose sur un seul message, notre approche n'est pas compositionnelle et nous n'effectuons pas de reformulation complète telle que préconisée par les approches génératives. Puisque les décisions sur l'élagage et la substitution de passages sont supervisées par l'usager du système, notre approche met donc l'accent sur la réutilisation d'une solution tandis que la révision de cette solution est laissée au soin de l'usager.

5.9 Conclusion

Dans cette section, nous avons présenté une approche pour réutiliser des réponses antécédentes afin de répondre à de nouvelles requêtes. Nous avons proposé deux approches pour sélectionner les portions pertinentes des messages antécédents. L'une des approches, par condensation, tend à sous estimer la pertinence des passages et s'avère donc conservatrice dans ses choix de phrases à retenir. Nous recommandons plutôt l'utilisation de l'approche par regroupement qui offre une meilleure performance en termes de précision et de rappel. L'approche par condensation pourrait s'avèrer utile pour les applications construites à partir d'un large corpus de messages. Par la suite, nous avons exploré l'utilisation de techniques d'extraction d'entités nommées pour déterminer les portions variables des réponses. L'efficacité de cette étape repose principalement sur la disponibilité d'outils pour repérer les entités. Les résultats que nous avons obtenus indiquent que l'identification de rôle est assez facile à mettre en oeuvre lorsque les entités sont bien déterminées au préalable.

Nous avons présenté une première tentative au niveau de l'adaptation CBR textuel et plusieurs voies se dégagent pour des travaux futurs. Il nous semble que l'idée de patrons dynamiques est une métaphore suffisamment générique pour s'appliquer à d'autres contextes que la réponse au courriel. Elle permet de préserver la forme narrative des solutions et permet d'éviter les approches génératives qui, dans un contexte textuel, sont difficiles à réaliser. Il faudrait toutefois valider cette approche sur d'autres corpus pour mieux cerner ses avantages et ses limitations.

Nous avons choisi, pour ces travaux, de cerner les problèmes reliés à la réutilisation d'un seul cas. Toutefois, une approche compositionnelle qui prend en compte plusieurs cas offrirait sûrement un meilleur recouvrement des thèmes présents dans les requêtes. L'utilisation de plusieurs cas permettrait de mettre en oeuvre des systèmes de votes pour la sélection des passages. Toujours dans un contexte de réutilisation multi-cas, l'identification des portions variables pourrait s'effectuer par la comparaison des énoncés à différents niveaux de granularité et pourrait donner lieu à l'utilisation de techniques syntaxiques pour identifier des passages autres que les entités

nommées. Ceci permettrait de surpasser la principale limitation de la démarche que nous avons suivie, à savoir une identification manuelle des rôles du domaine. Finalement, l'approche par regroupement nous donne deux groupes qui confirment ou infirment la pertinence des énoncés. La même intuition pourrait être utilisée pour construire un tableau et obtenir des règles qui permettraient de catégoriser les phrases en fonction des mots de requêtes. Cette approche devrait pallier le déséquilibre causé par le faible nombre d'exemples positifs qui supportent un énoncé de réponse.

Chapitre 6 . Conclusion et perspectives futures

Dans ce dernier chapitre, nous résumons les principaux résultats présentés dans cette thèse et nous donnons un aperçu de pistes que nous jugeons prometteuses pour la poursuite de travaux dans cette voie de recherche.

6.1 Conclusion

Dans cette thèse, les travaux que nous avons présentés se situent à deux niveaux :

- Une perspective *tâche* qui consiste à répondre à de nouvelles requêtes étant donné un corpus de messages antécédents que l'on peut exploiter ; et

- Une perspective *résolution de problème* qui a mené à l'étude de techniques de raisonnement à base de cas textuels et à leur application.

Afin de mettre en correspondance ces deux perspectives, nous avons proposé des extensions aux approches actuelles du CBR. Nous avons évalué leur applicabilité à un corpus de messages courriels du domaine du service aux investisseurs. Les extensions proposées se situent à plusieurs niveaux. Pour la phase de recherche, l'idée principale qui se dégage dans nos travaux est que l'exploitation de relations entre les descriptions des problèmes et de leurs solutions correspondantes peuvent contribuer à améliorer la performance de ce processus CBR. Cette approche s'avère particulièrement avantageuse lorsque les similitudes entre les solutions sont plus grandes qu'entre les problèmes. Des associations de mots, capturées à l'aide de modèles de cooccurrences et de traduction, permettent d'incorporer les solutions dans la phase de recherche et ainsi de tirer profit de leur cohésion. Par expérimentation, nous avons déterminé que l'utilisation de ces associations permet d'augmenter la précision de la phase de recherche de notre module CBR. Nous avons également proposé un cadre d'évaluation qui permet de prédire, et ce lors de la conception du système, les gains potentiels de performance.

Nous nous sommes par la suite attaqués au problème de la réutilisation de solutions textuelles. En l'absence de modèles théoriques pour aborder cette

problématique, la tâche de réponse au courrier électronique constitue une métaphore intéressante pour une approche pragmatique car elle exige la modification du contenu circonstancié des messages. A partir d'observations sur notre corpus de messages, nous avons formulé ce problème comme la sélection des énoncés pertinents au contexte d'une nouvelle requête et la révision d'informations spécifiques d'une réponse. La sélection permet de modifier la structure d'une réponse pour en préserver la pertinence. Deux approches ont été comparées pour aider l'usager dans ce processus de sélection. L'identification des spécificités d'un message à partir de ses entités nommées permet à l'usager de valider la véracité des énoncés. Une approche d'extraction d'information a été préconisée pour mettre en oeuvre cette étape.

6.1.1 Avantages de l'approche

Tel que souligné dans les chapitres précédents de cette thèse, notre approche de réponse au courrier électronique comporte plusieurs avantages. La réutilisation de messages antécédents est naturelle car ces messages sont habituellement disponibles sur les logiciels de courriels d'une entreprise et peuvent être facilement accumulés durant les opérations quotidiennes. Le schéma de réutilisation que nous préconisons respecte l'utilisation courante que font les gens des logiciels clients de courrier électronique. Ce choix permet d'y intégrer aisément les composantes de sélection de messages antécédents et de suggestion de modifications (dans notre cas, l'intégration s'effectue à la fenêtre de composition de message). Cette intégration n'exige pas la modification substantielle de l'environnement de travail d'un préposé à la clientèle et n'entraîne pas de chambardement dans le processus d'affaire de l'entreprise.

Le raisonnement à base de cas a longtemps été proposé comme une alternative avantageuse aux systèmes à base de règles car il requiert peu de connaissances du domaine. Fidèle à cette philosophie, la conception d'un module CBR textuel selon notre approche ne requiert pas de modélisation intensive du cadre applicatif. Nous avons résisté à la tentation d'entreprendre une démarche d'acquisition et de modélisation de connaissances du service aux investisseurs pour préserver la généricité de l'approche.

Ainsi les modèles de traduction, les modèles de cooccurrences et le regroupement de cas similaires sont obtenus par le traitement du contenu de la base de cas et se révèlent de précieuses ressources pour alimenter les phases de notre approche CBR. Leur construction est menée lors de la conception initiale du système, ce qui limite le temps requis lors du traitement de nouveaux problèmes.

Les extensions que nous avons apportées au cycle CBR s'inspirent toutes du fait qu'un cas textuel est constitué d'un problème et d'une solution et que les liens entre ces deux composantes peuvent être exploités. Qu'ils prennent la forme d'associations de mots, de support entre les descriptions ou de relations avec les entités nommées, ces liens rapprochent les travaux en CBR textuel d'une vision plus classique du raisonnement où les solutions jouent un rôle tout aussi important que la description des problèmes. Nous ne prétendons pas que notre approche en est une de résolution de problème puisque nos extensions s'appuient principalement sur des techniques NLP pour repérer des informations dans des textes (c.-à-d. la recherche d'information, le résumé de texte et l'extraction d'information). Néanmoins, nous avons initié des travaux qui vont au-delà de la recherche sur des descriptions de problèmes et nous repoussons ainsi les frontières actuelles du CBR textuel.

6.1.2 Limitations de l'approche

Les techniques statistiques que nous proposons permettent d'éviter une modélisation cognitive du cadre applicatif. Toutefois pour des domaines présentant moins de cohésion que celui du service aux investisseurs, l'ajout de connaissances du domaine peut s'avérer nécessaire pour atteindre une performance satisfaisante du système.

Nous avons positionné nos travaux dans la phase de conception initiale d'un module CBR. Ainsi nous avons évacué tous les aspects portant sur l'évolution du système lors des opérations quotidiennes et sur sa maintenance pour préserver sa performance. Ces aspects peuvent s'avérer importants sur deux plans. Pour des domaines dynamiques comme celui du service aux investisseurs, la nature des requêtes

change avec le temps. Notre approche ne tient pas compte non plus de l'évolution chronologique des requêtes, de la modification des interventions des analystes et de la mise en veilleuse de certaines formulations de réponses. De plus, certains sujets de requête sont épisodiques (par ex. des requêtes suite à l'acquisition d'une nouvelle filiale). Toutefois nous ne proposons pas de techniques pour estimer la pérennité d'un thème ou d'un cas.

Dans notre utilisation des modèles de traduction, nous travaillons avec des séquences de lemmes (les racines morphologiques) plutôt qu'avec les phrases originales. De plus, plusieurs termes sont retirés des séquences en raison de leur faible fréquence ou de leur catégorie lexicale. Ces choix ont peu d'incidence sur un modèle IBM1 car celui-ci ne prend pas en compte l'ordre ou le nombre de termes. Toutefois, l'utilisation de séquences limite l'application d'autres modèles IBM (ou tout autre modèle statistique de traduction) en raison de leur dépendance à la longueur des phrases et à la position des mots dans les phrases.

Finalement, la réutilisation d'informations instanciées exigent un travail manuel pour identifier les rôles et les règles qui identifient des valeurs de substitution. Nos règles dépendent en partie de la tâche de réponse (par ex. l'émetteur et le récepteur d'un message) et en partie du domaine applicatif (par ex. une référence temporelle comme *le quatrième trimestre*). Des approches adaptatives d'extraction d'information pourraient permettre de capturer certaines des règles régissant l'identification et la modification des messages. Nous n'avons pas exploré cette piste de recherche dans nos travaux.

6.2 Travaux futurs

Nous proposons dans cette dernière section quelques voies de recherche à envisager pour des travaux futurs. Dans un premier temps, nous identifions d'autres options pour guider le processus de réponse au courrier électronique. Par la suite, des pistes plus techniques pour la recherche en CBR textuel sont proposées.

6.2.1 Extensions au processus de réponses au courriel

Dans nos travaux, nous formulons le problème de réponse comme la synthèse de tout un bloc de réponse étant donné un texte de requête. Toutefois, plusieurs personnes préfèrent rédiger une réponse en découpant le message de requête en segments individuels et en associant une portion de réponse à chacun des segments. L'exemple de la Figure 44 illustre cette approche par le découpage de la requête en différents thèmes (thème 1, thème 2) et l'adjonction de fragments de réponse. Cette approche présente des défis lors de la construction du module de réponse pour déterminer la granularité de fragmentation des solutions, la structure de la base de cas (par ex. la représentation hiérarchique des cas (Smyth & al. 2001) et la mise en correspondance des fragments de requêtes et de réponses.

Dan,

> Hello, can you tell me when you will be releasing
> your next earnings report.

The release date for the earnings is on 26 January 2000.

> Also when your fiscal year ends.

The year ended on 31 december 1999.
Thank you for your interest in our corporation.

Figure 45 : Réponse par découpage de la requête

Une autre piste serait de faire le pont entre notre approche et celles basées sur des formulaires pour la rédaction de requête. Les formulaires structurent partiellement le contenu des requêtes, ce qui peut simplifier par la suite les estimations de similarité. Toutefois aucune technique n'est proposée pour réutiliser les réponses antécédentes. De nouveaux schémas pourraient être proposés pour exploiter la structuration des requêtes dans la sélection des passages pertinents et la suggestion de modifications.

Dans notre schéma de réponse, l'usager invoque directement le module de réponse lorsque des recommandations sont désirées. Toutefois des formes d'interaction plus riches devraient être étudiées. Des travaux récents en CBR sur les systèmes à initiative mixte (Aha 2002, Aha 2003) pourraient constituer un bon point de départ. Finalement le

déplacement du module de réponse du côté du serveur de courrier électronique (comme pour les logiciels pour filtrer les pourriels - *spams*) pourraient entraîner des choix technologiques différents des nôtres.

6.2.2 Extensions au cycle de raisonnement à base de cas textuels

Le CBR textuel est relativement récent et plusieurs avenues de recherche méritent d'être explorées afin de mieux définir ses frontières et de quantifier ses approches. Contrairement aux modèles structurel et conversationnel, il n'existe pas de vision unifiée du modèle CBR textuel mais plutôt un ensemble de travaux disparates. Une telle vision unifiée permettrait de mieux regrouper les différents travaux et de déterminer les principales lacunes du modèle. Le domaine présente également certaines déficiences d'un point de vue méthodologique. Tel que mentionné dans notre revue de la littérature du domaine de recherche (au chapitre 2), des expérimentations sont nécessaires pour obtenir une meilleure caractérisation des facteurs influençant le choix d'une approche particulière.

Nous présentons dans les prochains paragraphes des points qui, à notre connaissance, n'ont pas fait l'objet de recherche en CBR textuel.

Construction de cas ("authoring")

C'est probablement à ce niveau que les contributions les plus significatives peuvent être apportées. La littérature CBR actuelle propose des représentations de cas comportant des mots-clés, des termes complexes, des catégories et des paires attributs-valeurs. Toutefois, pour une application particulière, il n'existe pas de méthodologie pour déterminer le niveau de représentation adéquat ni de critères pour faire ce choix.

Cette lacune risque de s'avérer encore plus importante avec l'avènement du web sémantique. Nous faisons référence plus particulièrement à la construction de bases de cas à partir de documents semi-structurés ayant fait l'objet d'annotations. Des approches seront nécessaires pour déterminer l'importance relative du contenu textuel par rapport à sa contextualisation sémantique et pour les mettre à contribution dans le processus de

résolution de problème. Pour les documents non-structurés, les techniques d'extraction d'information et de fouille de textes (*text mining*) pourraient jouer un rôle important dans le processus de structuration de cas.

Recherche sur la base de cas

Bien que ce thème ait largement été étudié, certaines idées restent à explorer. Premièrement, plusieurs des méthodes actuelles préconisent une représentation de type "mots en vrac" (*bag of words*) ne tenant pas compte de l'ordre des termes d'indexation. Or, pour des applications de résolution de problèmes, la préservation d'informations linguistiques comme la négation de propositions ou des séquences particulières de mots peut jouer un rôle sur le choix des solutions proposées.

Il serait intéressant d'étudier la synergie entre les approches sémantiques découlant d'une modélisation du domaine et celles qui s'appuient sur des modèles statistiques de langue. Par exemple, l'enrichissement de ressources linguistiques (telles que WordNet) par des techniques de regroupement (*clustering*) de mots pourrait favoriser la définition de métriques sémantiques pour de nouveaux domaines applicatifs.

Au chapitre 3, nous comparons trois mesures de similarité : le *tf*idf*, la similarité de solutions par modèle de cooccurrences et l'utilité d'une solution par un modèle de traduction. Comment combiner ces trois mesures en une seule estimation de similarité ? On retrouve dans la littérature CBR des travaux qui agglomèrent plusieurs métriques de similarité (locale ou globale) à l'aide de sommes pondérés et de formulation multi-attributs. Toutefois, le domaine de l'aide à la décision multicritère propose un nombre considérable de méthodes de surclassement (par ex. Prométhée et Electre) qui permettraient de combiner plusieurs métriques de similarité mesurées sur des échelles non commensurables.

Nous avons mené nos travaux sur un seul corpus de message. Une comparaison plus exhaustive devrait être effectuée afin d'obtenir une estimation plus claire des améliorations qui peuvent être atteintes pour différentes bases de cas dont les caractéristiques varient. La taille de la base de cas est d'une importance primordiale et

nous nous attendons à ce que des listes d'associations plus représentatives soient générées à partir d'un corpus plus large. Il serait intéressant de mener des expérimentations sur des textes qui varient selon leur taille, leur dépendance au domaine et leur niveau de structuration.

Réutilisation et adaptation

Hormis l'approche de réutilisation que nous proposons dans cette thèse, l'adaptation de solutions textuelles est une tâche ardue pour laquelle on ne retrouve ni paradigme, ni modèle, ni approche. La communauté CBR y prête peu attention en raison du manque d'incitation économique et des difficultés techniques qu'elle présente. Mais le problème demeure important d'un point de vue académique et pour des applications comme le service personnalisé à la clientèle.

En partant des travaux que nous avons accomplis, nous envisageons de poursuivre nos recherches sur les trois thèmes suivants :

- Comment sélectionner des énoncés provenant de plusieurs cas et les fusionner au sein d'une nouvelle solution ? Ceci correspondrait à une approche d'adaptation de type *compositionnelle*. Ces recherches pourraient s'appuyer sur des travaux en résumé de texte à partir de documents multiples, mais elles devraient être étendues à des formulations qui prennent en compte les descriptions de problèmes (i.e. de type *query-biased*).

- Comment exploiter les travaux sur l'acquisition de connaissances d'adaptation en CBR structurel (Leake & al. 96, Kinley 2001, Hanney & Keane 1997, Jarmulak *et al*. 2001) pour apprendre comment substituer les valeurs d'entités nommées dans les descriptions de solutions textuelles ?

- Comment ne plus limiter les portions variables à des entités nommées ? La comparaison des descriptions de cas dans la base pourrait permettre de déterminer les séquences de termes qui varient pour des problèmes différents et qui méritent d'être révisées.

Maintenance

La maintenance de bases de cas est habituellement guidée par des critères tels que la redondance entre cas ou le recouvrement des solutions (*coverage* et *reachability*). Ces critères se prêtent bien au CBR structurel car les cas y sont structurés sans ambiguïté, les attributs et leur domaine étant définis à partir d'une description du domaine. Toutefois, pour appliquer ces approches de maintenance à des cas textuels, il serait important d'étendre ces critères pour tenir compte de la synonymie et la polysémie des différents termes indexant les cas. La modélisation de l'évolution chronologique de la base de cas et la détection d'imperfection (par ex. la redondance, l'incohérence) dans les descriptions textuelles pourraient jouer un rôle important pour des domaines dynamiques comme celui du service aux investisseurs.

6.2.3 Autres applications potentielles

Les extensions que nos proposons au cycle de raisonnement à base de cas textuel peuvent être appliquées à d'autres contextes que celui de la réponse au courrier électronique. Le thème commun à ces applications dont la tâche en est une de rédaction est la forme narrative des solutions qui joue un rôle important. On pense entre autre aux applications suivantes :

- La production de foires aux questions (*FAQ*s) à partir d'un corpus de messages. Ces FAQs pourraient être produits en identifiant les réponses fréquentes et les généralisant par le retrait de phrases superflues (non pertinentes et accessoires) et de passages spécifiques;

- La rédaction de curriculum vitae et leur mise en correspondance avec des descriptions d'emploi par l'identification d'associations entre énoncés ;

- La rédaction de fiches et de rapports en milieu de travail. Par exemple, plusieurs domaines dont le commandement et contrôle militaire sont menées à partir de listes d'activités (*checklists*) pour guider la conduite des opérations. Une rédaction plus dynamique de ces fiches favoriserait une meilleure

description des tâches à accomplir et pourrait les rendre plus spécifiques aux différentes situations qui sont confrontées.

Finalement, des domaines misant sur la prestation de service par des modes électroniques, comme la gestion de connaissance ou le gouvernement en ligne (*e-governement*), pourraient bénéficier de l'approche proposée dans nos travaux.

Bibliographie

Aamodt A., Plaza E., 1994. Case-base reasoning : foundational issues, methodological variations, and system approaches, *AI Communications*, vol. 7, no. 1, pp. 39-59.

Aha D.W., Breslow L.A., Muñoz-Avila H, 2001. Conversational case-based reasoning, *Applied Intelligence*, vol 14, pp. 9-32.

Aha D. W., Breslow L. A., 1997. Refining conversational case libraries, *Proceedings of the Second International Conference on Case-Based Reasoning (ICCBR-97)*, Springer-Verlag, pp. 267-278.

Aleven V., Ashley K. D., 1996. How Different is Different? Arguing about the Significance of Similarities and Differences, *Proceedings of the Third European Workshop on Case-Based Reasoning (EWCBR-96)*, Lecture Notes in Artificial Intelligence 1168, Springer Verlag, pp. 1-15.

Bélanger L., 2003. *Le traitement automatisé des courriels pour les services aux investisseurs: une approche par la question-réponse*, rapport interne, Département d'informatique et de recherche opérationnelle, Université de Montréal.

Bergmann R., Breen S., Göker M., Manago M., Wess S., 1998. *Developing Industrial Case-Based Reasoning Applications: The INRECA Methodology*, Lecture Notes in Artificial Intelligence 1612, Springer, Berlin.

Branting L. K., Lester J., 1996. Justification Structures for Document Reuse, *Proceedings of the Third European Workshop on Case-Based Reasoning (EWCBR-96)*, Lecture Notes in Artificial Intelligence 1168, Springer Verlag, pp. 76-90.

Brown P., Cocke J., Della Pietra S., Della Pietra V., Jelinek F., Mercer R., Roossin P., 1990. A Statistical Approach to Machine Translation, *Computational Linguistics*, vol 16, no. 2, pp. 79-85.

Brown P. F., Della Pietra S. A., Della Pietra V. J., Mercer R. L., 1993. The mathematics of statistical machine translation: Parameter estimation, *Computational Linguistics*, vol 19, no. 2, pp. 263-311.

Brüninghaus S., Ashley K. D, 2001. The Role of Information Extraction for Textual CBR, *Proceedings of the Fourth International Conference on Case-Based Reasoning (ICCBR-01)*, Lecture Notes in Artificial Intelligence 2080, Springer Verlag, pp. 74-89.

Brüninghaus S., Ashley K. D., 1999. Bootstrapping Case Base Development with Annotated Case Summaries, *Proceedings of the Third International Conference on Case-Based Reasoning (ICCBR-99)*, Lecture Note in Computer Science 1650, Springer Verlag, pp. 59-73.

Brüninghaus S., Ashley K. D., 1997. Finding Factors: Learning to Classify Case Opinions Under Abstract Fact Categories, *Proceedings of the Sixth International Conference on Artificial Intelligence and Law (ICAIL-97)*, pp. 123-131.

Buckley C., 1985. *Implementation of the SMART Information Retrieval System*, Rapport technique 85-685, Université Cornell.

Burke R.; Hammond K.; Kulyukin V.; Lytinen S.; Tomuro N.; and Schoenberg S., 1997. *Question Answering from Frequently-Asked Question Files: Experiences with the FAQ Finder System*, Technical Report TR-97-05, Department of Computer Science, University of Chicago.

Burke R., Hammond K., Kozlovsky J., 1995. Knowledge-based Information Retrieval for Semi-Structured Text, *Working Notes from AAAI Fall Symposium on AI Applications in Knowledge Navigation and Retrieval*, AAAI, pp. 19-24.

Cardie C., 1993. Using Decision Trees to Improve Case-Based Learning, *Proceedings of the Tenth International Conference on Machine Learning*, pp. 25-32.

Cardie C., 1996. Automating Feature Set Selection for Case-Based Learning of Linguistic Knowledge, *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pp. 113-126.

Cheetham W., 2003. Lessons Learned using CBR for Customer Support, *Proceedings of the 16th Florida Artificial Intelligence Research Symposium (FLAIRS 2003) Conference*, Ste-Augustine, Florida, pp. 114-118.

Ciravegna F., Basili R., Gaizauskas R., 2000. *Proceedings of the Workshop on Machine Learning for Information Extraction*, http://www.isi.edu/~muslea/ .

Ciravegna F., Kushmerick N., Mooney R., Muslea I.; 2001. *Proceedings of IJCAI-2001 Workshop on Adaptive Text Extraction and Mining*, http://www.smi.ucd.ie/ATEM2001/ .

Cooper E., 1996, Improving FAQfinder's Performance: Setting Parameters by Genetic Programming, *AAAI Spring Symposium on Machine Learning in Information Access*, http://www2.parc.com/istl/projects/mlia/papers/cooper.ps .

Cowie J., Lehnert W., 1996. Information Extraction, Communications of the ACM, vol. 39, no 1, pp. 80-91.

Croft B., Callan J., Lafferty J. (éditeurs), 2001. *Proceedings of Workshop on Language Modeling and Information Retrieval*, Carnegie Mellon University, http://la.lti.cs.cmu.edu/callan/Workshops/lmir01/ .

Cunningham H., Maynard D., Bontcheva K., Tablan V., 2002. GATE: A Framework and Graphical Development Environment for Robust NLP Tools and Applications, *Proceedings of the 40th Anniversary Meeting of the Association for Computational Linguistics (ACL'02)*, pp. 168-175.

Daelemans W., Zavrel J., van der Sloot K., van den Bosch A., 2000. *TiMBL: Tilburg Memory Based Learner - version 4.0 - Reference Guide*, ILK Technical Report ILK01-04.

Daniels J., Risland E., 1998. Locating Passages Using a Case-Base of Excerpts, *Proceedings of the Seventh International Conference on Information and Knowledge Management (CIKM '98)*, pp. 38-44.

Daniels J., 1996. *Retrieval of Passages for Information Reduction*, Thèse de doctorat, Université du Massachusetts, Amherst.

Deerwester S., Dumais S. T., Furnas G. W., Landauer T., Harshman R., 1990. Indexing by Latent Semantic Analysis, *Journal of the American Society for Information Science*, vol. 41, no. 6, pp. 391–407.

Diaz-Agudo B., Gonzalez-Calero P. A., 2001. Knowledge intensive CBR made affordable, *Proceedings of the Workshop Program at the 4th International Conference on Case-Based Reasoning (ICCBR-01)*, Technical Note AIC-01-003, Navy Center for Applied Research in Artificial Intelligence, Washington, DC, USA.

Dubois J., 2002. *Classification automatique de courrier électronique*, Mémoire de maîtrise, Département d'informatique et de recherche opérationnelle, Université de Montréal.

Fox S., Leake D., 2001. Introspective Reasoning for Index Refinement in Case-Based Reasoning, The Journal of Experimental and Theoretical Artificial Intelligence, vol. 13, no. 1, pp. 63-88.

Fuchs B., Mathon A., Mille A., 2001. Representing CBR knowledge with the Rocade System, *Proceedings of the Workshop Program at the 4th International Conference on Case-Based Reasoning (ICCBR-01)*, Technical Note AIC-01-003, Navy Center for Applied Research in Artificial Intelligence, Washington, DC, USA.

Fuchs B., Lieber J., Mille A., Napoli A., 2000. An algorithm for adaptation in case-based reasoning, *Proceedings of the 14th European Conference on Artificial Intelligence (ECAI'2000)*, Amsterdam, pp. 45-49.

Fuchs B., Lieber J., Mille A., Napoli A., 1999. Vers une théorie unifiée de l'adaptation en raisonnement à partir de cas, *Actes de la Conférence française sur le raisonnement à partir de cas (RàPC-99)*, Palaiseau, France, pp. 77-85.

Gartner Research, 2000. *E-mail Response Management: Perspective*, http://cnscenter.future.co.kr/resource/rsc-center/gartner/email.pdf.

Grefenstette G., 1992. Use of syntactic context to produce term association lists for text retrieval, *Proceedings of the 15th International Conference on Research and Development in Information Retrieval (SIGIR'92)*, Copenhagen, Denmark, pp. 89-97.

Gutwin C., Paynter G.W., Witten I.H., Nevill-Manning C., Frank E. 1999. Improving browsing in digital libraries with keyphrase indexes, *Decision Support Systems*, vol. 27, no. 1-2, pp. 81-104.

Hanney K., Keane M., 1997. The Adaption Knowledge Bottleneck: How to Ease it by Learning from Cases, *Proceedings of the Second International Conference on Case-Based Reasonining (ICCBR-97)*, Springer Verlag, pp. 359-370.

Hofmann T., 1999. Probabilistic Latent Semantic Indexing, *Proceedings of the 22nd International Conference on Research and Development in Information Retrieval (SIGIR'99)*, pp. 50-57.

Jarmulak J., Craw S., Rowe R., 2001. Using Case-Base Data to Learn Adaptation Knowledge for Design, *Proceedings of the 17th International Joint Conference on Artificial Intelligence (IJCAI-01)*, Morgan Kaufmann, pp. 1011-1016.

Jordan B., 2001. Skills-based routing, an industry survey, *CRM Xchange*, http://www.crmxchange.com/whitepapers/iex-industry-survey.html

Kinley A., 2001. *Learning to Improve Case Adaptation*, Thèse de doctorat, Département d'informatique, Indiana University.

Kolodner J., 1993. *Case-Based Reasoning*, Morgan Kaufmann.

Koole G., Pot A., Talim J., 2003. Routing heuristics for multi-skill call centers, *Proceedings of the 2003 Winter Simulation Conference*, New Orleans, pp. 1813-1816.

Koole G., Mandelbaum A., 2002. Queueing models of call centers: An introduction., *Annals of Operations Research*, vol. 113, pp. 41-59.

Kosseim L., Beauregard S., Lapalme G., 2001. Using Information Extraction and Natural Language Generation to Answer E-mail, *Natural Language Processing and Information Systems*, Lecture Notes in Computer Science 1959, Springer-Verlag, pp. 152-163.

Kosseim, L., Lapalme. G., 2001. Critères de sélection d'une approche pour le suivi automatique du courriel, *Actes de la 8ème conférence annuelle sur le Traitement Automatique des Langues Naturelles (TALN-2001)*, Tours, France, pp. 357-371.

Kosseim L., Lapalme G., 1998. Exibum: Un système expérimental d'extraction d'information bilingue, *Actes de la Rencontre Internationale sur l'extraction, le filtrage et le résumé automatiques (RIFRA-98)*, pp. 129-140.

Kusui D., Shimazu H., 2001. Transforming Mailing Lists into Case Bases, *Proceedings of the Fourth International Conference on Case-Based Reasoning (ICCBR-01)*, Lecture Notes in Artificial Intelligence 2080, Springer Verlag, pp. 690-701.

Lamontagne L., Lapalme G., 2003a. Using Statistical Models for the Retrieval of Fully-Textual Cases, *Proceedings of the 16th Florida Artificial Intelligence Research Symposium (FLAIRS-2003)*, AAAI Press, pp.124-128.

Lamontagne L., Lapalme G., 2003b. Applying Case-Based Reasoning to Email Response, *Proceedings of Fifth International Conference on Enterprise Information Systems (ICEIS-03)*, Angers, France, pp. 115-123.

Lamontagne L., Lapalme G., 2002. Raisonnement à base de cas textuels – état de l'art et perspectives, *Revue d'Intelligence Artificielle*, Hermes, Paris, vol. 16, no. 3, pp. 339-366.

Lamontagne L., 2001. Raisonnement à base de cas textuel pour la réponse automatique au courrier électronique, Proposition prédoctoral, Université de Montréal.

Lapalme G., Kosseim L., 2003. Mercure : Toward an automatic e-mail follow-up system, *IEEE Computational Intelligence Bulletin*, vol. 2, no. 1, p. 14-18.

Lenz M., Glintschert A., 1999. On Texts, Cases, and Concepts, *Proceedings of the 5th Biannual German Conference on Knowledge-Based System (XPS-99)*, Lecture Notes in Artificial Intelligence 1570, Springer Verlag, pp. 148-156.

Lenz M., Bartsch-Spörl B., Burkhard H.-D., Wess S. (Eds.), 1998. *Case-Based Reasoning Technology - From Foundations to Applications*, Lecture Notes in Artificial Intelligence 1400, Springer Verlag.

Lenz M., Burkhard H.-D., 1997. CBR for Document Retrieval - The FAllQ Project, *Proceedings of the Second International Conference on Case-Based Reasoning (ICCBR-97)*, Lecture Notes in Artificial Intelligence 1266, Springer Verlag, pp. 84-93.

Leake D. B., Wilson D. C., 1999. Combining CBR with Interactive Knowledge Acquisition, Manipulation and Reuse, *Proceedings of the Third International Conference on Case-Based Reasoning (ICCBR-99)*, Lecture Notes in Artificial Intelligence 1650, Springer Verlag, pp. 203-217.

Leake D. B., Wilson, D. C., 1999b. When experience is wrong, *Proceedings of the Third International Conference on Case-Based Reasoning (ICCBR-99)*, Lecture Notes in Artificial Intelligence 1650, Springer-Verlag, pp. 203-217.

Leake D. B. (éditeur) 1996. *Case-Based Reasoning: Experiences, Lessons, and Future Directions*, AAAI Press/MIT Press, Menlo Park, CA, 1996.

Leake D. B., Kinley, A., and Wilson D., 1996. Acquiring Case Adaptation Knowledge: A Hybrid Approach, *Proceedings of the Thirteenth National Conference on Artificial Intelligence (AAAI-96)*, AAAI Press, Menlo Park, CA, pp. 684-689.

Leake D. B., Smyth B., Yang Q., Wilson D., 2001, Special Issue on Maintaining Case-Based Reasoning Systems, *Computational Intelligence*, Vol. 17, No. 2, 2001.

Lytinen S., Tomuro N., Repede, T. 2000. The Use of WordNet Sense Tagging in FAQFinder, *Proceedings of the Workshop on Artificial Intelligence and Web Search (AAAI-2000)*, Austin.

Macklovitch E., Simard M., Langlais P., 2000, TransSearch: A Free Translation Memory on the World Wide Web, *Proceedings of Second International Conference On Language Resources and Evaluation*, Athènes, Grèce, pp. 1201-1208.

Manco G., Masciari E., Tagarelli A., 2002. A Framework for Adaptive Mail Classification, *Proceedings of 14th Conference on Tools with Artificial Intelligence (ICTAI-2002)*, pp. 387-394.

Manning C., Schütze H., 1999. *Foundations of Statistical Natural Language Processing*, Cambridge, Massachusetts, MIT Press.

Marquez L. P., Rodriguez H., 2000. A Machine Learning Approach to {POS} Tagging, *Machine Learning*, vol. 39, no. 1, pp. 59-91.

McSherry D., 2001. Improving the build quality of CBR systems: the case-authoring challenge, *Proceedings of the Workshop Program at the 4th International Conference on Case-Based Reasoning (ICCBR-01)*, Technical Note AIC-01-003,

Navy Center for Applied Research in Artificial Intelligence, Washington, DC, USA.

Minor M., Staab S. (éditeurs) 2002. *First German Workshop on Experience Management*, Lecture Notes in Informatics P-10, Bonner Köllen Verlag.

Mittal V., Berger A., 2000. Query-relevant summarization using FAQs,. *Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics (ACL-2000)*, Hong Kong, pp. 294-301.

Mullins M., Smyth B., 2001. Visualisation Methods in CBR, *Proceedings of the Workshop Program at the 4th International Conference on Case-Based Reasoning (ICCBR-01)*, Technical Note AIC-01-003, Navy Center for Applied Research in Artificial Intelligence, Washington, DC, USA.

Och F. J., Ney H., 2000. Improved Statistical Alignment Models, *Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics (ACL-2000)*, pp. 440-447.

Racine K., Yang Q., 1997. Maintaining unstructured case bases, *Proceedings of the Second International Conference on Case-Based Reasoning (ICCBR-97)*, Lecture Notes in Artificial Intelligence 1266, Springer Verlag, pp. 553-564.

Reinartz T., Iglezakis I., Roth-Berghofer T., 2001. Review and Restore for Case Base Maintenance, *Computational Intelligence*, vol. 17, no. 2, pp. 214-234.

Richter M., 1998. Introduction, dans *CBR Technology : From Foundations to Applications*, chapitre 1, Springer, Berlin, pp. 1-15.

Riesbeck C., Schank R., 1989. *Inside Case-Based Reasoning*, Erlbaum.

Riloff E, 1996. Automatically Generating Extraction Patterns from Untagged Text, *Proceedings of the Thirteenth National Conference on Artificial Intelligence (AAAI-96)*, pp. 1044-1049.

Roark B., Charniak, E., 1998. Noun-phrase co-occurrence statistics for semi-automatic semantic lexicon construction, *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics*, pp. 1110-1116.

Salton G., McGill M., 1984. *Introduction to Modern Information Retrieval*, McGraw-Hill Book Company.

Segal R., Kephart J., 2000. Incremental Learning in SwiftFile., *Proceedings of the Seventh International Conference on Machine Learning*, pp. 863-870.

Shimazu H., Kusui D., 2001. Detecting Defect Sign Cases, *Proceedings of the Fourth International Conference on Case-Based Reasoning (ICCBR-01)*, Lecture Notes in Artificial Intelligence 2080, Springer Verlag, pp. 611-621.

Smyth B., McClave P., 2001. Similarity vs. Diversity, *Proceedings of the Fourth International Conference on Case-Based Reasoning (ICCBR-01)*, Lecture Notes in Artificial Intelligence 2080, Springer Verlag, pp. 347-361.

Smyth B., Keane M. T., Cunningham P., 2001. Hierarchical Case-Based Reasoning: Integrating Case-Based and Decompositional Problem-Solving Techniques for Plant-Control Software Design, *IEEE Transactions on Knowledge and Data Engineering*, vol. 13, no. 5, pp. 793-812.

Smyth B., McKenna E., 1999. Building Compact Competent Case-bases, *Proceedings of the Third International Conference on Case-based Reasoning (ICCBR-99)*, Springer Verlag, pp. 329-342.

Smyth B., Keane M., 1998. Adaptation-Guided Retrieval : Questioning the similarity assumption in reasoning, *Artificial Intelligence*, vol. 102, no. 2, pp. 249-293.

Smyth B., McKenna E., 1998. A portrait of case competence: Modelling the competence of case--based reasoning systems, *Proceedings of the Fourth European Workshop on Case-Based Reasoning (EWCBR-98)*, Springer Verlag, pp. 208-220.

Smyth B., Keane M. T., 1995. Experiments on Adaptation-Guided Retrieval in a Case-Based Design System, *Proceedings of the First International Conference on Case-Based Reasoning (ICCBR-95)*. Lecture Notes in Artificial Intelligence, Springer-Verlag, pp. 313-324.

Soderland S., 1999. Learning information extraction rules for semi-structured and free text, *Machine Learning*, vol. 34, pp. 233-272.

Turney P.D., 2000. Learning algorithms for keyphrase extraction, *Information Retrieval*, vol. 2, no. 4, pp. 303-336.

Varma A., 2001. Managing Diagnostic Knowledge in Text Cases, *Proceedings of the Fourth International Conference on Case-Based Reasoning (ICCBR'2001)*, Lecture Notes in Artificial Intelligence 2080, Springer-Verlag, pp. 622-633.

Voorhees E. M., 1994. Query expansion using lexical-semantic relations, *Proceedings of 17th International Conference on Research and Development in Information Retrieval (SIGIR-94)*, pp. 61-69.

Wang Y., Yang Q., Zhang Z., 1999. Real-Time Scheduling for Multi-Agent Call Center Automation, *Proceedings of the European Artificial Intelligence Planning Conference '99*. United Kingdom, pp. 187-199.

Ward G., 1994. *Grady Ward's Moby Lexicon*, University of Sheffield. http://www.dcs.shef.ac.uk/research/ilash/Moby/.

Watson I, 1997. *Applying Case-Based Reasoning: Techniques for Enterprise Systems*, Morgan Kaufmann Publishers Inc.

Weber R., Martins A., Barcia R, 1998. On legal texts and cases, *Textual Case-Based Reasoning: Papers from the AAAI-98 Workshop*, Rapport technique WS-98-12, AAAI Press, pp. 40-50.

Weber R., 1998. *Intelligent Jurisprudence Research*, Thèse de doctorat, Université fédérale de Santa Catarina, Brésil.

Wilson D. C., Bradshaw S., 2000, CBR Textuality, *Expert Update*, vol. 3, no. 1, pp. 28-37.

Yang Q., Wu J.; 2000. Keep It Simple: A Case Base Maintenance Policy based on Clustering and Information Theory, *Proceedings of the Canadian AI Conference 2000*, Montréal Canada, pp. 102-114.