RAISONNEMENT À BASE DE CAS TEXTUEL POUR LA RÉPONSE AUTOMATIQUE AU COURRIER ÉLECTRONIQUE

Proposition de projet doctoral

Luc Lamontagne

16 novembre 2001

1.0 Introduction

Ce document décrit la démarche que nous proposons pour aborder le problème de la gestion de réponse automatique au courrier électronique à l'aide de techniques de raisonnement à base de cas textuel.

Nos motivations pour mener cette recherche se situent à deux niveaux. Premièrement, la gestion de réponse au courrier électronique présente d'importants enjeux techniques et commerciaux. D'un point de vue technique, ce problème constitue un défi de taille car il exige des systèmes qui combinent la compréhension ainsi que la génération de textes. D'un point de vue commercial, l'importance croissante du courrier électronique requiert des systèmes facilement déployables qui garantissent de bons niveaux de performance dans le suivi du service à la clientèle.

Deuxièmement, le raisonnement à base de cas (CBR) est l'une des voies les plus prometteuses pour aborder ce problème. La conception d'un système CBR de gestion de réponse s'appuie principalement sur un corpus de messages antécédents, une ressource qui est à la fois représentative du domaine de discours et de divers problèmes résolus par voie de courriel. De plus, contrairement aux approches de catégorisation de texte, le schéma de raisonnement "recherche et adapte" préconisé par les techniques CBR permet de concilier l'analyse des questions et la synthèse des réponses à ces questions. Récemment quelques travaux ont proposé des approches pour permettre l'application du CBR traditionnel à des cas provenant de documents textuels. Nos recherches viseront à enrichir ces approches et à proposer de nouvelles avenues pour tenir compte des spécificités de notre application. Les prochains paragraphes décrivent plus précisément le contexte de nos travaux de recherche.

1.1 Gestion de la réponse au courrier électronique

Selon Forrester Research, on prévoit que douze milliards de messages électroniques seront échangés cette année. Ce large volume donne une indication de l'utilisation accrue de ce mode de communication au sein des entreprises. Toujours selon Forrester Research, un sondage indique que plus de 70% des entreprises jugent que le courrier électronique est important ou très important pour leur stratégie de marketing et de vente.

De plus, pour pallier l'augmentation du volume de requêtes et du temps de réponse dans les centres d'appels des entreprises, on anticipe qu'une large part des communications actuellement effectuées par voie téléphonique seront redirigées vers des moyens électroniques comme le courriel et les services de "chat". Toutefois, le Gartner Group [Gar00] estime que seulement 10% des entreprises sont préparées pour gérer adéquatement le volume de courrier découlant de l'interaction avec leur clientèle. L'insertion de nouvelles technologies d'information est donc préconisée pour faire face à ce volume de messages tout en maintenant la qualité du service à la clientèle.

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" nécessitent des techniques de routage (re-direction vers les préposés à la clientèle), de "queueing" (regroupement de messages en file prioritaire) et de réponse automatique. Les premiers systèmes de gestion de réponses sont apparus vers 1997. Actuellement, le domaine est très volatile avec un grand nombre de compagnies ayant fait l'objet d'acquisition au cours des derniers mois (Inference, Aptex, EchoMail, Genesys…). On prévoit que ce secteur d'activité va croître en 2002 pour atteindre des revenus de $210 millions répartis sur une clientèle de 25,000 entreprises. Actuellement, moins de 10% de ce marché potentiel a été pénétré.

L'utilisation de ces systèmes offre de nombreux avantages. Elle permet 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, elle entraîne une augmentation de la productivité et une diminution de la redondance en facilitant le traitement des messages répétitifs, la distribution du travail entre les préposés, et un meilleur suivi des messages. 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 (élicitation des règles d'affaires, des canevas de messages, etc.).

1.2 Adéquation du raisonnement à base de cas

Les techniques de raisonnement à base de cas sont déjà largement utilisées pour des applications de 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'application de ces systèmes au service à la clientèle par courrier électronique est donc naturelle. Plusieurs questions étant répétitives, l'utilisation de réponses précédentes pour répondre à de nouvelles questions semble tout indiquée. De plus, ces systèmes offrent des capacités d`apprentissage "online" car une nouvelle paire de question-réponse peut être immédiatement réutilisée par le système pour les requêtes subséquentes. Cette approche est plus accessible que la conception de systèmes experts, car la définition d'un ensemble de règles pour guider le choix des réponses est une tâche ardue. En effet, il est difficile de prévoir toutes les questions possibles et de développer un modèle du domaine pour y répondre.

A date, aucune publication sur la gestion de réponse au courrier électronique n'a été répertoriée dans la littérature CBR scientifique. Par ailleurs, il existe dans le domaine commercial deux systèmes de gestion de réponse qui utilisent des techniques CBR: eGain et FirePond.[1] Ces outils utilisent un modèle de raisonnement CBR conversationnel (modèle présenté à la section 2.2 de ce document). De plus, plusieurs des compagnies oeuvrant dans ce domaine utilisent de documents de type FAQ, i.e. des questions fréquentes auxquelles on associe une réponse standard générée manuellement. Malheureusement leur documentation technique ne fournit pas de précision sur les mécanismes de gestion des FAQ. Par ailleurs, certains 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.

1.3 Domaine du service aux investisseurs

Pour notre étude, nous nous intéressons particulièrement aux messages électroniques échangés dans le cadre du service aux investisseurs ("investor relations"). Le service aux investisseurs est le processus par lequel une compagnie communique avec ses investisseurs. Ceci comprend, entre autres, la dissémination des nouvelles corporatives et l'assistance aux investisseurs. Ce service est important car les investisseurs professionnels estiment que la qualité de l'information et des

[1] La technologie de eGain découle de leur acquisition de Inference Corp alors que celle de FirePond fait suite à leur acquisition de Brightware. Il est important de mentionner que Brightware était à l'origine un "spin-off" de Inference Corporation. Inference était le concepteur du principal outil commercial CBR, k-commerce (anciennement CBR3). Ayant eu accès à la technologie de Inference, des composantes de l'outil de Inference ont été intégrées au logiciel Art-Enterprise de Brigthware.

réponses qui leur sont fournies compte pour 10% dans leurs décisions d'investissement. Pour obtenir leur information, les investisseurs préfèrent les technologies d'information, et plus particulièrement les technologies web. Par exemple, un sondage indique que 66% des investisseurs préfèrent consulter un rapport annuel en ligne plutôt qu'une version papier. En réponse à cette demande, les entreprises offrent de plus en plus de services en ligne tels que le "webcasting", le "fax-on-demand" ou le courrier électronique. Ainsi, la grande majorité des entreprises rendent disponible une multitude d'information sur leur site web tels que la valeur de l'action, les rapports financiers, des annonces d'événements corporatifs, etc.

Nous menons nos recherches dans le contexte des opérations du service aux investisseurs de Bell Canada Enterprise (BCE). Le site web de BCE contient une foule d'information sur les aspects financiers de cette corporation et de ses filiales. Si des investisseurs désirent obtenir de plus amples renseignements, BCE met à leur disposition un groupe d'analystes pour répondre à leurs requêtes acheminées par courrier électronique. Actuellement, les analystes de BCE reçoivent un grand nombre de requêtes. Plusieurs de ces questions sont répétitives tandis que d'autres font référence à des informations disponibles sur le site web ou dans des documents tels les rapports annuels.

BCE a donc identifié le besoin de développer un système d'information qui facilitera la gestion et le suivi des messages et qui assistera les analystes dans leurs interactions avec les investisseurs. Notre étude vise à démontrera les avantages d'une approche de raisonnement à base de cas textuel comme solution potentielle à ce problème: nous adapterons des techniques CBR pour exploiter un corpus de messages de BCE dans le but de répondre à de nouvelles requêtes des investisseurs.

1.3.1 Exemples du domaine

Afin d'illustrer notre approche CBR de gestion de réponse, nous présentons quelques exemples simples des messages portant sur la divulgation de résultats financiers. Le tableau I contient des questions-réponses provenant du corpus BCE utilisé comme base de cas pour nos exemples. Ces messages traitent de dates de divulgation de rapports financiers (C1, C2, C4), d'appels conférence (C2, C3) et de fin d'année fiscale (C1,C3).

Tableau I – Base de cas constituée de quelques messages antécédents

Cas	Question	Réponse.
C1	Hello Can you tell me when you will be releasing your next earnings report also when your fiscal year ends Best Regards, Mark Strasse	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…
C2	Dear Investor Relations: I would like to find out when the year end earnings will be? Also is there going to be a conference call? Your shareholder	Dear shareholder, The release date is on January 26th, 2000 after the Board meeting which is scheduled for 10:00. The conference call will be the same day at 17:00. We do not have the particulars for the conference call as of yet. Thank you for your interest in our corporation.
C3	I seem to have misplaced the number for Tuesday's conference call. Could you please	The conference call for the BCE earnings release is on Wednesday January 26, 2000 at 17:00 Eastern

	send me the number?	Time. 1-800-999-9999 between 16:50 and 16:58. Thank you for your interest in our corporation.
C4	When will financials be released?	Good afternoon, The second quarter will be released on the 26 July 2000. Thank you for your interest in our corporation.

Etant donnée la base de cas présenté au tableau I, supposons maintenant que le système reçoive la question suivante:

Question1: Can you tell me when you are reporting next. Thanks, Elliott.

Habituellement, une approche de raisonnement à base de cas s'effectue en deux étapes: la recherche de cas similaires et la réutilisation de ces cas. Dans ce cas précis, le processus de recherche évaluerait que les messages C1, C2, et C4 sont les plus similaires à la *Question1*. Supposons pour les fins de cet exemple que le système ne retiendrait qu'un seul cas: C1.

Etant données les caractéristiques textuelles de notre application, nous avons établi que la réutilisation d'un cas nécessite en priorité l'identification des portions des réponses antécédentes pouvant être utiles[2]. Pour le cas C1, le système déterminerait que la portion traitant de la date de divulgation des profits ("release date for the earnings…") est utile. Ensuite, il établirait que les portions de texte portant sur la fin d'année fiscale ("The year ended…") et sur la date du rapport trimestriel ("26 January 2000") devraient être modifiées ou tout simplement élaguées. Les passages à modifier/élaguer sont représentés par la notation « texte ».

Donc, la réponse proposée par le système CBR pour cette question serait la suivante:

Réponse1: 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…

Précisons que l'exemple précédent est basé sur la sélection d'un seul cas. Par ailleurs, une réponse pourrait être construite à partir de plusieurs messages antécédents. En voici un exemple:

Question2: Hello, I am writing to find out when you are reporting the 2nd quarter earnings and to obtain the number if you are having a conference call. Thank you…

Contrairement à la *Question1*, une réponse composée à partir des C2 et C3 serait, avant les modification et élagage, la suivante:

[2] Les opérations de substitution et d'élagage ne sont pas présentées dans nos exemples afin d'en préserver la clarté. Par ailleurs, nous en discuterons dans la présentation des thèmes de recherche à la section 6 de ce document.

Réponse2: «Dear shareholder»,

The release date is on «January 26th, 2000» after the Board meeting which is scheduled for «10:00».

«The conference call will be the same day at 17 :00. We do not have the particulars for the conference call as of yet.»

The conference call for the BCE earnings release is on «Wednesday January 26, 2000» at «17:00 Eastern time.».

«1-800-999-9999» between «16:50 and 16:38.».

Thank you for your interest in our corporation.

Une analyse préliminaire du corpus de BCE a révélé plusieurs caractéristiques qui rendent complexe l'utilisation traditionnelle des techniques CBR. Une liste de ces caractéristiques est présentée à la section 4 du document. En voici un exemple:

Question3: Hi, Can you tell me when the next three financial earnings reports are due for BCE ? Thanks, Paul

Cette question contient implicitement des questions multiples (les dates de divulgation de chacun des trois prochains trimestres). Idéalement, notre processus de réutilisation de cas serait capable de repérer cet énoncé de questions multiples et de proposer une réponse du type:

Réponse3: Dear «Mr. Strasse»,

«The year ended on 31 december 1999. »

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

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

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…

Les exemples précédents illustrent, en partie, le comportement des étapes de recherche et de réutilisation (adaptation) d'un système CBR. Ils permettent également de noter que le choix d'une base de cas peut entraîner certaines difficultés, notamment:

- les messages comportant plusieurs thèmes sont difficiles à traiter. Une plus fine granularité de cas simplifierait les étapes de recherche et de réutilisation. Par exemple, la *Réponse2* contient deux phrases redondantes portant sur la date de l'appel conférence. Nous pourrions éviter cette redondance en limitant chaque cas de la base à un seul thème;
- la similarité des messages repose sur quelques termes et le choix de la représentation interne influence la qualité de la comparaison des cas. Par exemple, la similarité entre la *Question1* et les cas C1, C2 et C4 est évaluée à partir des termes {when, report, reporting, next, releasing, released, earnings, year, end, find, out, financials};
- il existe des redondances entre les messages du corpus. L'élimination de certains cas permettrait d'améliorer l'efficacité de la recherche tout en limitant la dégradation de la qualité

des réponses. Par exemple, le message C4 est un sous-ensemble du message C1 et pourrait être retiré de la base de cas sans grande conséquence.

Lors de la conception d'un système CBR, ces difficultés sont adressées par le processus d'authoring qui guide la création de la base de cas exploitée par les différentes étapes du raisonnement. Nous reviendrons sur ces aspects à la section 6 lors de la présentation de nos thèmes de recherche.

1.4 Plan du document

Notre document comporte trois parties. La première partie présente une revue de littérature des techniques de raisonnement à base de cas (section 2) et des principaux travaux sur l'exploitation de documents textuels (section 3). La deuxième partie décrit les propriétés d'un sous-ensemble de notre corpus de messages (section 4) ainsi que des options pour l'application du CBR au problème de gestion de réponse (section 5). Finalement, la dernière partie décrit les thèmes que nous aborderons dans nos recherches (section 6), les échéanciers de nos travaux (section 7) et les contributions que nous apporterons au cadre applicatif de la gestion de réponse et au domaine du CBR textuel (section 8).

2.0 Modèles de raisonnement à base de cas

2.1 Principes généraux du CBR

Le raisonnement à base de cas (CBR) est une approche de résolution de problèmes qui utilise des expériences passées pour résoudre de nouveaux problèmes [Lea96]. L'ensemble des expériences forme une base de cas. Typiquement un cas contient au moins deux parties: une description de la 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 (e.g. succès ou échec). Les techniques CBR permettent de produire de nouvelles solutions en extrapolant sur les situations similaires au problème à résoudre.

Les fondements du CBR proviennent de travaux en science cognitive menés par Roger Schank et son équipe de recherche durant les années 80 [Rie89]. Ces travaux, ayant pour but de déterminer le rôle de la mémoire dans le raisonnement humain, 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), contient les descriptions d'expériences passées et de situations stéréotypées. La première implantation de système CBR, développée par Janet Kolodner, est basée sur cette structure [Kol93].

Au début de la dernière décennie, le domaine a connu un regain de popularité. Depuis le milieu des années 90, le CBR s'est révélé une précieuse technique pour la mise en oeuvre d'applications commerciales [Wat98]. Actuellement le CBR est l'une des techniques de l'intelligence artificielle la plus largement répandue. Quelques logiciels commerciaux ont été développés pour mettre en pratique les principes de base du CBR tels que les produits K-Commerce de Inference et Orenge de Empolis Knowledge.

L'approche CBR peut être utilisée dans différents contextes applicatifs incluant les applications suivantes:

- Pour résoudre des problèmes de classification, de diagnostic, de configuration, de design et de planification.
- Pour la conception de système d'aide à la décision ("decision support"). Actuellement les systèmes de "help-desk" sont la principale application commerciale du CBR. Ce sont des systèmes aviseurs qui fournissent aux décideurs/analystes des recommandations sur les actions à entreprendre dans des situations routinières.
- Pour préserver et exploiter la connaissance des entreprises. Le domaine de la gestion de connaissance est actuellement en pleine expansion.
- Pour des applications de type "recherche d'information" sur des domaines restreints. Une application de ce type est l'exploitation de documents structurés tels les "frequently-asked questions" (paires de questions-réponses).

L'approche CBR offre de nombreux avantages. Elle permet d'éviter les problèmes d'acquisition de connaissance ("knowledge bottleneck") qui rendent difficile la construction de bases de connaissances de taille importante. Pour certaines applications, l'approche CBR est plus simple à mettre en oeuvre que les approches basées sur un modèle du domaine (e.g. base de règles). De plus, le CBR est particulièrement bien adapté pour les applications ayant les caractéristiques suivantes:

- la tâche est accomplie par des humains expérimentés dans leur domaine et ces expériences sont disponibles dans une base de données, dans des documents ou chez un expert humain;
- une analyse détaillée du domaine n'est pas nécessaire pour obtenir des solutions satisfaisantes et la tâche n'exige pas une solution optimale ("satisficing solution");
- un modèle du domaine ne peut être élaboré parce que le domaine est mal formalisé (peu de documentation, expert non disponible) ou parce qu'il n'existe pas de principes généraux qui sont éprouvés (e.g. comment investir à la bourse);
- les situations sont répétitives et les solutions sont réutilisables. Ces situations, dites monotones, sont telles que de petites différences dans le problème entraînent de petites différences dans la solution. De plus, une solution valide à un moment le demeure à un autre moment.

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. Un modèle générique a été proposé par Aamodt et Plaza [Aam94] pour décrire les différentes étapes du processus de résolution CBR (Figure 1).

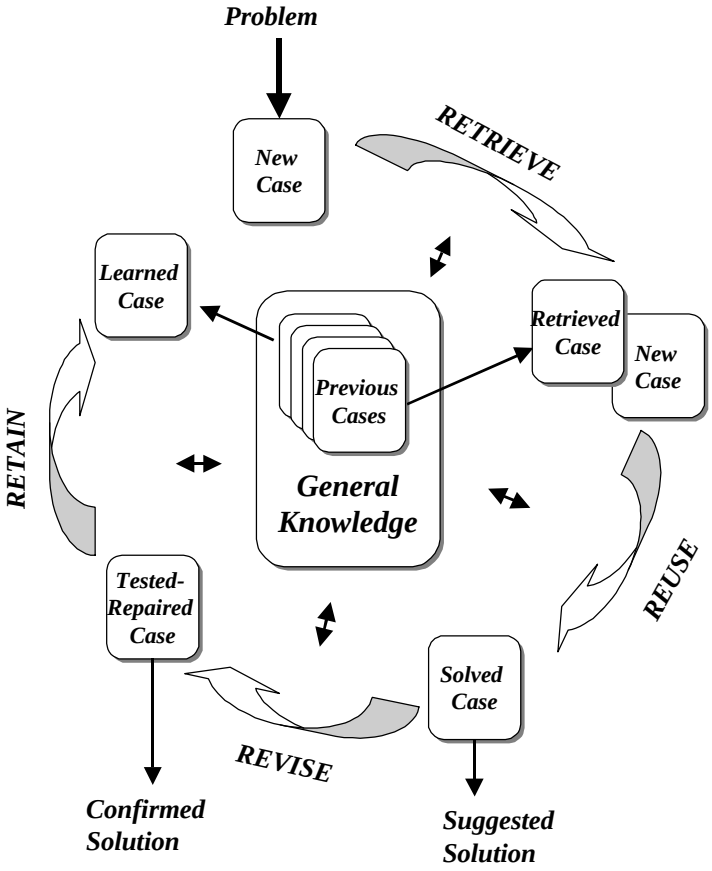


Figure 1 – Modèle de processus CBR de Aamodt et Plaza

Processus

Dans le modèle présenté à la figure 1, on note les principaux processus CBR suivants: la recherche ("retrieval"), l'adaptation ("reuse") et la maintenance ("retain")[3].

- *Recherche ("retrieval")*: cette phase permet de déterminer les cas de la base qui sont les plus similaires au problème à résoudre. La procédure de recherche de similarité est implantée par une approche basée sur les plus proches voisins ("k-nearest-neighbors") ou sur la construction par induction d'une structure de recherche. L'approche des plus proches voisins utilise des métriques de similarité pour mesurer la correspondance entre chaque cas et le problème à résoudre. L'approche par induction génère un arbre qui partitionne les cas selon différents attributs et qui permet de guider le processus de recherche. L'approche par induction équivaut à indexer hiérarchiquement la base de cas. Cette approche est particulièrement avantageuse pour des bases comprenant un grand nombre de cas.

- *Adaptation ("reuse")*: suite à la sélection des 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 2):
    - l'adaptation structurelle: on obtient une nouvelle solution en modifiant des solutions antécédentes et les réorientant afin de satisfaire le nouveau problème.

    - l'adaptation dérivationnelle: pour chaque cas passé, on garde une trace indiquant comment la solution a été générée. Pour un nouveau problème, une nouvelle solution est générée en appliquant une de ces suites d'étapes.

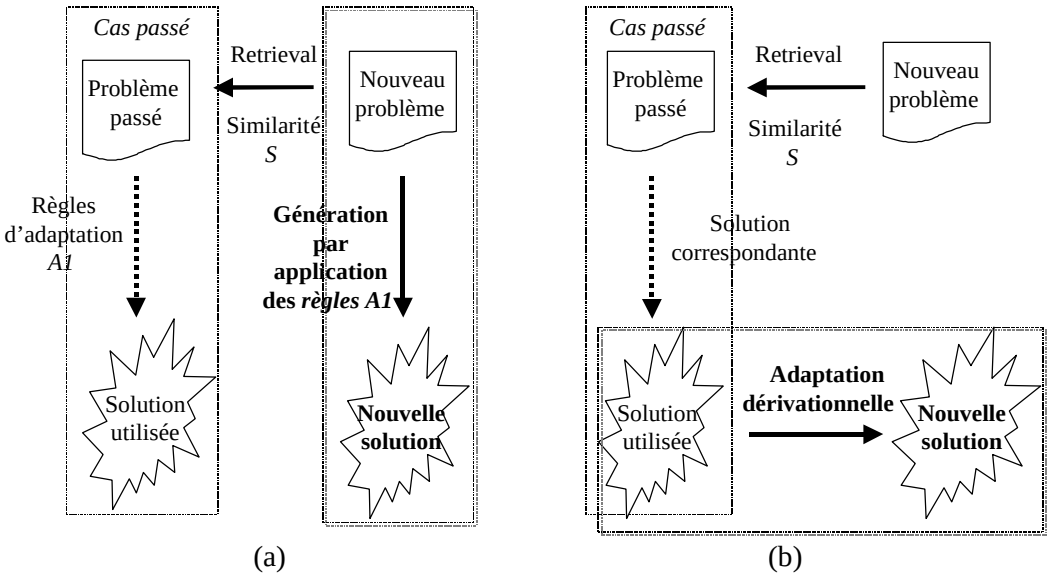


Figure 2 – Adaptation de type a) structurel, b) dérivationnel

[3] Une autre phase du modèle, la révision de solution ("revise") n'est pas présentée 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.

Peu de systèmes CBR font de l'adaptation complètement automatique. Pour la plupart des systèmes, une intervention humaine est nécessaire pour compléter une solution partielle ou tout simplement pour générer une solution entièrement à partir d'exemples. Comme cette phase est fortement dépendante de l'application, il est difficile d'en dégager des principes généraux. Il est important de mentionner que les tâches de classification ne nécessitent pas d'adaptation (nombre prédéterminé de catégories). Pour les tâches de type "synthèse" (e.g. design ou planification) le développement d'un module d'adaptation automatique est souvent coûteux en terme de temps et d'efforts. Le degré d'intervention humaine dépend alors des bénéfices en terme de qualité de solution que peut apporter l'automatisation de la phase d'adaptation.

- *Maintenance ("retain")*: la phase d'intégration de la nouvelle solution dans la base de cas et de la modification du contenu et de la structure du système CBR. Une stratégie simple pour aborder cette phase 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 (e.g.: indexation) pour en faciliter l'exploitation. On peut également altérer le cas en modifiant les attributs et leur importance relative. Cet aspect de recherche est actuellement l'un des plus actifs dans le domaine du CBR.

Connaissances

Les différentes connaissances utilisées par un système CBR sont regroupées en quatre catégories ("knowledge containers"):

- Vocabulaire d'indexation: un ensemble d'attributs descriptifs ("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 (indexation).
- Base de cas: l'ensemble des expériences structurées qui seront exploitées durant 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 fonctions des index 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. Puisque ces modèles reposent sur des principes variés, nous décrivons dans les sections suivantes les différences les plus importantes.

2.2.1 Modèle structurel

Le modèle structurel a émergé des premières vagues 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 doit être capable d'élaborer un modèle de données du domaine applicatif. Tel qu'illustré à la figure 3, les cas sont complètement structurés et sont représentés par paires <attribut, valeur> (similaire à un "frame" ou à un objet). 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 (représentation plate - "flat") ou sur plusieurs niveaux (hiérarchie d'attributs).

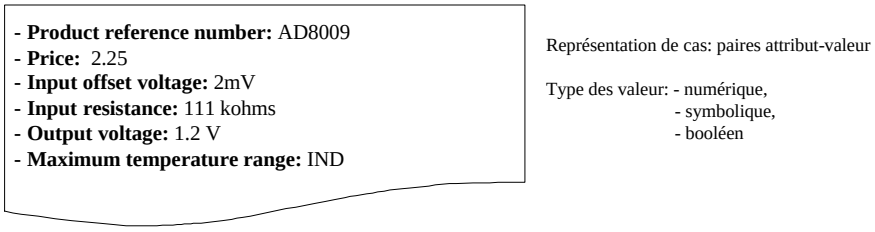


Figure 3 – Exemple de structuration d'un cas en CBR structurel

La similarité entre deux cas est mesurée en fonction de la distance entre les valeurs de mêmes attributs. Pour des attributs dont l'échelle est nominale ou ordinale, la similarité est habituellement déterminée par la correspondance exacte entre les valeurs des attributs. Pour les autres échelles (e.g. attributs définis sur les réels), il est possible d'évaluer un degré partiel de similarité entre deux valeurs (e.g. similarité entre 24,5 km et 25,2 km). 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 [Leake96] identifie environ dix techniques permettant de générer des solutions par substitution, transformation partielle ou dérivation complète. Ces techniques sont habituellement mises en oeuvre par des systèmes à base de règles, ce qui nous ramène aux problèmes d'acquisition de connaissance et d'absence de principes généraux pour certains domaines. Pour en limiter les difficultés, certaines approches évitent l'adaptation en sélectionnant, durant la phase de recherche, des cas qui nécessiteront peu d'adaptation.

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. Pour obtenir des solutions satisfaisantes, l'usager doit avoir a priori une bonne idée de tous les facteurs pouvant influencer la résolution de son problème. Toutefois pour certains problèmes, il est difficile de déterminer à l'avance les aspects de la situation sont importants.

Par ailleurs, cette exigence présuppose une expertise du domaine d'application, ce qui n'est pas le cas chez tous les usagers de systèmes CBR. Des usagers novices éprouvent parfois des difficultés à bien caractériser une situation à l'aide de valeurs numériques ou symboliques. Par exemple, les préposés des centres d'appels ne connaissent pas tous les aspects techniques des produits vendus par la compagnie, et leurs clients encore moins. Le modèle conversationnel a donc été proposé pour surmonter ces difficultés.

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 [Aha01].

Structure de cas

Un cas du modèle conversationnel consiste en trois parties (figure 4):

- un problème P : une brève description textuelle de la nature du problème exprimée habituellement en quelques lignes. Cette description est utilisée pour faire une première recherche dans la base de cas.
- une série de questions et de réponses Q_A : des index, exprimés sous forme de questions ont pour but d'obtenir plus d'information sur la description du problème. Chaque question a un poids qui représente son importance par rapport au cas.
- une action A: une description textuelle de la solution à mettre en oeuvre pour ce problème. Cette description n'est pas structurée ("free-text"). Par contre, certains systèmes permettent d'associer des informations additionnelles qui accompagne le cas (e.g. fichiers en attachement).

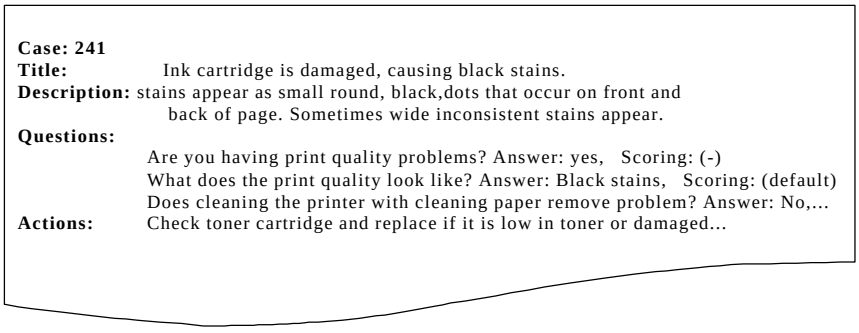


Figure 4 – Exemple de cas pour le modèle conversationnel

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

Schéma de résolution pour le CBR conversationnel

En décrivant progressivement les différents aspects du problème, on espère réduire le nombre d'attributs nécessaires pour trouver des solutions au problème courant. 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.
- 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. Il peut également être acheminé par courrier électronique à un analyste qui pourra le résoudre manuellement et l'ajouter à la base de cas.

Calcul de la similarité et des priorités

La similarité initiale entre les cas de la base et le problème à résoudre est déterminée à l'aide d'une représentation vectorielle. La similarité est estimée par la comparaison des descriptions textuelles de la partie "problème" qui sont converties soit en vecteurs de n-grammes (de caractères) ou de mots-clés. Certains systèmes font une analyse linguistique superficielle qui consiste à enlever les mots outils ("stop words") ou à faire une analyse morphologique simple ("stemming"). Le système NACODAE [Aha00] fait également du "part-of-speech tagging" pour reconnaître les noms et les verbes et leur associer des synonymes. Ceci équivaut à faire l'expansion de la description de problème.

Les questions présentées à l'usager sont sélectionnées en fonction de leur fréquence dans les cas les plus similaires au problème à résoudre. Ces fréquences sont pondérées par le poids de chaque cas.

Une fois que des réponses ont été fournies par l'usager, la valeur de similarité *score(Q,C)* entre chaque cas *C* et le problème *Q* est mise à jour à l'aide de la formule suivante:

$$score(Q,C) = \frac{same(Q_{qa}, C_{qa}) - diff(Q_{qa}, C_{qa})}{|C_{qa}|}$$

Adaptation

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"). 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.

Applications

Le modèle conversationnel est actuellement le plus répandu parmi les applications commerciales. Ce modèle CBR est souvent adopté pour concevoir des systèmes de type “help-desk” (“hotline”, CRM, etc.). Le modèle est particulièrement bien adapté pour les applications de service à la clientèle qui exigent une interaction avec le client.

Cette approche offre de nombreux avantages pour la mise en pratique. Premièrement, les problèmes sont décrits en langue naturelle, ce qui rend plus facile l’utilisation du système par un usager n’ayant pas de connaissance du CBR. De plus, les cas sont structurés à partir d’un ensemble de paires <questions, réponses> qui permettent de guider les interactions avec l’usager du système.

Puisque la base de cas permet aux analystes peu expérimentés de répondre aux appels de routine, les préposés plus expérimentés peuvent se concentrer sur les cas inhabituels et plus complexes. Traditionnellement, le client interagit directement avec l’analyste qui utilise le système CBR pour l’aider à résoudre le problème de ce client. Dans le schéma de type web de la figure 5, le client utilise lui-même le système CBR conversationnel pour tenter d’obtenir une solution à son problème. L’analyste n’intervient que pour les situations que le système ne peut résoudre de manière satisfaisante.

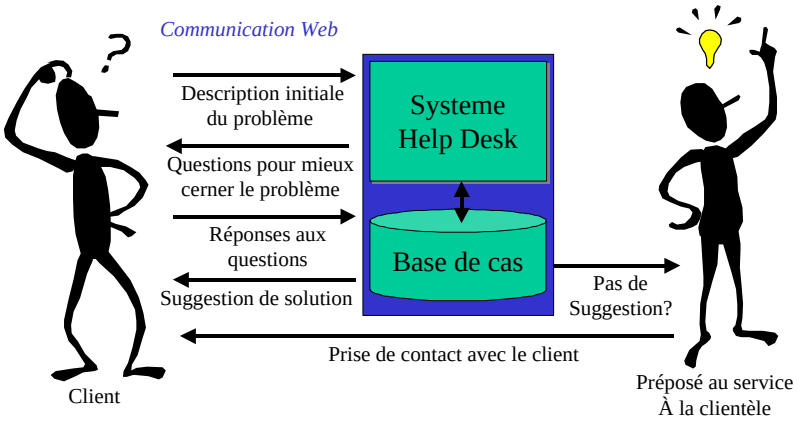


Figure 5 – Interaction client-analyste pour le service à la clientèle

2.2.3 Modèle textuel

Les travaux sur le raisonnement à base de cas textuel 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 qui a un seul attribut dont la valeur est textuelle tandis qu'un cas textuel semi-structuré est un cas dont un sous-ensemble des 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 soit: 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 date, aucune représentation standard ne s'est dégagée pour le modèle textuel. Les approches actuelles misent uniquement sur les mécanismes de recherche dans 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 d'index basés sur les caractéristiques du domaine (concepts, catégories, sujets, mots-clé, etc.). Pour ce type de travaux, on vise à structurer le plus possible les cas textuels pour pouvoir 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 applicatif est restreint. Le projet SMILE [Bru97] présenté à la section 3.1 en est un exemple.
- *Extension du modèle de recherche d'information:* pour ce type de travail, les efforts sont consacrés à élaborer des mécanismes de recherche plus sophistiqués tout en gardant le processus d'indexation le plus simple possible. Dans ce cadre, les cas sont indexés selon une approche basée sur la fréquence de mots-clés. 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 de recherche vectoriel. Cette approche semble plutôt valide pour les applications génériques qui veulent conserver une indépendance par rapport au domaine d'application. Le projet FAQFinder [Bur95] présenté à la section 3.3 en est un exemple.

Ces deux pôles sont en fait des stéréotypes auxquels empruntent la plupart des approches actuelles. Nous présentons à la section 3 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 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 comparer, par mots-clé ou n-grammes 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 naturelle est y utilisée uniquement dans le but de rendre les questions plus intelligibles à l'usager du système.

3.0 Travaux pertinents en CBR textuel

Dans cette section, nous présentons des travaux sur le CBR textuel qui diffèrent par le niveau de structuration des cas, la complexité des métriques de similarité et le mécanisme de recherche sur la base de cas.

3.1 SMILE - factorisation des cas par la catégorisation de textes

Ces travaux, menés par Steffanie Brunninghaus et Kevin Ashley de l'Université de Pittsburgh [Bru97], 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 des "facteurs".

Les cas étudiés relèvent du domaine légal impliquant des causes de fraude reliées aux secrets industriels d'entreprises. Un cas est constitué d'une description de la cause, des plaidoyers et du jugement. Ces travaux ont été initiés par Aleven [Ale96] qui propose un système tutoriel, basé sur le CBR structurel, pour enseigner aux étudiants de droit comment mener une argumentation pour ce type de cause. 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 ("issues").

Dans les travaux de 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* [Bru97]: chaque texte est représenté comme un vecteur de fréquence de mots. On utilise des techniques d'induction de graphe de décision (ID3) pour apprendre comment attribuer les facteurs légaux à chacun des textes. 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", "Libbow", "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* [Bru99]: divers passages sont étiquetés manuellement pour indiquer la présence de facteurs dans le texte (voir figure 6). Ces passages servent de corpus d'entraînement pour apprendre comment catégoriser les portions de textes. Un thésaurus est également utilisé pour identifier les correspondances entre des termes de différents passages. Ceci permet de détecter la présence de facteurs dans des textes 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 certains 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* [Bru01]: plus récemment, Brunninghaus a proposé d'utiliser le système d'extraction d'information AutoSlog [Ril96] 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 débutent et aucune expérimentation n'a encore été menée.

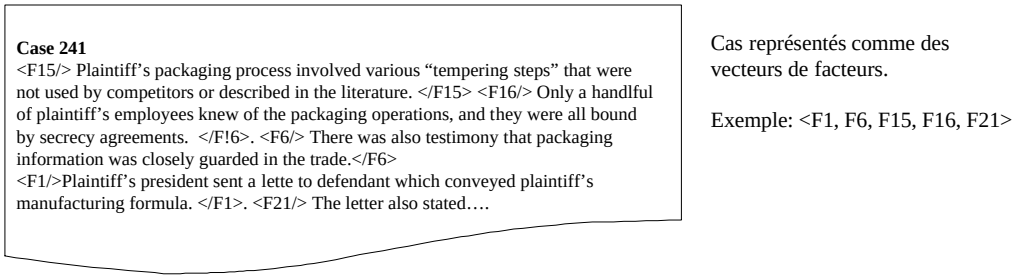


Figure 6 – Étiquetage de passages selon des catégories (facteurs)

Tout comme pour les travaux de Aleven, la similarité entre les cas est établie simplement en fonction du nombre de facteurs en commun et/ou qui diffèrent. Ces travaux ne proposent pas de techniques pour faire l'adaptation de textes légaux.

3.2 DRAMA – cas partiellement textuels

Le projet DRAMA [Lea99][Wil00] 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"[4]), d'attributs descriptifs (e.g. caractéristiques du moteur) et d'explications textuelles (annotations) donnant des précisions sur les choix des concepteurs et sur les caractéristiques des composantes de l'avion. Les cas contiennent des parties structurées (diagrammes et attributs valués) et des parties non structurées (textes). Les auteurs les qualifient de cas "weakly-textual", i.e. des cas dont une partie est textuelle mais dont la portion la plus importante pour l'application est non-textuelle.

Ce système intègre un module CBR avec des outils interactifs pour la capture de connaissances. Les outils interactifs permettent de définir et de consulter les cartes conceptuelles à différents niveaux d'abstraction. Le module CBR permet de rechercher les designs les plus pertinents et de les adapter manuellement.

Le mécanisme de recherche combine des fonctions de similarité sur les attributs structurés et sur les annotations textuelles. Afin de simplifier la recherche textuelle, les auteurs utilisent des principes de recherche d'information. Ces techniques offrent l'avantage de comparer de courts textes tout en permettant l'intégration facile de ces résultats avec les mesures de similarité non-textuelles pour les cartes conceptuelles et les attributs structurés.

Plus précisément, la recherche comporte les étapes suivantes:

- chacun des attributs textuels est converti individuellement en vecteur de termes (modèle vectoriel); puisque les descriptions textuelles sont courtes, la sélection de termes ne repose pas sur un calcul de fréquence mais plutôt sur les syntagmes nominaux ("noun phrases") de type *Nom-Nom*, *Adj-Nom* ou *Nom-Prep-Nom*. Les syntagmes sont identifiés avec l'aide d'un dictionnaire, le lexique Moby [War94].

[4] 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.

- un poids, similaire au tf*idf[5], 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.

En résumé, le système DRAMA est principalement un système CBR structurel dont quelques attributs sont textuels. 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 parties textuelles en fonction du nouveau design.

3.3 FAQ-Finder – exploitation de questions-réponses

FAQFinder [Bur95] [Bur97] est un système de questions-réponses basé sur les "Frequently-Asked Questions" (FAQs) de USENET. Un FAQ est une réponse à une question fréquemment posée à un groupe d'intérêt (e.g. groupe de programmation Java). Un FAQ est considéré comme un cas CBR car il contient la description d'un problème (la question) et d'une solution (la réponse) (figure 7).

Le système est conçu pour recevoir en entrée une question en langue naturelle et identifier les FAQs de USENET qui sont les plus similaires à cette question.

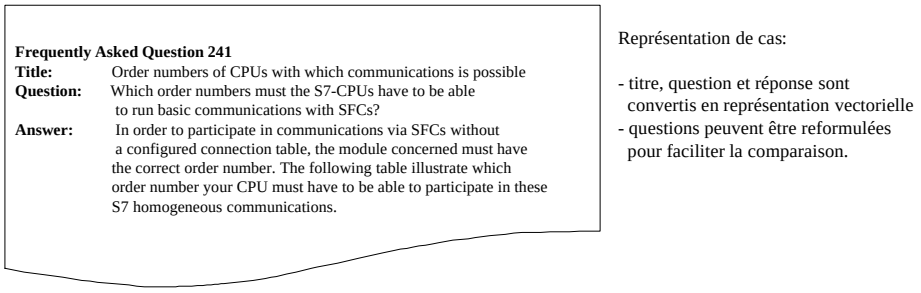


Figure 7 – Structuration des frequently-asked questions

La recherche de réponses pertinentes dans FAQFinder est effectuée en 2 étapes:

- la première étape permet de choisir, à partir de tous les fichiers FAQ 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 les langages de programmation Lisp et Java. Cette étape adopte une approche de recherche d'information et utilise le système SMART [Sal85]. 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). 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 quelques milliers de fichiers de FAQs et d'en retenir quelques dizaines seulement.

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

- 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 similaires de type tf*idf. Différentes fonctions ont été testées dans leurs expérimentations.
    - Métrique sémantique: en utilisant le thésaurus WordNet, la distance sémantique entre chaque paire de mots est estimée. Pour évaluer cette distance, un algorithme de type "marker-passing" (proposé pour des travaux de réseau sémantique) est utilisé. Cet algorithme correspond à peu près à une approche de type "edge-counting" qui estime la distance entre deux concepts à partir du nombre de liens qui les séparent.
    - 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 FAQ de UseNet. 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 (rappel pour statistique seulement - 55%, sémantique seulement – 58%, combiné - 67%). La qualité des résultats est limitée par l'utilisation de WordNet qui est une ressource linguistique trop générale pour ce type d'application.

Plusieurs tentatives ont été menées pour améliorer la performance du système. Le "part-of-speech tagging" a été utilisé pour identifier les termes importants des questions. Pour la deuxième étape, des techniques pour reconnaître le type de question et pour en faire la reconversion ont été appliquées. En exprimant la question sous différentes formes, les auteurs pensaient pouvoir améliorer les performances du système. Par exemple, la question "How often should the oil of my car be changed" peut être remplacé par la question "When should I make an oil change of my car". Ceci correspond à paraphraser des questions à partir de canevas prédéterminés. Finalement des expérimentations ont été menées afin de déterminer automatiquement la pondération de chacune des fonctions de similarité par apprentissage automatique (algorithmes génétiques). 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).

3.4 CBR-Answers – réseau pour la recherche de cas

Ce système, développé par Mario Lenz [Len97] [Len99], est basé sur une structure de réseau appelée "case retrieval net". Le réseau est utilisé pour "compiler" la base de cas et pour effectuer la recherche. Les valeurs de similarité sont propagées dans les nœuds du réseau et permettent de déterminer la pertinence de chacun des cas. Un exemple de réseau est présenté à la figure 8.

Le réseau contient un ensemble de nœuds appelés “information entity” (IE). Les IEs décrivent des éléments de documents tels que des mots-clés, des termes complexes, des paires attributs-valeurs, des catégories du domaine et des identifiants de cas.

Les liens du réseau décrivent soit a) l’appartenance d’un “information entity” à un cas, soit b) une relation de similarité entre deux entités. 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.

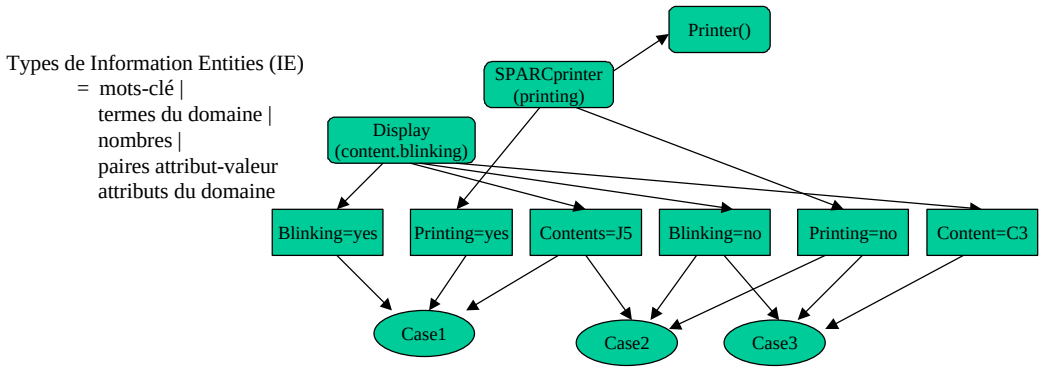


Figure 8 – Exemple de réseau de type “case retrieval net”

Une procédure est proposée pour l’indexation des cas et la construction du réseau [Len98]. Cette structuration repose uniquement sur une analyse lexicale des textes.

- Chaque terme des documents est étiqueté selon sa catégorie lexicale (“part-of-speech”);
- Les textes sont normalisés en remplaçant les mots-clés par leur racine morphologique (“stemming”);
- Les mots ayant une racine commune sont regroupés. Les racines forment les index des cas.
- Les termes sont classés selon trois catégories:
  - Inutile: ce que l’on retrouve dans une liste de mots-outils (“stoplist”) avec en plus des déterminants, des auxiliaires, des verbes et des pronoms.
  - Utile: les adverbes, les adjectifs, les verbes et les noms.
  - Potentiellement utile: les termes qui n’appartiennent pas aux catégories précédentes.
- On élimine les termes inutiles et on détermine la fréquence des termes utiles et potentiellement utiles.
- On fait la vérification manuelle des termes utiles et potentiellement utiles.

La recherche dans un “case retrieval net” suit les étapes suivantes:

- Les termes de la requête sont découpés en mots clé et associés à des “Information entities” du réseau.
- Les entités du réseau sont ensuite activées.
- Les valeurs de similarité sont propagées parmi les différents IEs du réseau pour déterminer les IEs similaires.
- 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 le 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 Seimens AG). Ces systèmes permettent la recherche de documentation technique (e.g. spécifications de composantes, 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 de recherche des case retrieval net. Une étude a été menée avec Simatic pour évaluer la contribution des différents niveaux de "Information entities". Avec un corpus de 500 documents, l'utilisation de simples mots-clé offre la plus faible performance en terme de courbe précision-rappel et l'ajout successif de chaque couche d'entités (thesaurus, glossaire et thème) 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.

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

Ces travaux ont été menés par Jody Daniels et Edwina Rissland de l'Université du Massachusett à Amherst [Dan96], [Ris95]. Ils ont développé SPIRE, un système hybride de CBR et de recherche d'information. 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. Dans ce schéma, le module CBR agit donc 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 étude menée à partir de 20 documents et 10 requêtes indique que cette phase du système offre une performance équivalente à un humain; chacun offrant de meilleurs passages pour un même nombre de requêtes.

3.6 PRUDENCIA – structuration de cas

PRUDENCIA est un système qui facilite la recherche documentaire en jurisprudence légale [Web98a][Web98b]. Il permet de rechercher des situations, décrites dans des documents textuels, qui sont similaires à une nouvelle cause juridique.

Weber propose 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 (e.g. "heading", "abstract", "body", "closing"). 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.

A partir d'exemples indiquant comment ces champs apparaissent dans les textes, un certain nombre de méthodes ont été développées pour extraire du texte des informations qui alimentent les formulaires. On mentionne les méthodes suivantes:

- Directe: quelques attributs sont explicites et leurs valeurs sont situées à des positions fixes dans le texte;
- Par mots-clé: on recherche dans les sous-sections des mots-clé contenus dans une liste d'expressions pour chaque champ du formulaire. Pour faciliter la recherche, un "stemming" est appliqué à chaque mot et un dictionnaire de synonymes est utilisé;
- Par patron: des expressions régulières permettent d'obtenir des numéros d'articles de loi (e.g. "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 manuellement suite à une comparaison avec d'autres cas structurés (similaire à une classification à base de cas).

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]). Weber mentionne que le temps de recherche est d'environ 45 secondes (un temps de réponse que nous jugeons lent). Malheureusement aucune évaluation de la précision et du rappel du système n'est présentée.

3.7 Discussion des travaux en CBR textuel

Les deux tableaux suivants résument les principaux travaux en CBR textuel présentés dans cette section. Le Tableau I décrit les particularités de la tâche accomplie par le système CBR et le Tableau II présente les particularités techniques de l'approche CBR préconisée.

Tableau I – Caractéristiques des domaines des systèmes CBR textuel

Travaux	Tâche	Caractéristiques du domaine	Caractéristiques des 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.
DRAMA	*Recherche* de dossiers de design et préservation de la connaissance des concepteurs.	Domaine aéronautique, ce qui laisse présager un vocabulaire restreint et un groupe de concepts relativement limités.	Les cas sont en partie structurés (diagrammes, attribut-valeur) et en partie textuels. Les attributs textuels sont courts et moins importants que les attributs structurés.
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 fichiers initiaux sont longs, contenant plusieurs centaines de FAQs. Chaque FAQ est relativement court et est structuré en deux parties: question et réponse.
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.
SPIRE	*Recherche* de passages pertinents dans un corpus de documents.	Domaine de la gestion de faillite personnelle. L'approche ne dépend pas des caractéristiques du domaine. Applicable à un domaine restreint seulement.	Aucune contrainte sur la nature des textes. Une base de cas donne une description de documents du corpus (cas structurels) et l'autre base contient des extraits textuels.
PRUDENCIA	*Recherche* de textes légaux.	Domaine de la jurisprudence légale.	Des textes légaux relativement longs et complexes. Tente de prendre en compte la structure rhétorique des textes. La base contient 3500 cas.

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). 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 retourne 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.

L'étendue des domaines d'application de ces systèmes varient également. Les approches de SMILE et PRUDENCIA présument que des connaissances du domaine sont disponibles. Par contre, il y a peu de contraintes pour FAQFinder (aucune modélisation des domaines des FAQs).

La plupart des textes utilisés dans ces applications ne sont pas structurés. Seules les questions-réponses de FAQFinder et les annotations de design de DRAMA offrent un découpage en plusieurs attributs. Mais le nombre d'attributs est faible et la longueur des textes est relativement courte.

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.

Tableau II – Caractéristiques techniques des systèmes CBR textuel

Travaux	Indexation	Structuration	Recherche
SMILE	Un groupe de facteurs provenant d'une modélisation manuelle du domaine.	Catégorisation des documents, de passages ou d'informations extraites. Utilise des techniques d'apprentissage automatique.	Comparaison de présence/absence de facteurs (abordé dans les travaux sur CATO [Ale97]).
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.
FAQFINDER	Les mots-clés des fichiers (approche IR). Les mots-clés font l'objet de stemming et sont filtrés par rapport à une liste de mots-outils ("stoplist").	Création semi-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 sur WordNet). Reformulation de questions (paraphrasage) pour faciliter la comparaison.
CBR-ANSWERS	"Information Entities" qui contiennent les mots-clés des fichiers, des syntagmes nominaux et des 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 le réseau.
SPIRE	Un groupe d'attributs ("features") du domaine sélectionnés par le concepteur.	Création manuelle de "frames" et d'extraits textuels.	Comparaison d'attributs pour le CBR et recherche documentaire par le système INQUERY.
PRUDENCIA	Des index fournis par un expert du domaine juridique.	Création semi-automatique de "frames"	Comparaison d'attributs avec mesures binaires et graduées.

Le Tableau II illustre qu'aucun des projets présentés ne propose de processus complexes dans le choix des index 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 de stemming.

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 oeuvrer sur les deux plans.

Quelle approche est la meilleure? Devrait-on miser sur une représentation conceptuelle de cas, sur des fonctions de similarité plus riches ou sur un formalisme de recherche plus élaboré? Il ne se

dégage malheureusement pas de réponse à cette question à partir des travaux répertoriés mais plusieurs facteurs jouent un rôle important dans le choix de l'approche.

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. Par contre, aucun travail de recherche n'a encore été effectué pour tenter de quantifier chacune de ces dimensions pour une variété de corpus.

Bien que déjà mentionné à de nombreuses reprises, il est important de souligner qu'aucun système 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 les applications de 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 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 de texte. Or il y a lieu de croire que la communauté CBR ne maîtrise pas actuellement les outils nécessaires pour aborder ces tâches.

3.8 Autres travaux pertinents

Cette section présente quelques travaux connexes qui décrivent les principaux enjeux de la construction de base de cas et qui illustrent l'exploitation du CBR et des techniques de traitement de la langue.

3.8.1 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 d'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 éliciter 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 IV). 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 IV 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-clé seulement (FAQFinder), catégories seulement (SMILE), termes complexes seulement (DRAMA), passages seulement (SPIRE).

Tableau IV – Explicitation du contenu textuel par niveau de structuration [6]

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 abstrait 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 [Gut99] [Tur00] ou la construction automatique de lexique [Roa98] [Gre92].
Mots-clé	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.

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 des différents niveaux pour structurer les cas textuels. Une étude a été menée par Lenz [Len98] 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 système CBR industriel [Inreca98]. 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. Certaines reposent sur la mise en correspondance d'une ontologie du domaine d'application avec une description des fonctions d'un système CBR [Fuc01] [Dia01]. 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 [McS01] [Smy99]. 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 [Mul01]. 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 [Smy99][Qia00] et pour varier le poids de chacun des attributs des cas [Aha97]. Ces deux dimensions permettent d'ajuster la performance du système (e.g. temps de réponse du processus de recherche) et de maintenir la qualité des solutions (e.g. la couverture du système).

[6] 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.

Ces techniques de maintenance sont insuffisantes pour le processus d'authoring de base de cas textuelles. 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) est centrale au processus d'authoring.

3.8.2 Utilisation du CBR en traitement de la langue naturelle.

Une approche de réutilisation de documents a été proposée par Branting et Lester [Bra96] 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 (e.g. testament), des textes qui sont habituellement longs et complexes. Ces recherches sont les seules que nous avons répertoriées dans la litérature 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 [Mar00]. Claire Cardie [Car93, Car96] 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 [Dea00], 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, utilisé pour la classification de nouvelles instances. Le système a été utilisé pour le "part-of-speech tagging" et le "chunking".

3.8.3 Extraction adaptative 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/syntaxique et de système à base de règles [Kos98]. 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 [Cir00, Cir01]. Ces techniques se révèlent particulièrement efficaces pour des textes structurés (près de 100% de rappel et précision sur quelques 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) [Sod99].

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 certaine 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.

4.0 Caractéristiques du corpus

Afin d'illustrer les particularités de notre domaine d'application, nous avons analysé un corpus de 102 messages regroupés sous le thème "financial". Nous décrivons les principales propriétés de ces messages par rapport au domaine applicatif, aux textes et à l'interaction entre l'analyste et l'investisseur.

4.1 Caractéristiques du domaine

Les messages acheminés aux analystes de BCE portent principalement sur les indicateurs financiers de BCE, sur la valeur de son titre boursier et sur les dates de divulgation des rapports financiers (e.g. rapports, appels conférence). Plus particulièrement, on note des questions ayant trait aux aspects suivants:

- *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 sur la valeur du titre en bourse et la variation de ce titre par rapport aux indices boursiers (e.g. indice TSE). Également, des demandes d'explication de la performance en bourse du titre de BCE ou des filiales.
- *Sources d'informations*: des 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 également des requêtes sur l'obtention de la date et des coordonnées de différents événements (divulgation de rapports, appels conférence téléphoniques, rencontres d'actionnaires).

4.2 Caractéristiques des textes

Nous avons identifié un certain nombre de caractéristiques que nous devrons prendre en compte lors de la conception d'un système informatisé de réponse au courrier électronique. Les principales caractéristiques sont:

a) Longueur de questions-réponses: 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 (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.

b) Uniformité des questions-réponses: les réponses sont écrites par un nombre limité d'analystes (5-10). La structure des réponses est répétitive et plusieurs réponses sont quasiment identiques (e.g. dates de divulgation de résultats financiers). Les questions provenant de différentes personnes, le style de rédaction varie d'une question à l'autre. On retrouve donc différentes formulations pour une même requête (paraphrasage). 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.

c) Généricité/spécificité des questions-réponses: la plupart des questions concernent une information précise (e.g.: la valeur d'un indicateur financier, une date, un numéro de téléphone) pour une période de temps ou pour un événement bien déterminé (e.g. "la fin du dernier trimestre" ou "le prochain rapport financier"). On retrouve quand même quelques questions génériques du genre "Pourquoi devrais-je investir dans BCE?". On retrouve un nombre considérable de réponses génériques qui n'abordent pas directement la question. Par exemple, on indique à l'investisseur que de nombreuses informations sont disponibles sur le site web. On note parfois des "méta-réponses", i.e. 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".

d) Structure des messages: une question est habituellement constituée de trois parties: une brève description du contexte, une ou plusieurs questions, et les coordonnées de l'investisseur (nom, affiliation, adresses postale et électronique). Certains demandeurs fournissent des justifications pour motiver leurs questions. Par exemple des investisseurs corporatifs indiquent qu'ils mènent une étude sur BCE ou des particuliers font part de leur intention d'investir dans l'entreprise. 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).

e) Questions 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. Certains demandent une même information pour différentes filiales ou plusieurs indicateurs financiers pour une même compagnie.

f) Dépendance inter-messages: les messages sont indépendants entre eux. Chaque question est, en principe, posée par une personne différente. Les suites d'échanges multiples avec le même investisseur ne sont pas fréquentes (observée à une seule reprise).

g) Temporalité: on note les trois dimensions temporelles suivantes dans les messages.

- *Temporalité des informations*: certaines informations varient dans le temps comme les bénéfices ou la valeur de l'action. Ces valeurs sont liées à l'activité économique de l'entreprise, un processus dynamique. La nature des réponses en est fortement dépendante. Par exemple, une question sur la prochaine rencontre des actionnaires peut entraîner des réponses différentes selon que la date ait été fixée ou non. Les références aux heures, dates ou périodes peuvent être explicites (e.g. "le ratio P/E du premier trimestre de l'an 2000") ou implicites (e.g. "…du dernier trimestre").
- *Temporalité d*es *événements*: plusieurs messages font référence à un événement précis. On note deux types d'événements: périodique et non-récurrent. Comme événement périodique, on retrouve la divulgation de rapports financiers et les appels conférence téléphonique. Les messages liés aux événements périodiques sont toujours pertinents, mais peuvent faire référence aux événements précédents, actuels ou futurs. Comme événements non-récurrents, on note des occurrences inhabituelles ou à périodicité variable qui sont difficiles à prévoir et qui amènent une forte concentration de messages dans un court laps de temps. A ce titre, on note la vente des parts de Nortel ou une baisse subite du cours de l'action. Il est également intéressant de constater que l'absence d'événement suscite également des questions. Par exemple, des investisseurs s'interrogent sur la stagnation (absence de croissance) du titre de BCE pour une période prolongée.
- *Délai de réponse*: quelques requêtes imposent des contraintes sur les délais de réponse. Des contraintes sont explicites ("j'aimerais planifier cette semaine…"), d'autres sont implicites (e.g.. obtenir le numéro de téléphone pour l'appel conférence du lendemain). Finalement, certaines contraintes de temps sont "soft" (e.g. "le plus tôt possible").

h) Traitement des réponses: la plupart des questions sont traitées par le premier récepteur du message quoique certains messages sont redirigés vers un autre analyste. On note également que des messages demeurent sans réponse.

4.3 Caractéristiques de l'interaction entre l'investisseur et l'analyste

Dans cette section, nous nous intéressons à la forme et à la suite d'échanges de messages. Habituellement, l'interaction entre l'investisseur et l'analyste financier se résume à l'échange d'une seule requête et d'une seule réponse. Donc, contrairement aux systèmes traditionnels de "help-desk" ou à l'échange de messages sur des "bulletin board", il n'y a pas de succession de messages qui permettrait à l'analyste de mieux cerner les objectifs et les besoins de l'investisseur.

Ceci nous amène à émettre un certain nombre d'hypothèses sur les aspects suivants:

- *explicitation du contenu de la requête*: on suppose que les questions des investisseurs sont facilement compréhensibles par rapport au domaine d'application et qu'elles fournissent tous les éléments nécessaires à la réponse. Le contenu porte principalement sur la nature de l'information à obtenir (e.g. ratio P/E pour un trimestre donné) ou sur l'action sollicitée auprès de l'analyste (e.g. ajout à une liste de distribution).
- *but de l'émetteur*: la plupart des messages donnent peu d'indication sur les motivations de l'investisseur, telles l'achat d'actions. On suppose donc que cet aspect ne joue pas un rôle important dans le processus de gestion de réponse.
- *description de la problématique*: on retrouve rarement une description explicite des problèmes confrontés par l'investisseur (e.g. un investisseur corporatif qui éprouve des difficultés à mener une analyse financière de BCE). Il est donc inutile de formuler les situations comme un problème de diagnostic comportant des descriptions de symptômes et des causes possibles.
- *événement déclencheur*: certaines questions font suite à un événement particulier. Par exemple, de nombreuses questions ont suivi le réajustement du titre de BCE par rapport à Nortel. La détection de l'occurrence de ces événements est donc très importante dans la gestion des réponses.
- *rôle de l'émetteur*: il semble que les réponses soient indépendantes du rôle de l'émetteur (e.g: un investisseur corporatif ou individuel).

5. 0 Conception d'un système CBR pour la gestion de réponses

Afin de mieux cerner les facettes de la mise en oeuvre d'un système CBR pour la réponse automatique au courrier électronique, nous décrivons dans cette section les fonctionnalités que notre système doit posséder, les différentes architectures logiciels (modèle et processus) qui s'offrent à nous, ainsi que les motivations de notre choix de modèle CBR.

Aucune publication n'a été répertoriée jusqu'à maintenant dans la littérature CBR sur la gestion de réponse au courrier électronique. Les travaux de recherche les plus pertinents, actuellement menés au laboratoire de recherche de NEC au Japon, portent sur l'analyse de messages électroniques par des techniques CBR pour la détection de nouveaux problèmes [Shi01]. Ces travaux misent sur l'accroissement de mots-clé comme indicateurs de problèmes potentiels. Des travaux sont également menés par cette équipe sur le partage de "mail folders" et de "bulletin boards"convertis en base de cas [Kus01].

5.1 Fonctionnalités et tâches du système

Un système de réponse automatique au courrier électronique doit offrir les fonctionnalités suivantes de gestion des messages:

- la préservation: réception des messages, consignation de leur contenu, consultation et recherche par mots-clé, par sujet ou par émetteur.
- le suivi: la possibilité de savoir à tout moment si la requête a été traitée et à qui en revient la responsabilité, ainsi que la possibilité d'effectuer une analyse statistique des messages et de la performance du système.
- la gestion des réponses: le filtrage des messages par catégorie ou par priorité, la proposition aux analystes de réponses potentielles, la possibilité pour l'analyste de valider ou modifier la réponse proposée. La gestion est un cycle qui comprend les fonctions de catégorisation, de prioritisation, de routage, de suggestion et de réponse**.**

La fonction de préservation des messages est naturelle pour un système CBR. Sa base de cas assure l'intégrité du contenu des messages antécédents. De plus la structuration des messages sous forme de cas permet de consulter différentes portions de l'entête des messages ainsi que leur contenu.

Pour effectuer le suivi des messages avec un système CBR, il est nécessaire d'ajouter aux cas des attributs descriptifs qui conservent une trace des différentes opérations effectuées sur les messages. Ces informations peuvent être par la suite exploitées pour des fins d'analyse.

La gestion de réponses est la fonction principale du système CBR. Cette fonction doit faciliter la génération d'une réponse adéquate pour une nouvelle question soumise par un investisseur. Elle peut mener soit à l'envoi d'un simple accusé de réception, d'une ancienne réponse contenue dans la base de cas ou d'une réponse conçue spécifiquement pour le nouveau message.

Nous pouvons formuler la tâche de la gestion de réponse selon deux types d'inférence [Sch00]. Le premier type d'inférence, dit d'analyse, mise principalement sur une forme de "compréhension" des questions. Par exemple, une approche de type "classification" nous amènerait à catégoriser les cas selon différentes classes et d'associer une réponse standard à chacune de ces classes. Nous pourrions retenir cette approche pour les questions routinières. L'autre type d'inférence, dite de synthèse, favorise la construction d'une nouvelle solution à partir de l'énoncé d'un problème. Par exemple, une approche de type "planification" ou "design" viserait à générer une réponse comportant une suite d'affirmations dont le contenu tente de satisfaire des buts, des contraintes et un contexte énoncés dans la question. Cette approche est plus riche et correspond mieux à la tâche

que nous voulons accomplir. Toutefois, elle demande une forte structuration des messages des investisseurs, ce qui la rend plus difficile à réaliser en pratique.

5.2 Composantes et architecture du système CBR

Tel que décrit à la section 2 de ce document, un système de raisonnement à base de cas comporte un certain nombre de processus et de connaissances ("knowledge containers"). Certains des processus du système sont exécutés lors des activités de résolution de problème ("on-line") tandis que d'autres le sont lorsque le système n'est pas en opération ("off-line") (Figure 9). Les fonctions "online" permettent de résoudre chaque problème individuellement, c'est à dire de gérer chacune des nouvelles questions des investisseurs. Les phases de recherche et d'adaptation permettent respectivement de sélectionner des messages précédents qui sont similaires à la nouvelle question et de modifier ces anciennes réponses en fonction du contexte de la nouvelle question. Les fonctions "off-line" permettent de mettre en fonction le système et de mettre à jour périodiquement le contenu de ses différentes bases de connaissances. Parmi ces fonctions, on note le processus d'authoring qui supporte la construction initiale de la base de cas et le processus de maintenance qui aide à raffiner et à optimiser le comportement du système après un certain temps d'utilisation.

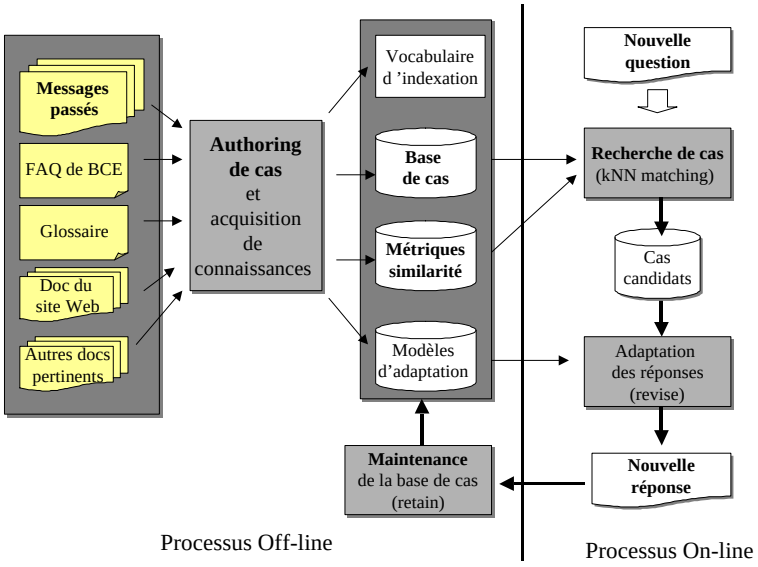


Figure 9 – Composantes du système CBR

Au niveau des connaissances du système, on retrouve la base de cas, le vocabulaire d'indexation, les métriques de similarité et les règles d'adaptation. Pour appuyer la construction des connaissances, on dispose des ressources suivantes:

- Messages antécédents: la source principale pour construire la base de cas.
- "FAQ": un groupe de questions-réponses fréquentes a été compilé manuellement par des analystes de BCE. Ces "FAQ" pourrait permettent de répondre à certaines questions d'intérêt général.
- Glossaire: on retrouve sur le site web de BCE une liste de termes financiers utilisés par les analystes. Le glossaire contient chacun des termes en français et en anglais. Il y a également ailleurs sur le web de nombreux glossaires reliés au domaine financier qui pourraient nous être utiles pour enrichir le vocabulaire d'indexation.

- Pages web: le site de BCE contient un ensemble de documents fournissant de l'information sur différents aspects financiers de la corporation. Comme la plupart des compagnies offrent un service électronique aux investisseurs, le web regorge de documents qui pourraient permettre de cumuler des informations et des statistiques sur des sites similaires.

On peut envisager différentes architectures pour combiner ces composantes et ces connaissances au sein d'un système de réponse automatique. Ces architectures varient en fonction du degré d'autonomie du système, du degré d'intervention de l'analyste et de l'étendue du processus CBR. Le Tableau III décrit les trois niveaux envisagés de système de réponse automatique.

Tableau III – Niveaux de l'architecture d'un système de réponse automatique

Caractéristiques	***Description des fonctionnalités***	***Surnom***
Système *embarqué* avec autonomie limitée	A ce niveau, l'analyste a le plein contrôle du système. L'analyste a la responsabilité de choisir les messages qu'il juge utiles et de combiner le contenu de différentes réponses jugées pertinentes. Les composantes CBR sont imbriquées dans le système d'édition de réponses. Sur demande, ces composantes effectuent la recherche de réponses potentielles. A ce niveau, on ne considère pas de fonction d'adaptation automatique. Ce système ressemble aux "help desks" couramment utilisés dans les applications de service à la clientèle.	Glaneur
Système *autonome* avec droit de regard de l'analyste	A ce niveau, le système CBR est en contrôle du processus de réponse automatique et réagit à l'arrivée de nouvelles questions d'investisseurs. Le système a l'entière responsabilité de construire une réponse initiale. Par la suite, le système offre à l'analyste la possibilité de modifier et de valider sa suggestion.	Rédacteur
Système à *initiative-mixte* misant sur la synergie entre l'analyste et les composantes CBR.	A ce niveau, le contrôle est partagé entre l'analyste et les modules CBR, et ce sans fixer un schéma d'interaction prédéterminé. Un formulaire électronique guide l'interaction entre le système et l'analyste pour remplir les différentes sections d'une réponse. Ce système à initiative mixte ("mixed-initative system") est une forme hybride qui se situe entre le glaneur et le rédacteur.	Éditeur intelligent

D'un point de vue CBR, chacun de ces niveaux présente des défis. Pour le glaneur, le contrôle appartient à l'analyste qui doit générer une réponse. Il est donc possible, lorsque le module CBR embarqué est invoqué, que l'analyste ait déjà commencé à répondre partiellement à la requête de l'investisseur. Le module CBR doit donc pouvoir tenir compte de l'état de la réponse dans son processus de recherche. Concrètement, le mécanisme de recherche doit être capable de focaliser sur les portions de cas qui sont encore d'intérêt pour l'analyste et non sur les portions qui ont déjà fait l'objet de réponse. Ceci peut se traduire par l'élimination de portions du cas à résoudre (la requête) ou par la modification du poids de certains de ses attributs. De plus, la portion de réponses déjà rédigées par l'analyste devrait également contribuer à mieux cerner les cas les plus pertinents.

Le rédacteur doit aller plus loin que le glaneur dans le processus de génération de réponse. Pour être utile, ce niveau doit offrir des capacités d'adaptation relativement élaborées qui lui permettent d'atteindre une plus grande qualité de réponse. Tel que mentionné précédemment, les travaux actuels en CBR textuel ne proposent pas d'approche pour l'adaptation de texte, ce niveau nous amène donc à explorer une nouvelle voie de recherche.

L'éditeur intelligent est un compromis entre la servitude du glaneur et l'ardeur du rédacteur. D'une part, l'éditeur intelligent doit savoir miser sur les forces respectives de l'analyste et des composantes CBR. D'autre part, l'intervention de l'éditeur intelligent dans le processus de

rédaction doit tenter de maximiser la contribution du système tout en minimisant le nombre d'intrusions inopportunes.

5.3 Choix du modèle CBR

Quel modèle CBR devrions-nous utiliser pour concevoir ce type d'application? Rappelons que les trois principaux modèles sont le structurel, le conversationnel et le textuel.

Le modèle structurel exige du concepteur un modèle du domaine bien défini. Ce modèle doit contenir une représentation des principales entités du domaine ainsi que les relations qui les lient entre elles. A priori, il semble difficile de concevoir un tel modèle pour notre domaine puisque le domaine de discours est large et que les messages ne sont pas toujours très descriptifs. De plus, même s'il était possible d'élaborer un tel modèle, il faudrait être capable d'extraire des messages les occurrences de ce modèle (i.e. les instances d'entités et de relations du modèle). Une approche manuelle d'extraction nous semble peu viable. Pour les nouveaux messages, l'extraction manuelle du contenu risque de prendre plus de temps que de rédiger directement la réponse. De plus, des problèmes de passage à grande échelle ("scalability") sont à prévoir pour la construction de bases de cas contenant plusieurs milliers de messages. Une approche complètement automatique serait possible seulement si le contenu des textes présente une certaine structuration et une régularité, ce qu'on ne retrouve pas dans tous les messages de notre corpus.

Le modèle conversationnel repose sur l'hypothèse que l'usager du système est disponible pour répondre aux questions du système. Dans notre cadre applicatif, il y a deux possibilités:

- l'investisseur est l'usager: on peut imaginer un scénario où l'usager pourrait avoir accès au système via une page web du site de BCE. L'usager pourrait interagir avec le système afin de bien définir ses questions. Le système pourrait alors répondre sur-le-champ ou relayer la question aux analystes. Cette solution est intéressante mais implique un mode d'opération différent du travail actuel des analystes.
- l'analyste est l'usager: dans ce scénario, l'analyste peut soit répondre lui-même aux questions du système ou bien les relayer aux investisseurs. Y répondre soit-même semble peu avantageux car répondre à une série de questions pourrait être plus coûteux que de générer une réponse de quelques phrases seulement. D'autre part, relayer des questions aux investisseurs semble inutile si leur message contient déjà toute l'information nécessaire pour y répondre (voir section 4.3).

Eu égard aux observations précédentes, l'approche CBR textuel nous semble la plus adaptée à notre problème. Elle mise principalement sur une analyse du contenu textuel des cas pour trouver des solutions. Puisque les questions sont suffisamment claires pour permettre aux analystes de générer une réponse, une analyse (statistique ou sémantique) de leur contenu permettrait d'accomplir, en partie, cette tâche avec un système CBR. Contrairement à l'approche structurelle, l'approche textuelle permet d'exploiter les textes sans structuration à outrance de la base de cas. Contrairement à l'approche conversationnelle, l'analyste n'a pas à être sollicité durant la phase de recherche.

En résumé, l'approche textuelle nous semble la plus adéquate car elle ne modifie pas le mode d'opération actuel du service aux investisseurs (schéma question-réponse unique) et car elle exploite le contenu textuel des messages disponibles. De plus, comme ce modèle est relativement nouveau, il offre de nombreuses perspectives de recherche et il nous permettra d'apporter des contributions originales aux niveaux des techniques d'authoring de la base de cas et des approches d'adaptation des solutions textuelles.

6.0 Démarche et thèmes de recherche

Notre principal objectif est de concevoir un système CBR de réponse au courrier électronique. Pour atteindre cet objectif, nous devons surmonter deux difficultés: a) atteindre une grande précision dans le choix des messages antécédents et b) élaborer des méthodes pour modifier les réponses antécédentes.

Tel que discuté à la section 3, les approches CBR textuel proposées dans la littérature sont actuellement déficientes à ces deux niveaux. Nous proposons de les enrichir à l'aide de deux prototypes: un premier misant uniquement sur l'authoring et la recherche de cas; et un deuxième misant sur l'adaptation des textes.

Premier prototype: Etude de l'authoring et la recherche de cas

La phase de recherche CBR a fait l'objet de nombreux travaux, résultant ainsi en une grande variété de techniques. Ces techniques étant insuffisantes pour garantir de bons niveaux de précision, le développement de meilleures approches d'authoring de cas nous permettra de réaliser des gains.

Le processus de structuration est particulièrement important pour le CBR textuel. D'une part, c'est à cette étape qu'on tente de surmonter les difficultés causées par les textes peu structurés. D'autre part, nous pouvons miser sur des techniques de traitement de langue naturelle pour appuyer cette structuration.

Notre prototype nous permettra d'étudier deux facettes:

- d'un point de vue applicatif, il correspond au niveau "glaneur" qui permet à l'analyste de consulter la base de cas et qui lui recommande les messages antécédents pertinents. Nous pourrons donc évaluer la capacité du système à repérer les messages permettant à l'analyste d'accomplir sa tâche.
- d'un point de vue technique, il nous permettra de mieux cerner les caractéristiques du corpus de BCE, d'étudier une approche de structuration des cas et d'évaluer les performances de la phase de recherche du système CBR.

Le développement de ce prototype comporte les trois étapes suivantes:

- le choix de la granularité des cas textuels: ce thème est présenté à la section 6.1.
- la structuration des cas textuels: ce thème est présenté à la section 6.2.
- l'évaluation du système et de la base de cas: ce thème est présenté à la section 6.3.

Deuxième prototype: Etude de l'adaptation des réponses antécédentes

Le processus d'adaptation doit fournir des réponses aux questions suivantes: quelles solutions méritent d'être adaptées, quels portions/passages de ces cas devraient être modifiés, et comment les modifier?

Le prototype sera bâti à partir du premier prototype. Il permettra d'étudier les facettes suivantes:

- d'un point de vue applicatif, il correspond au niveau "rédacteur" qui permet de générer de manière autonome de nouvelles réponses. Nous pourrons donc estimer l'utilité et la pertinence des modifications apportées aux messages antécédents.
- d'un point de vue technique, il nous permettra d'étudier diverses approches pour l'adaptation des cas textuels et d'évaluer les performances de cette phase.

Pour réaliser ce prototype, nous devrons principalement consacrer nos efforts aux techniques permettant d'identifier les passages à adapter. Ce thème est présenté à la section 6.4.

Hypothèses de travail

Afin de délimiter notre cadre de travail, notre approche est basée sur un certain nombre d'hypothèses:

- les cas sont représentés par des ensembles de triplets (attribut, valeur, poids) où:
  - les attributs peuvent être des mots-clé, des termes composés, des catégories et autres étiquettes provenant d'une modélisation du domaine. Les attributs ne sont pas reliés entre eux;
  - les valeurs sont booléennes (e.g. indiquant la présence de mots-clé), numériques (e.g. des indicateurs financiers) ou symboliques (e.g. nom de filiales);
  - les poids varient sur l'intervalle [0,1] (importance croissante de l'attribut).
- la phase de recherche s'appuie sur une combinaison linéaire de mesures de similarité statistique et sémantique. La similarité statistique est déterminée par l'occurrence conjointe de mots-clé, de termes complexes ou de catégories. La similarité sémantique est déterminée par la distance mesurée entre deux termes dans une taxonomie du domaine. On suppose que cette taxonomie est disponible avant la conception du système CBR.
- L'adaptation du contenu des textes se fait par la modification de passages comportant des entités nommées, dates, valeurs numériques…. On suppose que les modifications préservent la structure syntaxique des phrases. Les modifications sont dépendantes du domaine et sont effectuées soit:
  - manuellement par l'analyste,
  - par l'application de requêtes dans une base de données (e.g. pour obtenir la valeur du ratio profit-bénéfice du dernier trimestre),
  - par l'application de règles (e.g. substitution d'entités nommées).

  On suppose que la forme des requêtes, des règles et des équations peut être déterminée manuellement.

6.1 Choix de la granularité d'un cas textuel

Lors de cette étape, nous devons déterminer si un texte donné, comportant la requête d'un investisseur et la réponse d'un analyste, correspond à un cas (voir figure 10). A prime abord, cette correspondance semble naturelle. Toutefois les messages ayant des questions multiples et des réponses génériques nous amènent à mettre ce choix en doute.

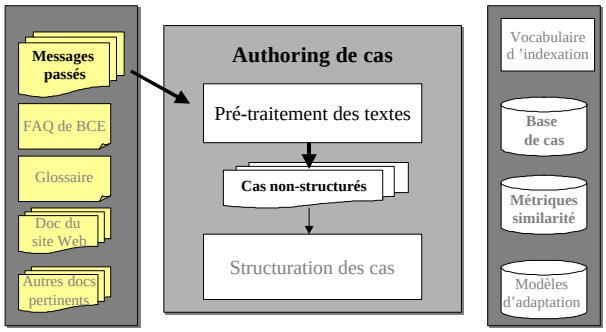


Figure 10 – Choix de la granularité des cas

Le traitement des messages aux questions multiples est particulièrement intéressant. Ce phénomène est peu fréquent dans les applications CBR. Habituellement les travaux en CBR émettent l'hypothèse du "single fault assumption", i.e. qu'une description ne concerne qu'un seul problème qui n'exige qu'une seule solution. Or, tel qu'indiqué à la section 4, certains de nos messages contiennent plusieurs thèmes. La Figure 11 illustre cette distribution de thèmes pour une vingtaine de messages tirés de notre corpus. Un **Q** indique la présence du thème dans la question, un **R** indique sa présence dans la réponse. On note que la distribution des thèmes dans le tableau n'est pas uniforme et que les messages forment des regroupements (représentés par les ellipses).

# Message	Date de divulgation	Appel conférence	Liste de distribution	Rumeurs financières	Information financière	Visite du Site web	Actionnaire	Autres questions	Remercie-ment	Échange de documents	Fluctuation des marchés	Infos sur les filiales
1		Q R							R			
2	Q R	Q R							R			
3	Q R	Q R							R			
4	Q R	Q R	Q						R			
5	Q R	Q R					Q		R			Q R
6	Q R	R							R			
7	Q R								R			
8	Q R								R			
9	Q			Q R				R	R			
10	Q R			Q								
11	Q R				Q R				R			
12			Q		Q R				R			
13					Q	R	Q	R	R			
14					Q R			R				
15					Q	R		R				
16						R		R		Q		
17										R		R
18					Q R	R		R				
19								R				Q R
20								R			Q R	

Figure 11 – Exemple de distribution des questions-réponses multiples et de regroupement de thèmes

Ceci nous amène à proposer, pour la conversion initiale des textes en cas, trois types de correspondence entre les textes de questions-réponses et les cas de la base:

- un cas correspond à un message qui peut contenir plusieurs questions et plusieurs réponses (correspondence 1:N);
- un cas ne peut contenir qu'une seule question et sa réponse correspondante (correspondence 1:1);
- des groupe de cas disjoints sont construits à partir de messages traitant d'un même thème (correspondence M:N, donc M cas couvrant N thèmes).

Correspondance 1: N

En premier lieu, nous retiendrons ce type de correspondance pour la conversion des textes en cas. Elle offre l'avantage que les textes sont directement exploitables sans transformation majeure. Toutefois, le désavantage est qu'il est difficile d'estimer la similarité entre messages à questions multiples. Reprenons un de nos exemples de la section 1 pour bien illustrer ce point.

> **Question2**: Hello, I am writing to find out when you are reporting the 2nd quarter earnings and to obtain the number if you are having a conference call. Thank you…

La Question2 contient deux sous-questions: une portant sur la date de divulgation de résultats financiers et l'autre portant sur l'appel conférence. Lors du processus de recherche, on note deux situations possibles:

a) la base de cas contient au moins un message portant sur ces deux thèmes. Nous avons alors la garantie qu'une réponse peut être construite à partir d'un seul cas. Cette situation ne pose pas de problème particulier.

b) la base de cas ne contient aucun message portant simultanément sur ces deux thèmes. Une réponse doit alors être construite à partir de plusieurs messages antécédents. Les difficultés surgissent lors de cette dernière situation.

Pour construire une réponse adéquate, la phase de recherche du système doit faire la recherche sur plusieurs messages et retenir au moins un de chaque thème. Or un critère de similarité n'offre aucune garantie que nous retrouverons des messages pour ces deux thèmes parmi les k plus proches voisins. Par exemple, si nous limitons le nombre de cas voisins à k = 2, il se peut que les deux messages portent sur un même thème au détriment de l'autre thème de la question. Si on augmente le nombre de voisins k, nous n'avons aucune garantie que des messages portant sur l'un ou l'autre des thèmes se retrouveront en tête de liste.

Si nous rencontrons cette difficulté dans nos travaux, nous explorerons la solution suivante. Elle consiste à modifier le critère de similarité afin de tenir compte de la diversité des cas. Ainsi, lors de la recherche, le premier cas retenu sera le plus similaire. Le second sera celui qui offre le meilleur compromis entre i) la similarité par rapport à la question, et ii) la distance par rapport au premier cas retenu. Nous espérons ainsi sélectionner un nombre limité de cas recouvrant un maximum de thèmes.

Si l'approche basée sur la diversité s'avère insuffisante, il faudra alors étudier les deux autres types de correspondance entre messages et cas.

Correspondance 1: 1

L'avantage de cette correspondance est que la similarité entre deux cas est établie sur la base d'un seul thème. Ainsi pour un nouveau message à questions multiples, des recherches individuelles sont menées pour chacune des sous-questions. On reporte ainsi la difficulté du traitement à la phase d'adaptation qui devra "recoller" les différentes réponses individuelles.

Cette approche nécessite l'étude des mécanismes pour découper les messages initiaux en sous-questions et en sous-réponses. Dans notre corpus, les réponses multiples sont habituellement réparties sur des phrases différentes. Toutefois, on retrouve des messages à questions multiples ayant les formes suivantes:

- les thèmes sont répartis sur des phrases différentes (la forme la plus fréquente).
- une même phrase contient une conjonction de sous-questions; la Question2 présenté dans cette section en est un exemple;
- une phrase contient une conjonction ou une énumération dans les syntagmes nominaux; par exemple, des questions du genre "j'aimerais obtenir la date de divulgation des prochains rapports financiers de Bell Canada International, Bell Emergis et Teleglobe".

Si cette étude s'avère nécessaire, nous aborderons le découpage des sous-questions à l'aide de techniques d'analyse syntaxique [Man99].

Correspondance M: N

Bien que nous n'envisagions pas cette option à ce stade-ci, il est intéressant de noter qu'une approche de construction de base de cas pourrait tirer profit des regroupements naturels de thèmes de messages (clusters). Dans notre exemple présenté à la figure 11, les regroupements correspondent aux ellipses.

Le regroupement des thèmes, combiné à une approche de segmentation, permettrait de réorganiser les messages originaux de manière plus compacte avec peu de redondance. Le désavantage est qu'après le "clustering", il peut s'avérer ardu de regrouper les diverses portions afin de produire des messages intelligibles.

6.2 Structuration de la base de cas

Dans cette étape, nous abordons le problème de la structuration des cas (figure 12) afin d'améliorer les performances de la phase de recherche du système CBR. La structuration consiste à convertir des textes en représentation de cas et à modifier cette représentation par le biais de différentes opérations.

Ce processus est subjectif et il est fort à parier que, pour un même corpus de textes, différentes personnes créeront différentes bases de cas. Les bases de cas varieront selon l'interprétation et l'importance attribuée à chacune des portions de textes. L'encodage de ces textes variera selon le schéma de représentation et le niveau de détail choisis.

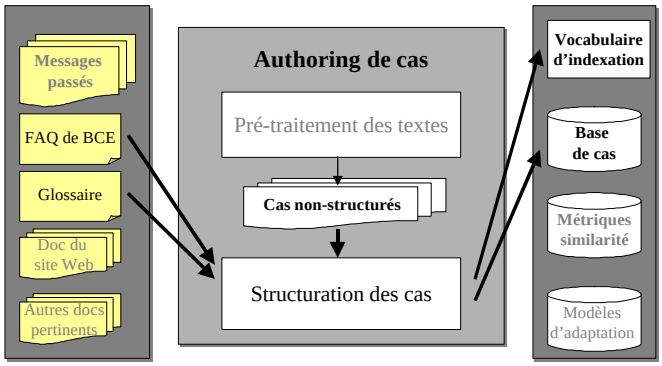


Figure 12 – Structuration de la base de cas

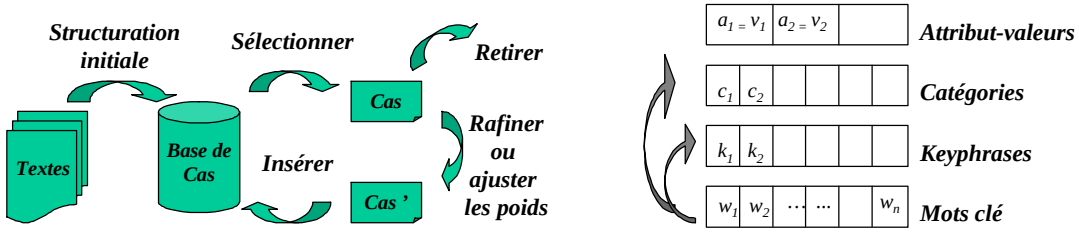
Ceci nous amène à nous demander si différentes bases de cas provenant d'un même corpus de textes sont équivalentes. Et pour une base de cas donnée, y-a-t-il des meilleures approches pour la construire?

Dans cette étape, nous proposons une approche qui guide le concepteur du système CBR dans sa démarche de structuration de la base de cas textuels. Cette approche vise ultimement à augmenter la précision de la recherche du système CBR. Pour ce faire, nous allons explorer trois aspects:

- l'algorithme de structuration
- les indicateurs pour guider la structuration.
- les opérateurs de structuration de cas;

6.2.1 Algorithme de structuration

L'approche de structuration que nous retenons est de type itératif (Figure 13-a). Celle-ci permet au concepteur d'entreprendre la structuration de la base de cas sans avoir eu à consulter et à analyser au préalable chacun des textes. De plus, nous n'imposons pas un niveau de structuration uniforme à tous les cas. Ceci permettra de mieux tenir de la distribution des messages sur les différents thèmes du domaine.



(a) (b)

Figure 13 – Structuration (a) étapes de structuration (b) raffinement de cas par niveau

La première étape consiste à obtenir une structure initiale de la base, i.e. une représentation simplifiée de chacun des cas. Les itérations subséquentes permettent d'augmenter progressivement le degré de structuration des cas. On peut donc conceptualiser l'approche de structuration comme une recherche vorace sur la base de cas. Un version simplifiée de l'approche, de type "leave-one-out", peut être décrite comme suit:

Structuration(*Textes*, *Sim*, *Eval*)
CB ← *structuration_initiale(Textes)*
Candidats ← {}
Tant que *éval*(*CB*) non satisfaisant
C ← *sélectionne_cas(CB)*
$CB \leftarrow CB - C$
Candidats← *recherche(C,CB, Sim)*
Si *redondant(C,Candidats)* alors laisser *C* hors de la base de cas CB
Sinon
Si *inconsistant(C,Candidats)* alors C' ← *raffiner(C)*;
Sinon C' ←*ajuster_poids(C)*.
$CB \leftarrow CB \bigcup C'$
Fin Tant que.

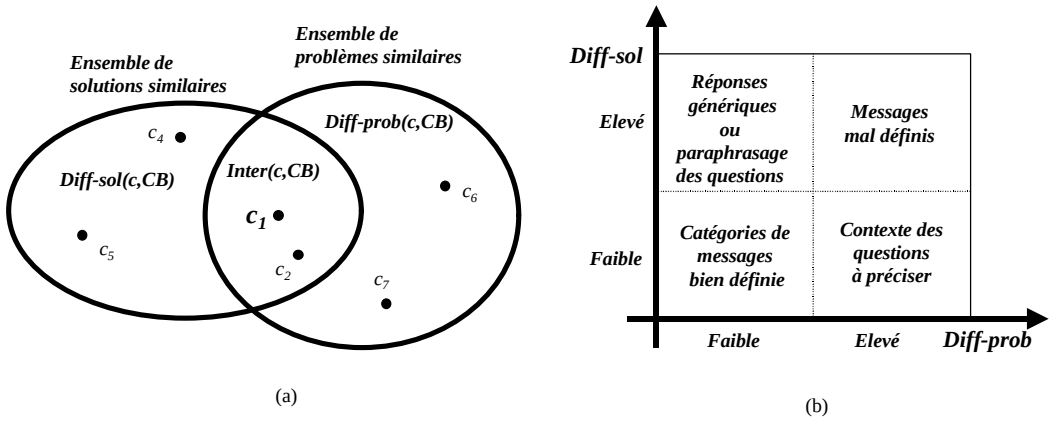
La structuration initiale des cas résultera en:

- une représentation vectorielle de ses mots-clés (avec des scores tf*idf);
- une catégorie de messages attribuée par un module de routage (projet de maîtrise de Julien Dubois);
- une représentation attribut-valeur de la date du message et du nom/adresse de l'investisseur.

Les itérations subséquentes s'appuient sur trois types d'opérateurs: retirer un cas de la base, raffiner les attributs d'un cas (voir figure 13-b), et ajuster les poids des attributs d'un cas.

6.2.2 Indicateurs pour guider la structuration

Des indicateurs sont nécessaires pour déterminer, à chaque itération de la démarche, les cas devant faire l'objet de modification (*sélectionne_cas*), le type de modification à apporter (*redondant*, *inconsistant*) et la pertinence de poursuivre la démarche (*éval*). Pour élaborer ces indicateurs, nous nous appuyons sur l'intuition qu'un cas mérite une meilleure structuration si sa solution ne peut pas être reconstruite à partir d'autres cas similaires.



(a) (b)

Figure 14 – Distribution de cas (a) ensembles de problèmes et/ou solutions similaires

(b) inteprétation des ensembles

Pour cerner plus précisément cette intuition, considérons deux ensembles liés à un cas c_1 (figure 14-a): l'ensemble des cas ayant des problèmes similaires et l'ensemble des cas ayant des solutions similaires. Ces deux espaces sont définis comme suit:

$$E_{\text{problème}}(c_1, CB) = \{ c \in CB : \text{sim}(c_1^{\text{prob}}, c^{\text{prob}}) > d_{prob} \}$$

$$E_{\text{solution}}(c_1, CB) = \{ c \in CB : \text{sim}(c_1^{\text{sol}}, c^{\text{sol}}) > d_{sol} \}$$

où *sim* est la similarité globale entre deux cas. A partir de ces ensembles, on détermine trois groupes de cas:

- Les cas ayant des solutions et des problèmes similaires à c_1

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

- Les cas ayant uniquement des problèmes similaires à c_1

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

- Les cas ayant uniquement des solutions similaires à c_1

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

Tel qu'illustré sur la figure 14-b, on peut considérer quatre situations qui nous aident à relier ces définitions à notre problème de réponse au courrier électronique:

- *Inter* contient la majorité des cas: ceci indique que la base contient des questions similaires à c_1 et que ces questions nous amènent vers des réponses analogues à celles de c_1. Donc c_1 appartient à une catégorie de questions relativement bien définies et exprimées avec un groupe restreint de mots. Les messages de la section 1 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 le retirer de la base de cas sans grande incidence. Sinon on peut faire un ajustement des poids de ses attributs.
- *Diff_sol* contient la majorité des cas: autrement dit, des solutions similaires à celle de c_1 sont utilisées pour différents problèmes. Par rapport à notre application de réponse au courrier électronique, ceci peut se traduire par deux possibilités:

- ces réponses sont génériques et sont appliquées à des questions différentes; des tests de généricité des messages tels que proposés par [Kos01] peuvent nous aider à valider cette hypothèse; Pour ce type de cas, nous jugeons qu'un réajustement des poids est suffisant.
- ces réponses font suite à des questions lexicalement différentes mais sémantiquement similaires. Puisque les questions proviennent de différents investisseurs, ceci suggère des textes paraphrasés. Dans cette éventualité, il est nécessaire de raffiner le contenu des questions. Puisque les réponses ne permettent pas de discriminer entre les questions, une catégorisation des messages serait donc difficile. Nous proposons alors un raffinement par extraction d'information.

- *Diff_prob* contient la majorité des cas: cette situation indique que des questions similaires donnent lieu à des réponses exprimées différemment. A la section 4, nous avons observé que les réponses sont rédigées par un groupe restreint d'analystes et que leurs textes sont assez uniformes. Donc si les réponses sont exprimées différemment, c'est qu'elle sont différentes. Il est alors nécessaire de mieux cerner le contexte des questions. Comme les réponses varient d'un message à l'autre, elle peuvent servir de base pour la catégorisation des questions.
- Les ensembles *Diff_prob* et *Diff_sol* contiennent 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 l'un des deux ensembles.

La base de cas résultant du processus de structuration sera évaluée en terme de mesures précision-rappel de la phase de recherche. De plus, nous comparerons les résultats des opérateurs de structuration avec ceux obtenus manuellement par un humain.

Afin de rendre opérationelle l'approche de structuration proposée, il sera nécessaire de fournir un certain nombre de mesures quantitatives. Nous présentons certaines de ces mesures à la section 6.3.

6.2.3 Opérateurs de structuration de cas

A ce stade-ci de nos recherches, nous considérons les techniques suivantes pour les opérateurs de structuration de cas:

- modification des poids: nous utiliserons une approche de type "pseudo-relevance feedback" où les cas appartenant à *Inter* influencent positivement les termes du cas c_1 tandis que les cas des deux autres ensembles exercent une influence négative;
- extraction d'information: le RALI dispose d'un certain nombre de logiciels qui nous permettront d'entreprendre cette tâche (Alembic, EXIBUM, TextPro). De plus, si le nombre de cas similaires est suffisamment grand, nous utiliserons une approche adaptative de recouvrement par règles (voir section 3.8.3).
- catégorisation: nous utiliserons un algorithme de classification supervisée de type KNN, ce qui nous permettra d'exploiter certaines portions de notre implantation CBR.
- Termes composées ("keyphrases"): nous utiliserons soit une approche non-supervisée par analyse des co-occurences, ou une approche supervisée basée sur la catégorisation de texte (voir section 3.8.1).

6.3 Evaluation de la base de cas et du système CBR

Tel qu'illustré sur la figure 15, nous développerons des indicateurs qui nous permettront d'évaluer nos prototypes selon différents axes:

Prototype 1:

- qualité de la base de cas (évaluation 1)
- performance du processus de recherche de cas (évaluation 2)

Prototype 2:

- performance du processus d'adaptation (évaluation 3)

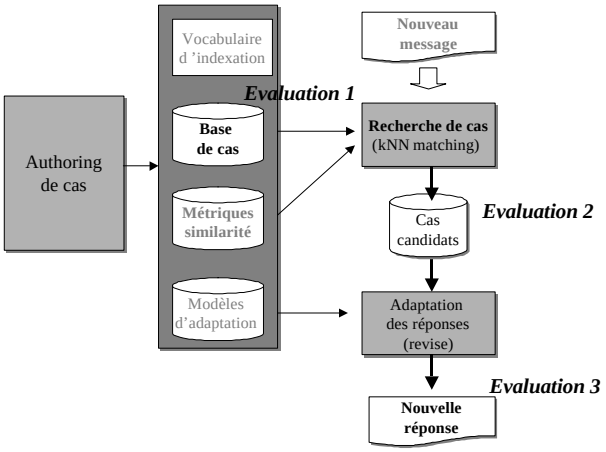


Figure 15 – Points d'évaluation du système CBR

Les paragraphes suivants donnent quelques précisions sur les indicateurs qui seront développés dans le cadre de notre projet.

Evaluation 1 - Qualité de la base de cas

La qualité de la base de cas peut être mesurée par rapport à un certain nombre de propriétés dont les suivantes:

- Rectitude ("correctness"): un cas est correct si sa solution permet de résoudre le problème qui y est décrit. Habituellement, cette propriété est supposée vraie dans la plupart des systèmes. Tel est le cas pour notre application de réponse au courrier électronique.
- Consistance: des cas ayant les mêmes descriptions de problème ne devraient pas avoir de solutions différentes.
- Unicité: un cas est unique si aucun autre cas de la base ne contient simultanément les mêmes descriptions de problème et de solution.
- 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).
- Cohérence: des cas sont cohérents si, pour une même solution, la description des problèmes ne varie pas selon un certain nombre d'attributs-valeurs.

Par rapport à la démarche de structuration proposée à la section précédente, ces propriétés nous permettrons de définir un certain nombre de fonctions et d'indicateurs. Afin d'illustrer l'utilisation de ces propriétés, nous donnons les exemples suivants:

redondant(C,Candidats): cet indicateur est basé sur la propriété d'unicité. On peut l'exprimer comme suit:

$$\max_{C' \in Candidats} \left(\frac{\left| A_C^{prob} \cap A_{C'}^{prob} \right|}{2 \times \max\left(\left| A_C^{prob} \right|, \left| A_{C'}^{prob} \right| \right)} + \frac{\left| A_C^{sol} \cap A_{C'}^{sol} \right|}{2 \times \max\left(\left| A_C^{sol} \right|, \left| A_{C'}^{sol} \right| \right)} \right)$$

où A_C^{prob} et A_C^{sol} sont respectivement les attributs de la description du problème et de la solution de C. On pourrait également pondérer les expressions par le poids des attributs.

consistant(C,Candidats): $$\min_{C' \in Diff\_Prob(C,Candidats)} \left(\frac{\left| A_C^{sol} \cap A_{C'}^{sol} \right|}{\max\left(\left| A_C^{sol} \right|, \left| A_{C'}^{sol} \right| \right)} \right)$$

cohérent(C,Candidats): $$\min_{C' \in Diff\_Sol(C,Candidats)} \left(\frac{\left| A_C^{prob} \cap A_{C'}^{prob} \right|}{\max\left(\left| A_C^{prob} \right|, \left| A_{C'}^{prob} \right| \right)} \right)$$

D'autres définitions seront nécessaires pour formuler les fonctions de sélection de cas (*sélectionne_cas)* et d'évaluation de base de cas (*éval*).

Evaluation 2 – Performance du processus de recherche

La distribution des cas est un premier indicateur de la performance du processus de recherche. Idéalement, on souhaiterait que les cas soient distribués uniformément dans l'espace des questions. Ceci nous assurerait qu'il existe, peu importe la nouvelle question à traiter, des réponses qui s'apparentent à cette dernière. On peut mesurer la distribution d'une base CB par sa densité de cas:

$$densité_{prob}(CB) = \frac{\sum_{c \in CB} \sum_{c' \in CB} sim\left(c_{prob}, c'_{prob}\right)}{|CB|^2}$$

La qualité des résultats de la recherche est un deuxième indicateur de performance. Cet indicateur détermine la proportion de questions pour lesquelles on trouve une réponse adéquate. Comme dans les systèmes de recherche ou d'extraction d'information, la qualité est mesurée par le taux de rappel et de précision. Toutefois, la définition de ces mesures varie selon qu'un seul cas (un seul thème) ou que plusieurs cas (plusieurs thèmes) sont nécessaires pour générer une solution adéquate.

Cas ayant:	**Précision**	**Rappel**
un seul thème	*1* Si le premier cas porte sur ce thème *0* Sinon	*1* Si au moins un cas porte sur ce thème *0* Sinon
N thèmes	*m/N* Si on retrouve m de ces thèmes parmi **les N premières** réponses retournées, $0 \leq m \leq N$	*m/N* Si on retrouve m de ces thèmes parmi **toutes les réponses** retournées, $0 \leq m \leq N$

La discussion sur l'évaluation du processus d'adaptation (Evaluation 3) est reporté à la section 6.4 .

6.4 Adaptation de cas textuels

Tel que mentionné à de nombreuses reprises, l'adaptation de cas textuels n'a pas encore été abordé dans la littérature CBR. Les principaux travaux en adaptation sont effectués dans le cadre du modèle CBR structurel qui offre l'avantage de décrire le problème et la solution d'un cas à l'aide d'attributs clairement identifiés. Il est alors possible de construire des modèles qui guident la modification des attributs de la solution à partir des attributs explicatifs du problème.

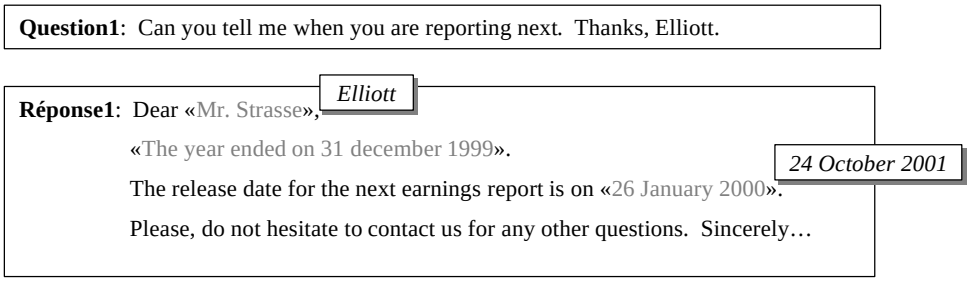


Figure 16 – Exemple de réponse modifiée

Pour des cas textuels, la modification des textes est possible mais doit être précédée d'un processus qui détermine ce que l'on doit modifier. Reprenons l'un de nos exemples de la section 1 (Figure 16) pour illustrer la nature des passages à modifier.

- Premièrement, on remarque que des noms de personne (*Mr. Smith*) et des dates (*26 January 2000*) sont très spécifiques au contexte du message antécédent. Habituellement ces informations doivent être modifiées pour de nouvelles requêtes, indépendamment de leur contexte. On peut généraliser cette observation aux entités nommées (lieux, organisations...) et aux descriptions numériques (e.g. numéro de téléphone).
- Deuxièmement, la pertinence de certains passages des réponses varient selon le contexte ou la temporalité de la requête. Par exemple, le passage sur la date de fin d'année fiscale (*The year ended…*) devient inutile puisque la nouvelle requête ne porte plus sur ce thème.

Ainsi nous proposons de découper la phase d'adaptation en 3 étapes (figure 17): i) l'extraction des entités nommées et des valeurs numériques, ii) l'identification des passages pertinents, iii) la substitution et l'élagage des passages à modifier.

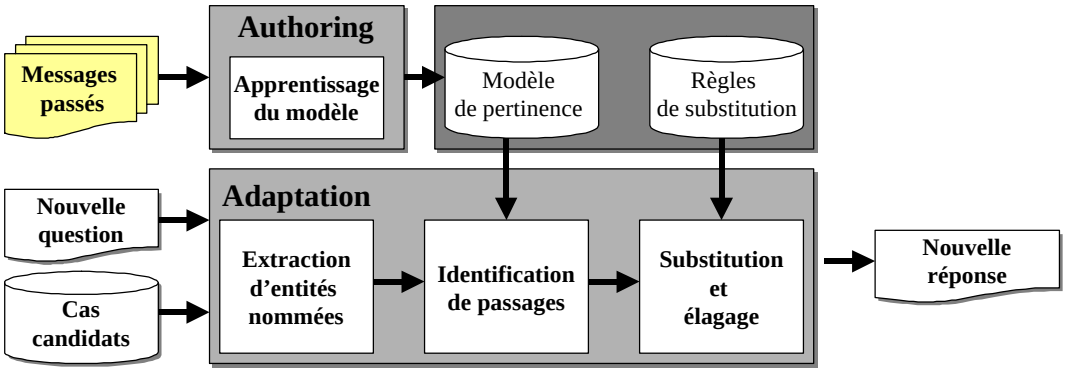


Figure 17 – Etapes du processus d'adaptation

Pour la première étape, le laboratoire RALI dispose d'un certain nombre de logiciels qui nous permettront de localiser les entités nommées de notre corpus de BCE.

La deuxième étape s'avère plus complexe, car elle exige de déterminer la pertinence de passages à partir de liens possibles entre questions et réponses. Or certaines techniques de traitement de la langue naturelle pourraient nous permettre de modéliser les similitudes entre des paires de questions-réponses, et les différences entre diverses réponses (ce qui varient d'une solution à l'autre). Les modèles résultants seraient exploités par la suite pour "étiqueter" les portions de textes selon leur pertinence/utilité. Nous étudions actuellement l'applicabilité de deux approches:

- "Relevance-based language models": ces travaux du domaine de la recherche d'information [Cro01] visent à proposer des modèles de langues pour estimer la pertinence d'un document étant donnée une requête (ou l'inverse). Notre intuition est que ces modèles ainsi que le critère d'ordonnancement "log-odds ratio" pourraient nous permettre d'établir la pertinence d'une portion de réponse.
- techniques "memory-based": des travaux de l'équipe de Dealemans [Dea00] ont démontré que ces techniques, analogue au raisonnement à base de cas structurels, offrent des performances intéressantes pour différentes tâches d'étiquétage ("POS tagging", "chunking"). Cette approche offre l'avantage d'être applicable malgré un nombre limité d'instances (le contenu de la mémoire du système).

Il est toutefois difficile à ce stade-ci de nos recherches d'établir la viabilité de ces approches. De plus amples travaux seront nécessaires pour mieux cerner les contributions potentielles de ces techniques à l'adaptation de cas textuels.

Tel que mentionné au début de la section 6, la troisième étape consiste à appliquer un certain nombre de règles qui permettront la modification des entités nommées et des valeurs numériques. Ces règles sont propre au domaine d'application et seront construites manuellement.

Evaluation 3 – Performance du processus d'adaptation - reporté à la section 6.4.

La couverture des cas est un indicateur de performance du processus d'adaptation. Tout comme pour la distribution des cas, cette mesure peut être formulée par la densité de ses solutions:

$$densité_{sol}(CB) = \frac{\sum_{c \in CB, c' \in CB} \sum sim(c_{sol}, c'_{sol})}{|CB| \times |CB| - 1}$$

Finalement, la justesse ("accuracy") du processus d'identification de passages peut être déterminée à partir de la proportion de mots étiquétés correctement.

7. 0 Echéancier des travaux

Nous proposons l'échéancier suivant pour effectuer nos travaux:

Période	Travaux
Sept 00	Début du doctorat
Avril 01	Examens écrits de connaissances générales
Déc 01 – Août 02	Elaboration du premier prototype de gestion de réponse, étude sur la granularité des cas et expérimentation avec l'approche de structuration.
Sept 02 – Mar 03	Etude sur l'identification de passages textuels à adapter, élaboration du deuxième prototype et évaluation globale du système.
Avril 03 – Août 03	Rédaction de la thèse et soutenance.

8.0 Conclusion

Nous avons présenté notre approche pour résoudre le problème de gestion de réponse au courrier électronique. Cette approche repose sur des techniques de raisonnement à base de cas textuel permettant d'exploiter un corpus de messages antécédents et de proposer des réponses à de nouveaux messages.

Notre revue de littérature du CBR textuel nous a permis de déterminer que les approches rapportées dans la littérature comportent des lacunes au niveau de la création de la base de cas textuels ("authoring" de cas) et au niveau de l'adaptation des solutions textuelles.

Le processus d'authoring de cas est important pour la performance du système et la qualité des solutions obtenues. Pour un même corpus de texte, il est possible de créer différentes bases de cas qui varient selon le nombre de cas et la structuration de chacun de ces cas. Une démarche permettant de guider ces choix s'impose pour les concepteurs de systèmes CBR textuel.

Pour ce qui est de l'adaptation, les approches actuelles du CBR textuel ne proposent pas de techniques dans ce sens. Deux constats expliquent cette lacune: leurs tâches n'exigent pas la modification du contenu des solutions, et leurs travaux exploitent peu les techniques de traitement de la langue naturelle (NLP). Par contre, il est crucial pour notre application de pouvoir réutiliser et modifier des réponses antérieures. D'où notre motivation de défricher la voie de l'adaptation des cas textuels.

Les contributions originales que nous entendons apporter se situent principalement au niveau:

- de la démarche de structuration des textes;
- de l'identification de portions de textes à adapter; et
- de la caractérisation de critères d'évaluation de bases de cas textuels.

En résumé, nous croyons que l'insertion de techniques NLP s'averera fructueuse et permettra d'étendre l'application des techniques de CBR textuel à des tâches de résolution de problème plus riches que la catégorisation de problème et la recherche de documents pertinents. Et ainsi de passer de la recherche au raisonnement.

Références

[Aam94] Aamodt, A., Plaza, E.; Case-base reasoning : foundational issues , methodological variations, and system approaches, AI-Communications, 7(1), 1994.

[Aha01] Aha, D.W., Breslow, L.A., & Muñoz-Avila, H. (2001). Conversational case-based reasoning. Applied Intelligence, 14, pp. 9-32.

[Ale96] Aleven, V., and K. D. Ashley. How Different is Different? Arguing about the Significance of Similarities and Differences. In Advances in Case-Based Reasoning: Proceedings of the Third European Workshop, EWCBR-96, edited by I. Smith and B. Faltings, 1-15. Lecture Notes in Artificial Intelligence, 1168. Berlin: Springer Verlag, 1996.

[Bra96] Branting, L. Karl and Lester, James. Justification Structures for Document Reuse. Proceedings of the Third European Workshop on Case-Based Reasoning (EWCBR-96), Lausanne, Switzerland, November 14-16, 1996, Lecture Notes in Artificial Intelligence 1168, pp. 76-90.

[Bru01] Stefanie Brüninghaus and Kevin D. Ashley (2001) The Role of Information Extraction for Textual CBR. To appear in: Aha, D.W., Watson, I. & Yang, Q. (Editors). Case-Based Reasoning Research and Development: Proceedings of the 4th. International Conference on Case-Based Reasoning (ICCBR-01). Vancouver, Canada, 30 July - 2 August 2001. Springer, Lecture Notes in Artificial Intelligence.

[Bru99] Stefanie Brüninghaus and Kevin D. Ashley (1999) Bootstrapping Case Base Development with Annotated Case Summaries In: Klaus-Dieter Althoff, Ralph Bergmann and L. Karl Branting (editors). Case-Based Reasoning Research and Applications. Proceedings of the Third International Conference on Case-Based Reasoning (ICCBR-99) Kloster Seeon, Muenchen, Germany. Lecture Notes in Computer Science 1650, Springer Verlag, Heidelberg, Germany. This paper won the Outstanding Research Paper Award.

[Bru97] Stefanie Brüninghaus and Kevin D. Ashley (1997) Finding Factors: Learning to Classify Case Opinions Under Abstract Fact Categories. In: L. Karl Branting (editor). Proceedings of the 6th International Conference on Artificial Intelligence and Law (ICAIL-97). Pages 123-131. Melbourne, Australia. Copyright 1997 by ACM, Inc.

[Bur95] Robin Burke, Kristian Hammond, and Julia Kozlovsky. “Knowledge-based Information Retrieval for Semi-Structured Text”, In Working Notes from AAAI Fall Symposium on AI Applications in Knowledge Navigation and Retrieval, pages 19-24. American Association for Artificial Intelligence, 1995.

[Bur97] Burke, R., Hammond, K., Kulyukin, V., Lytinen, S., Tomuro, N. & Schoenberg, S. Question Answering from Frequently-Asked Question Files: Experiences with the FAQ Finder System. AI Magazine, 18(2), pages 57-66, 1997

[Car93] Cardie, C. (1993) Using decision trees to improve case-based learning. In Proceedings of the Tenth International Conference on Machine Learning, pp. 25-32

[Car96] Cardie, C. (1996) Automating Feature Set Selection for Case-Based Learning of Linguistic Knowledge. C. Cardie. Proceedings of the Conference on Empirical Methods in Natural Language Processing, pp. 113-126, University of Pennsylvania, 1996

[Cir00] F. Ciravegna, R. Basili, R. Gaizauskas, Proceedings of the Workshop on Machine Learning for Information Extraction, 2000, http://www.isi.edu/~muslea/

[Cir01] F. Ciravegna, N. Kushmerick, R. Mooney, I. Muslea; Proceedings of IJCAI-2001 Workshop on Adaptive Text Extraction and Mining, 2001, http://www.smi.ucd.ie/ATEM2001/

[Cro01] Croft, B., Callan, J., Lafferty, J.; (2001) Workshop on Language Modeling and Information Retrieval, May 31-June 1, Carnegie Mellon University, Pittsburgh, Pennsylvania, USA, http://la.lti.cs.cmu.edu/callan/Workshops/lmir01/

[Dan96] Daniels, J. 1996. Retrieval of Passages for Information Reduction. PhD Thesis, University of Massachusetts, Amherst.

[Dea00] W. Daelemans, J. Zavrel, K. van der Sloot, and A. van den Bosch, (2000) TiMBL: Tilburg Memory Based Learner - version 4.0 - Reference Guide, ILK Technical Report ILK01-04, 2000.

[Dia01] Diaz-Agudo, Gonzalez-Calero; (2001) Knowledge intensive CBR made affordable, Workshop on Workshop On Case-Based Reasoning Authoring Support Tools, ICCBR'01.

[Fuc01] Fuchs, B., Mathon, A., Mille, A.; (2001) Representing CBR knowledge with the Rocade System, Workshop on Workshop On Case-Based Reasoning Authoring Support Tools, ICCBR'01.

[Gar00] Gartner Research (2000), E-mail Response Management: Perspective, 17 mars 2000, http://cnscenter.future.co.kr/resource/rsc-center/gartner/email.pdf.

[Gut99] Gutwin, C., Paynter, G.W., Witten, I.H., Nevill-Manning, C., and Frank, E. (1999) Improving browsing in digital libraries with keyphrase indexes, Decision Support Systems, 27(1-2), pp. 81-104.

[Gre92] G. Grefenstette (1992). Use of syntactic context to produce term association lists for text retrieval, Proceedings of SIGIR'92, Copenhagen, Denmark.

[Han00] Han, J., Kamber, M.; (2000) Data Mining: Concepts and Techniques, Morgan Kaufmann Publishers, August 2000.

[Kol93] Kolodner, J.; (1993) "Case-Based Reasoning", Morgan Kaufmann.

[Kos01] 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), pp. 357-371, Tours, France.

[Kos98] 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. Novembre. Sfax, Tunisie.

[Len99] Lenz, Mario and Glintschert, Alexander: On Texts, Cases, and Concepts., Proceedings of XPS-99, Springer Verlag, LNAI

[Len98] Lenz, Mario, Bartsch-Spörl, Brigitte, Burkhard, Hans-Dieter, Wess, Stefan (Eds.): Case-Based Reasoning Technology - From Foundations to Applications. Lecture Notes in Artificial Intelligence 1400, Springer Verlag, 1998

[Len97] Lenz , Mario, Burkhard, Hans-Dieter, CBR for Document Retrieval - The FAllQ Project. in: D. Leake, E. Plaza (Eds.): Case-Based Reasoning Research and Development, Springer Verlag, LNAI 1266, 1997.

[Lea99] Leake, David B. and Wilson, David C.. (1999). "Combining CBR with Interactive Knowledge Acquisition, Manipulation and Reuse." In Proceedings of the Third International Conference on Case-Based Reasoning. pp. 203-217. Berlin. Springer-Verlag.

[Lea96] Leake, D. B., editor. (1996) Case-Based Reasoning: Experiences, Lessons, and Future Directions. Menlo Park, CA: AAAI Press/MIT Press, Menlo Park, CA.

[Man99] Manning, C., Schütze, H., (1999) Foundations of Statistical Natural Language Processing, The MIT Press, Cambridge, Massachusetts.

[Mar00] Marquez, L. Padro, H. Rodriguez", "A Machine Learning Approach to {POS} Tagging", Machine Learning", 39(1), pp. 59-91, 2000

[McS01] McSherry, D.; (2001) Improving the build quality of CBR systems: the case-authoring challenge, Workshop on Workshop On Case-Based Reasoning Authoring Support Tools, ICCBR'01.

[Mil01] Miller,R., Myers, B., (2001) Outlier Finding: Focusing User Attention on Possible Errors, UIST 2001, Orlando, FL, November 2001

[Mul01] Mullins, M., Smyth, B.; (2001) Visualisation Methods in CBR, Workshop on Workshop On Case-Based Reasoning Authoring Support Tools, ICCBR'01.

[Rac97] Racine, K., & Yang, Q. (1997). Maintaining unstructured case bases. Proceedings of the Second International Conference on Case-Based Reasoning (pp. 553--564)

[Rie89] Riesbeck, C., Shank, R., "Inside Case-Based Reasoning", Lawrence Erlbaum Associates, 1989.

[Ris95] Rissland, Edwina L., and Daniels, Jody J. Using CBR to Drive IR. In the Proceedings of the 14th International Joint Conference on Artificial Intelligence, Montreal, Canada, 1995.

[Ril96] Riloff, E. (1996) "Automatically Generating Extraction Patterns from Untagged Text", Proceedings of the Thirteenth National Conference on Artificial Intelligence (AAAI-96) , 1996, pp. 1044-1049.

[Roa98] 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.

[Sch00] Schreiber, A., Akkermans, J., Anjewierden, A., de Hoog, A., Shadbolt, N., Van de Velde, W., Wielinga, B; (2000) Knowledge Engineering and Management: The CommonKADS Methodology, MIT Press.

[Shi01] Hideo Shimazu, H., Kusui, D.; (2001) Detecting Defect Sign Cases, ICCBR'01, pp. XX-XX.

[Shi01] Kusui, D., Hideo Shimazu, H.; (2001) Transforming Mailing Lists into Case Bases, ICCBR'01, pp. XX-XX.

[Sod99] Soderland, S. (1999). Learning information extraction rules for semi-structured and free text. Machine Learning, 34

[Smy99] Smyth, B. & McKenna, E. (1999). Building Compact Competent Case-bases. Proceedings of the third International Conference on Case-based Reasoning. Munich, Germany. pp. 329-342. Springer Verlag

[Tur00] Turney, P.D. (2000), Learning algorithms for keyphrase extraction, Information Retrieval, 2 (4), pp. 303-336.

[Qia00] 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 Montreal Canada, May 2000.

[Wat98] Ian Watson, I, (1997) Applying Case-Based Reasoning: Techniques for Enterprise Systems, Morgan Kaufmann Publishers Inc., July 1997

[Web98a] Weber, R., Martins, A., and Barcia, R. (1998). On legal texts and cases. In M. Lenz and K. Ashley eds. Textual Case-Based Reasoning: Papers from the AAAI-98 Workshop, 40-50. Technical Report, WS-98-12. Menlo Park, CA: AAAI Press.

[Web98b] Weber, R.; Martins, A.; Mattos, E.; Bueno, T.; Hoeschl, H.; Pacheco, R.; Barcia, R. (1998). Reusing Cases to the Automatic Index Assignment from Textual Documents. 6th German Workshop on Case-Based Reasoning - Foundations, Systems, and Applications. Berlin, March 6-8, 1998.

[Wil00] Wilson, David C. and Bradshaw, Shannon. (2000). "CBR Textuality." Expert Update. Vol. 3., No. 1. pp. 28-37