Université de Montréal

Classification automatique de courrier électronique

par

Julien Dubois

Département d'informatique et de recherche opérationnelle

Faculté des arts et des sciences

Mémoire présenté à la Faculté des études supérieures

en vue de l'obtention du grade de M. Sc.

en informatique

juin, 2002

Université de Montréal
Faculté des études supérieures

Ce mémoire intitulé :

Classification automatique de courrier électronique

présenté par :
Julien Dubois

a été évalué par une jury composé des personnes suivantes :

Miklós Csürös
président-rapporteur

Guy Lapalme
directeur de recherche

Balázs Kégl
membre du jury

Résumé

Ce mémoire porte sur l'aspect classification de Merkure, un projet de réponse automatique au courrier électronique élaboré pour l'entreprise BCE. À la suite de l'analyse des corpus de courriels de BCE, nous avons expérimenté avec quelques classifieurs (Bayes, kppv, Ripper) et plusieurs formes de prétraitement (troncature des mots et élimination des numéraux, des mots vides de sens et des mots rares). En dépit des difficultés générées par nos corpus, nous avons atteint une efficacité de 80% sur notre classification finale. Dans le cadre de Merkure, ces résultats sont bons puisque d'autres modules utilisant des techniques plus approfondies viendront compléter le module de classification.

Mots-clés: classification, courrier électronique, traitement des langues naturelles, intelligence artificielle.

Abstract

This masters thesis presents the classification part of Merkure, an automatic email answering project designed for the BCE company. Following the analysis of BCE email corpora, we experimented with a few classifiers (Bayes, knn, Ripper) and many ways of preprocessing (word stemming, and removal of numerals, stop words and rare words) the data. In spite of many problems arising from the corpora, we achieved 80% efficiency on our final classification. Within the framework of Merkure, these results are good, knowing that other modules using more sophisticated methods will complement the classification module.

Keywords: classification, email, natural language processing, artificial intelligence.

À mes parents, Monique et Roger

Remerciements

L'auteur tient à remercier Guy Lapalme, pour ses conseils, Leila Kosseim, pour ses idées, et les autres membres du laboratoire RALI, en particulier Philippe et les 3 Luc. Merkure a aussi été rendu possible grâce à la collaboration financière des Laboratoires Universitaires Bell (LUB) et du Conseil de recherche en sciences naturelles et en génie du Canada (CRSNG).

Table des Matières

Liste des Figures

Liste des Tables

Chapitre 1

Introduction

1.1 Le courrier électronique

Parmi la gamme des nouvelles réalités que rend possibles l'internet, le courrier électronique est sans doute celle qui changera le plus nos habitudes de vie. Tous s'accordent pour dire qu'il s'agit du "killer-app of the internet" et avec raison, car la croissance de l'internet est directement reliée à l'importance grandissante du courrier électronique. Plusieurs sites web lui sont maintenant consacrés[1] et il est même possible d'obtenir un diplôme entièrement en ligne[2]. Presque tous les gens qui ont accès à l'internet ont au moins une adresse de courrier électronique qu'ils vérifient quotidiennement. D'après la firme IDC, environ 5 milliards de courriels sont envoyés chaque jour, et ce chiffre atteindra les 15 milliards d'ici la fin de l'année 2002.

En le comparant aux autres façons de communiquer (par écrit, par téléphone et en personne), on s'aperçoit que les nombreux avantages du courrier électronique surpassent de loin ses inconvénients. Sa grande force réside dans son médium de transport. La rapidité à laquelle les courriels circulent, combinée à la possibilité de les envoyer à plusieurs personnes en même temps, améliore considérablement la productivité des groupes de travail séparés par des endroits différents et des fuseaux horaires opposés. Grâce à l'internet, le meilleur moyen pour transmettre rapidement et efficacement un

[1] email.about.com, www.emailtoday.com, www.junkemail.org

[2] www.msubillings.edu, www.rsu.edu

message vers une autre partie du globe est d'envoyer un courriel. Une rencontre en personne est souvent impossible, un envoi postal peut prendre plusieurs jours, et un appel téléphonique peut coûter très cher. Il est vrai qu'il est difficile de calculer exactement les frais reliés à un courriel, mais la majorité des gens n'achètent pas un ordinateur uniquement pour le courrier électronique et considèrent ce service comme gratuit, ou presque.

La nature informatique des courriels offre une gamme d'avantages incomparables, dont l'envoi de documents électroniques en attachement. La conservation et l'archivage des messages sont beaucoup plus faciles à effectuer qu'avec les communications écrites et téléphoniques. De plus, seul le courrier électronique permet d'effectuer un traitement rapide, efficace et automatique sur les messages comme la recherche par mots clés, le triage automatique par sujet et le filtrage des messages importants. Ces fonctionnalités sont tellement utilisées que certaines personnes vont même jusqu'à se servir du courrier électronique comme aide-mémoire et pour la planification de tâches. On parle alors de la surcharge du courrier électronique [48].

Malheureusement, il n'y a pas que des avantages au courrier électronique. Le défaut principal qu'on lui reproche est de ne pas transmettre toute la gamme d'informations visibles et audibles dont dépendent inconsciemment les gens dans une rencontre en personne [17]. On lui attribue aussi une facilité à engendrer de mauvaises interprétations, à provoquer des échanges violents et à entretenir un effet de dépersonnalisation [26, 43]. En plus, un grand nombre de personnes développent une personnalité électronique complètement différente pour leurs communications en ligne, et ce phénomène va même jusqu'à l'émergence de plusieurs personnalités dans certains cas [31]. Ces défauts sont beaucoup moins présents dans le courrier traditionnel parce que ce dernier est utilisé principalement pour des échanges officiels nécessitant le respect de certains protocoles.

Pour les entreprises, le courrier électronique apporte une toute nouvelle dimension au service à la clientèle. D'après Forrester Research[3], 38% des compagnies considèrent très importante l'utilisation du courrier électronique, en plus des 32% qui la jugent importante. Du point du vue du marketing, les compagnies peuvent rejoindre plus

[3] www.forrester.com

facilement leurs clients et leur envoyer des annonces de produits, des offres spéciales, etc. à moindre coût. Elles peuvent également personnaliser le contenu selon les préférences de chaque client, et fournir l'information en totalité ou en partie, avec des hyperliens pour avoir plus de détails [10]. Mais plus important encore, le courrier électronique offre la possibilité aux entreprises de se rendre plus accessibles à leurs clients.

Lorsque vient le temps pour un client de contacter une entreprise, il peut généralement le faire par téléphone et aboutir sur un système d'aiguillage automatique. Après quelques choix qui ne sont pas toujours évidents, le client se fait mettre en attente pendant plusieurs minutes supplémentaires avant de pouvoir parler à un préposé. Une alternative plus attrayante est d'envoyer un courriel. Le client peut le faire lorsqu'il en a le temps, peu importe l'heure du jour ou de la nuit, et à son rythme. Il n'a pas besoin d'attendre activement une réponse comme au téléphone, et peut continuer ses activités en vérifiant périodiquement son courrier. Aussi, il peut conserver la réponse pour une référence future. En contrepartie, il y a toujours un doute à savoir si le courriel va bel et bien se rendre jusqu'à la bonne personne et si celle-ci va y répondre clairement et rapidement.

Du côté de l'entreprise, il est plus facile de garder les traces des communications par courriel avec les clients, que ce soit pour compiler des statistiques ou pour garder un historique de chaque client. Il est aussi possible d'envoyer des instructions complexes accompagnées de documents audio et vidéo pour faciliter la compréhension. De plus, au lieu d'assigner plusieurs préposés aux lignes téléphoniques en permanence, il est possible de distribuer le travail de répondre aux courriels à plusieurs personnes dont ce n'est pas la tâche principale, ou qui sont séparées géographiquement ou dans le temps.

Les avantages du courrier électronique, combinés au fait que de plus en plus de gens l'adoptent comme principal moyen de communication, font en sorte qu'une proportion croissante de la population préfère envoyer un courriel plutôt que d'utiliser le téléphone lorsqu'il s'agit de communiquer avec une compagnie.

1.2 Les systèmes commerciaux

Si une entreprise propose des adresses électroniques pour communiquer avec elle mais ne gère pas ses courriels d'une façon efficace, elle peut facilement se retrouver dans une situation pire que de ne pas offrir le service du tout. En effet, il n'y a rien de plus frustrant pour un client que de ne pas se faire répondre à l'intérieur d'un délai raisonnable. C'est pour cette raison que plusieurs logiciels de gestion de courrier électronique ont récemment fait leur apparition. La plupart de ces logiciels possèdent plusieurs caractéristiques intéressantes:

- réception des courriels
- routage automatique par mots clés
- utilisation de patrons de réponse figés pour les questions les plus fréquentes
- accès à des bases de données
- correcteur orthographique
- intégration du service à la clientèle
- historique et archivage des messages

Toutes ces options sont attrayantes mais de toute évidence, elles n'aident que très peu à répondre automatiquement au courrier électronique. Et c'est précisément ce que recherchent vraiment les entreprises recevant un grand volume de courriels. Pour pallier à cette lacune, plusieurs logiciels ont intégré des modules de réponse au texte libre [21]. Les méthodes employées sont variées: classifieur bayesien, réseau de neurones, raisonnement à base de cas, prise de décision à base de règles, etc. Tous ces algorithmes nécessitent une certaine base de connaissances qui doit être construite manuellement. Ces bases de données sont statiques, sauf dans le cas du réseau de neurones, qui est continuellement mise-à-jour automatiquement. Une autre constante est que les courriels reçus sont généralement répartis en trois catégories:

simples Courriels contenant un message simple auquel le système n'a pas de problème à répondre.

complexes Courriels avec un message plus complexe que le système dirige vers un préposé en lui suggérant une ou plusieurs réponses.

inconnus Courriels portant un message que le sytème ne reconnaît pas ou auquel il n'est pas capable de répondre.

Il y a aussi une autre caractéristique commune à tous les systèmes sauf un: une seule approche (classifieur bayésien, réseau de neurones, etc.) est employée pour tenter de répondre au courrier électronique. Ce n'est pas un manque en soi, mais il a été suggéré plusieurs fois que l'utilisation combinée de plusieurs techniques améliore les résultats [23, 28, 30].

1.3 Le projet de réponse automatique au courriel

Bell Canada Entreprises (BCE) est une compagnie canadienne dont le but est d'offrir des services de communication tels que le téléphone et l'internet, autant aux particuliers qu'aux entreprises. Son service à la clientèle se doit donc d'être des plus compétitifs. Au lieu d'acheter un système commercial au prix fort élevé, la compagnie s'est plutôt tournée du côté des Laboratoires Universitaires Bell (LUB), un médiateur entre l'industrie du multimédia et le milieu universitaire. Une entente a ensuite été conclue avec le laboratoire de Recherche Appliquée en Linguistique Informatique (RALI) de l'Université de Montréal pour mettre sur pied un projet de réponse automatique au courriel, nommé Merkure. Dans un premier temps, nous avons concentré nos efforts sur le courrier électronique du département des relations aux investisseurs de BCE. Ce service s'adresse aux investisseurs enregistrés et potentiels, et s'occupe de l'envoi des rapports annuels et trimestriels, des notes de presse, etc. Il est aussi très sollicité pour toutes sortes de questions financières (valeur des actions, plans d'achat et de réinvestissement, explications sur un événement récent, etc.) et administratives (ajout ou retrait à une liste de distribution, changement d'adresse, perte de certificat, etc.).

Bien que le service ne soit offert qu'à des fins informatives et que BCE ne se tient pas responsable des pertes encourues suite à une réponse tardive, la rapidité et l'exactitude sont essentielles pour préserver de bonnes relations avec les investisseurs.

1.3.1 Les corpus

Depuis le mois de septembre 2000, nous avons obtenu trois corpus de courriels envoyés à BCE, et un quatrième est en constante évolution. Nous les avons nommés selon le rang de leur acquisition, soit BCE-1, BCE-2, BCE-3 et BCE-4. La table 1.1 résume les informations concernant les corpus, que nous verrons plus en détail au chapitre 4.

corpus	**format**	**date d'obtention**	**courriels**	**domaine**
BCE-1	électronique	septembre 2000	141	général
BCE-2	papier	octobre 2000	865	général
BCE-3	électronique	décembre 2000	1629	relations aux investisseurs
BCE-4	électronique	depuis avril 2001	8153	général

Tab. 1.1: Corpus de courriels provenant de BCE

Comme nous les avons reçus en premier, les corpus BCE-1 et BCE-2 ont servi à une analyse préliminaire et à la rédaction de notre proposition de projet. Par la suite, nous les avons laissés de côté et nous nous sommes concentrés sur BCE-3. C'est ce corpus qui constitue la principale source de données pour l'analyse et les expérimentations mentionnées dans ce mémoire. Nous avons aussi utilisé BCE-4 pour valider nos résultats obtenus avec BCE-3, puisque BCE-4 est une accumulation des messages envoyés quotidiennement aux adresses de BCE. Le nombre de courriels indiqué pour BCE-4 représente le nombre de courriels reçus en date du mois d'avril 2002.

1.3.2 La proposition de projet

L'analyse préliminaire des 2 premiers corpus a révélé que le département des relations aux investisseurs reçoit un mélange de plusieurs types de message[4]. Certains sont des

[4] Nous utilisons le terme *message* plutôt que *question* car un courriel peut avoir un but communicatif différent d'une question (plainte, demande de documents, etc). Aussi, nous préférons le terme *suivi* à *réponse* parce qu'il existe plusieurs actions possibles, comme les transferts et les accusés de réception.

messages courts et précis nécessitant une réponse factuelle. D'autres sont des messages plus longs et dont la réponse doit être formée à partir de plusieurs sources d'informations différentes. Puis, certains messages reviennent en grand nombre et n'ont besoin que d'un accusé de réception ou d'un transfert. Pour cette raison, et contrairement aux systèmes commerciaux, l'accent a été mise sur une combinaison d'approches complémentaires plutôt que d'employer une seule méthode. Cela a pour but de traiter le mieux possible toutes les sortes de messages et d'augmenter le niveau de confiance général du système. Tel qu'illustré à la figure 1.1, ce dernier est constitué de cinq modules distincts:

extraction d'entités nommées Identifie certaines expressions sémantiques comme les noms de personnes et de compagnies, les dates, les numéros de compte, etc. Utile pour personnaliser les réponses en fonction du client et de la situation, et pour repérer des informations importantes pour le système question-réponse. Peut aussi être utile au module de classification en faisant ressortir certaines informations typiques d'une sorte de suivi.

classification Répond aux questions en se fiant sur un ensemble de classes message-suivi prédéfinies. Utile pour les messages fréquents et stéréotypés. Peut aussi effectuer un traitement préliminaire pour alimenter seulement les modules concernés par un type de message. C'est sur ce module que se concentrera ce mémoire.

question-réponse (QR) Répond aux questions assez courtes et factuelles à l'aide d'une recherche d'information. Cette approche s'adapte facilement aux nouvelles situations car la réponse est générée dynamiquement à partir d'une ou plusieurs bases de données, comme le site web de BCE. Tant que la base de données est à jour, les réponses le sont aussi.

raisonnement à base de cas (RBC) Répond aux questions en se fiant à un répertoire de cas, constitués de paires message-suivi. Utile pour les questions peu fréquentes et dont la réponse ne fait pas partie de la base de données.

évaluation et combinaison des suivis Combine les résultats des modules de classification, question-réponse et raisonnement par cas, et identifie le suivi le plus

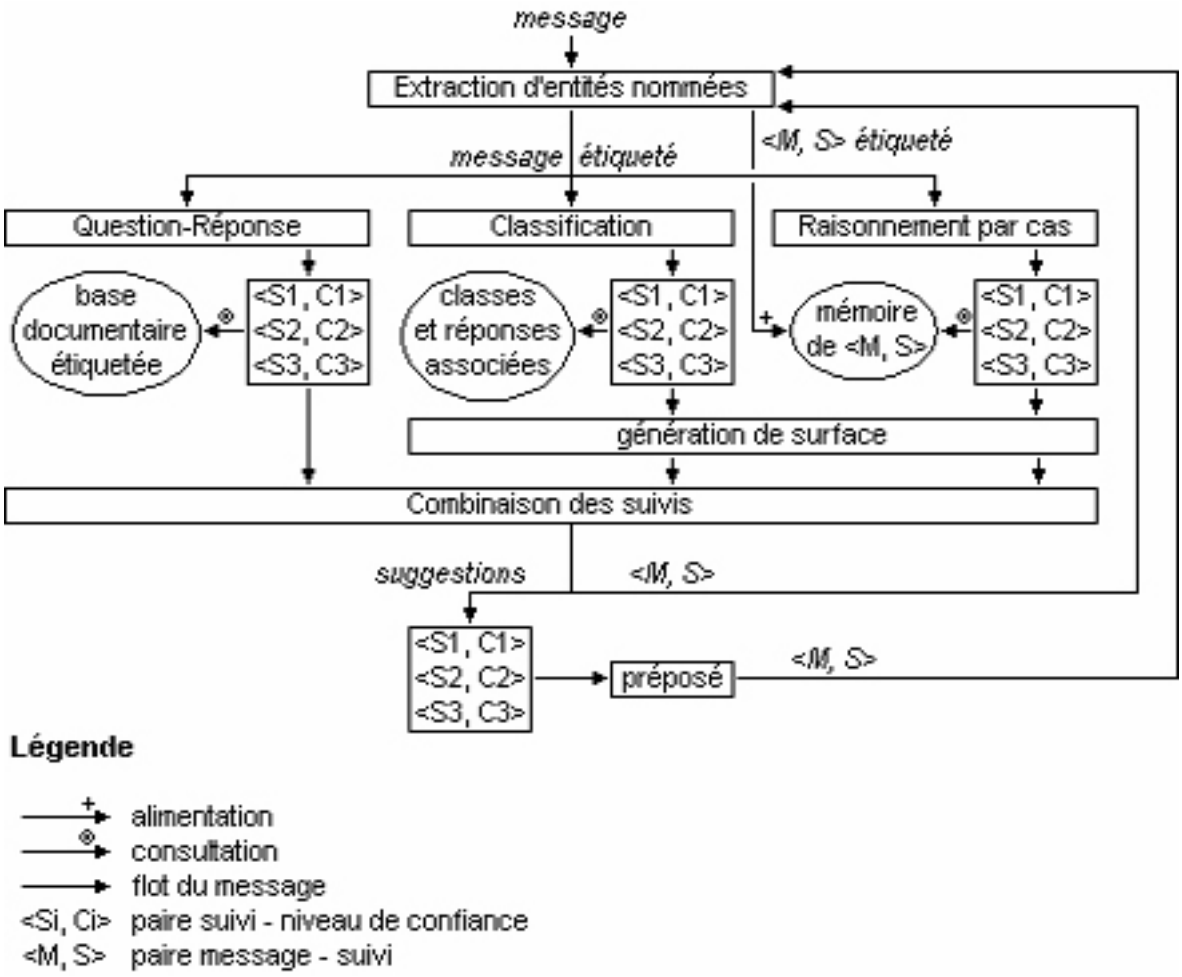

Fig. 1.1: Architecture du système

approprié en fonction du type de message.

Lorsqu'un message est reçu, il est d'abord analysé par l'extracteur d'entités nommées qui identifie les expressions sémantiques importantes. Le message est ensuite envoyé aux trois modules d'analyse (classification, question-réponse et raisonnement par cas) qui vont tenter de répondre au message. Chacun de ces modules fournit une liste (qui peut ne contenir qu'un élément) de suivis potentiels, indépendamment des autres. Puis, les suivis sont passés au module d'évaluation et combinaison, qui détermine le meilleur parmi les choix offerts. Si le niveau de confiance (réglable) n'est pas atteint, la liste des suivis potentiels est envoyée au préposé qui peut en choisir un de ceux-là ou un autre qui n'est pas dans la liste. Dans les deux cas, le choix final est étiqueté par l'extracteur

d'entités nommées et la paire message-suivi est insérée dans la base de connaissances du raisonnement par cas.

1.4 Le module de classification

Ce mémoire porte sur le module de classification de Merkure. Celui-ci peut servir à plusieurs fins. La classification de textes est généralement associée à la séparation des documents selon leur contenu. Dans le cas de Merkure, il est possible de classer les messages selon les aspects suivants:

type de message Chaque approche possède un point fort, un type de message pour lequel il est particulièrement adapté. Il est donc envisageable d'introduire une étape intermédiaire, entre l'extraction d'entités nommées et l'analyse du contenu par les différentes méthodes, dont le but est de déterminer le module le plus adéquat pour chaque type de message. Le type d'un message est basé sur des caractéristiques générales comme le nombre de mots, le pourcentage de numéraux par rapport au nombre de mots, le ratio de pronoms personnels sur le total des mots, etc. Une telle classification permet d'éviter un traitement inutile par le ou les modules moins bien adaptés lorsque le niveau de confiance est très élevé.

type de suivi Il existe plusieurs suivis possibles: réponse, transfert, accusé de réception ou même, dans le cas de remerciements, ne rien faire. Encore une fois, une étape intermédiaire située juste après l'extraction d'information peut éliminer une bonne quantité de traitements inutiles de la part des modules d'analyse. Par exemple, lorsqu'un message ne requiert qu'un transfert ou un accusé de réception, il est inutile de le soumettre à un traitement poussé comme le font les modules de question-réponse et de raisonnement à base de cas.

contenu du message Classifier les messages par rapport à leur contenu peut paraître inutile puisque les grandes classes contiennent généralement plusieurs sous-classes hétérogènes de messages. Par exemple, une catégorie `actions privilégiées` pourrait contenir trois sous-ensembles distincts (série P, série Q et série S)

ayant chacun quelques messages qui leur sont spécifiques. Par conséquent, associer un suivi à une classe peut sembler tout aussi inutile et inapproprié. Il existe toutefois certains messages au contenu identique (ou presque) qui surviennent assez fréquemment pour en faire une catégorie, mais cela n'arrive que rarement. Par contre, certains messages reçus en très grand nombre peuvent former une classe paramétrée, c'est-à-dire que tous les messages font la même requête et tous les suivis sont similaires, à l'exception de certains paramètres ou variables. Par exemple, les demandes de rapports annuels et trimestriels sont, à la base, identiques. Le suivi aussi est le même pour les deux sortes de rapports, et la seule chose qui change est le type de rapport, qui représente la variable. Une classification en fonction du contenu est pratique lorsque quelques catégories paramétrées et classes exclusivement homogènes dont définies. Dans le cas contraire, il peut tout de même être utile de classifier les messages selon leur contenu pour obtenir des statistiques.

ton du message La majorité des messages sont plutôt neutres. Mais il arrive que certaines personnes montrent clairement ce qu'elles ressentent (colère, bonheur, énèrvement, etc.) dans leur message. Dans de tels cas, il est possible de personnaliser la réponse en fonction de l'état de la personne, comme le ferait un être humain.

Ces différentes formes de classification des messages ne sont pas nécessairement mutuellement exclusives. Au contraire, elles sont assez étroitement reliées. Il est tout-à-fait normal que les messages ayant un contenu similaire soient appariés au même type de suivi. Dans le même ordre d'idées, le type de message donne généralement une bonne indication du type de suivi qui sera nécessaire, et donc du contenu.

Nous pensons être capable de classer les courriels du département des relations aux investisseurs en plusieurs catégories selon une combinaison des trois premières façons de classer les messages. La classification principale sera faite en fonction du contenu, en utilisant les types de message et de suivi si nécessaire. Nous croyons pouvoir atteindre un bon niveau d'efficacité en se basant principalement sur les mots pour les caractéristiques

d'apprentissage.

1.5 Plan du mémoire

Dans un premier temps, nous survolerons dans le chapitre 2 les algorithmes d'apprentissage et les caractéristiques que nous utiliserons pour nos expérimentations. Nous verrons ensuite les caractéristiques techniques et linguistiques propres au courrier électronique. Le chapitre 3 décrira brièvement les systèmes informatiques de BCE et de Stylus qui gèrent le site web et le courrier électronique. Les modifications proposées et le processus d'implantation d'un prototype y seront aussi présentés. Le chapitre 4 exposera ensuite une analyse détaillée du corpus BCE-3, sur lequel s'appuient nos expérimentations, qui seront présentées et discutées au chapitre 5. Finalement, nous terminerons avec un résumé des avenues explorées et des possibilités d'extension.

Chapitre 2

Travaux connexes

La classification de textes est loin d'être une nouveauté. L'importance croissante de l'éducation, l'apparition de nouvelles disciplines et l'abondance grandissante de la littérature en général sont tous des facteurs qui ont favorisé son utilisation. Au milieu du $20^{ième}$ siècle, un nombre toujours plus imposant de textes disponibles devaient être classés manuellement. Ce travail devenait de plus en plus fastidieux, proportionnellement à la quantité des nouveaux ouvrages publiés. L'intégration des ordinateurs dans la société a permis de se poser une question qui permettrait peut-être de transformer des heures de travail manuel en quelques minutes de traitement par ordinateur: serait-il possible d'automatiser la classification de textes avec une intervention minimale d'une personne? De nos jours, cette question a pris une importance capitale en raison de l'essor fulgurant de l'Internet, l'hôte du "World Wide Web" et du courrier électronique.

La classification de textes consiste à séparer des documents textuels en plusieurs catégories distinctes, selon leur contenu. Chaque catégorie regroupe donc des documents au contenu similaire. Le but principal de cette tâche est d'avoir une représentation sommaire du sujet de chaque document afin de pouvoir rapidement et facilement identifier les documents connexes à un certain sujet. Généralement, un document présente des signes d'appartenance plus marqués pour une certaine classe. Par contre, ça ne veut pas dire que le document n'est apparenté à aucune autre catégorie. Il est donc possible d'utiliser une classification stricte, c'est-à-dire que chaque document n'appartient qu'à une classe, ou souple, où les documents peuvent faire partie de plus d'une catégorie.

Cette deuxième sorte de classification est un peu plus compliquée à gérer, mais peut être plus utile que la première dans bien des cas.

La classification de textes n'est pas vraiment un domaine distinct. Elle consiste plutôt en une fusion des algorithmes d'apprentissage et des techniques de recherche d'information, tout en étant reliée de près au "clustering". Celui-ci est l'action contraire de la classification, c'est-à-dire qu'il sert à trouver des regroupements adéquats pour un ensemble de documents. Au lieu d'approcher la classification de textes comme un tout, il est plus simple de la considérer comme deux sous-problèmes distincts mais interconnectés: le choix d'un ou de plusieurs algorithmes d'apprentissage et le choix des caractéristiques à utiliser. Pour ce qui est de la classification du courrier électronique, les mêmes principes s'appliquent que pour la catégorisation de documents publiés. Il y a cependant quelques ajustements à faire en raison des différences entre un texte publié et celui rencontré dans un courriel. De plus, tandis qu'une publication ne contient que du texte, le courrier électronique est pourvu de multiples extensions qui compliquent le traitement automatique.

2.1 Les algorithmes d'apprentissage

L'apprentissage consiste à modifier ses réponses futures en se servant de ses expériences précédentes. Un bon apprentissage ne doit pas se limiter à apprendre par coeur, mais doit permettre une adaptation à de nouveaux cas. Les algorithmes d'apprentissage fonctionnent selon le même principe de base [44, 45]. Les expériences précédentes sont représentées par un ensemble de données $D_n = (z_1, z_2, \ldots, z_n)$ dont on suppose généralement que les données ont été tirées indépendamment d'une même distribution inconnue $P(Z)$. C'est l'hypothèse que les données sont identiquement et indépendamment distribuées. L'ensemble $\mathcal{F}$ regroupe les fonctions, c'est-à-dire les solutions possibles. Il est possible de mesurer l'erreur d'une fonction $f \in \mathcal{F}$ sur un $Z = z$ particulier à l'aide une fonction de coût $L(z, f)$. Par conséquent, il est possible de mesurer l'erreur moyenne pour un f particulier sur un ensemble d'exemples D_n:

$$\hat{R}(f, D_n) \quad = \quad \frac{1}{n} \sum_{i=1}^{n} L(z_i, f)$$

Il s'agit du *risque empirique*. Le *risque espéré*, ou *erreur de généralisation*, mesure l'erreur espérée avec f sur un exemple tiré aléatoirement de $P(Z)$:

$$R(f) \quad = \quad \int L(z, f)P(z)dz$$

Le problème de l'apprentissage est de trouver $f \in \mathcal{F}$ qui minimise l'erreur de généralisation $R(f)$. Malheureusement, il est impossible de calculer cette valeur alors que c'est elle qui est vraiment intéressante. Cependant, il est possible de l'estimer par le risque empirique. Le problème devient donc de minimiser ce risque empirique en trouvant la meilleure fonction pour l'ensemble D_n:

$$f^*(D_n) \quad = \quad argmin_{f \in \mathcal{F}} \hat{R}(f, D_n)$$

L'*erreur d'apprentissage* est la plus petite perte L moyenne sur les exemples d'apprentissage D_n obtenue en choisissant $f^* \in \mathcal{F}$:

$$\hat{R}(f^*(D_n), D_n) \quad = \quad min_{f \in \mathcal{F}} \hat{R}(f, D_n)$$

Le problème avec cette approche, c'est que $\hat{R}(f^*(D_n), D_n)$ est un estimé biaisé de $R(f^*(D_n))$. La fonction f^* est meilleure sur l'ensemble D_n qu'un autre ensemble tiré au hasard de la distribution $P(Z)$ puisque l'évaluation est faite à partir des mêmes données utilisées pour l'apprentissage. Normalement, en choisissant un autre ensemble D'_n, on obtient une autre fonction $f^*(D'_n)$. Généralement, $f^*(D_n)$ et $f^*(D'_n)$ ne peuvent pas être simultanément correctes partout où elles sont différentes, donc elles sont aussi différentes de la fonction qui minise le plus l'erreur de généralisation dans $\mathcal{F}$. Cependant, si la validation est effectuée sur un ensemble D' tiré indépendamment de l'ensemble d'apprentissage D_n, l'erreur d'apprentissage $\hat{R}(f^*(D_n), D')$ devient un estimé non biaisé de l'erreur de généralisation $R(f^*(D_n))$.

Concrètement, il existe plusieurs façons d'estimer l'erreur de généralisation. La plus simple est de séparer l'ensemble de données D un ensemble d'entraînement D_e et un ensemble de validation D_v. L'algorithme est entraîné avec l'ensemble d'apprentissage et l'erreur de généralisation est estimée par l'erreur empirique mesurée sur l'ensemble de validation:

$$R(f) \quad \approx \quad \hat{R}(f^*(D_e), D_v)$$

Lorsque les données ne sont pas assez nombreuses pour les séparer en deux ensembles, la validation croisée est une approximation plus fiable. Il s'agit de partitionner l'ensemble de données D en k sous-ensembles D_1, D_2, ..., D_k. Le choix de k est arbitraire, mais la valeur dix est la plus souvent utilisée. Puis, pour chaque $i = 1, 2, \ldots, k$, il faut entraîner l'algorithme sur toutes les données sauf celles contenues dans D_i, et mesurer l'erreur empirique sur l'ensemble D_i. Il faut donc répéter k fois le cycle d'apprentissage, et ensuite utiliser la moyenne des erreurs empiriques comme estimé non biaisé de l'erreur de généralisation:

$$R(f) \approx \frac{1}{k} \sum_{i=1}^{k} \hat{R}(f^*(D - D_i), D_i)$$

L'entraînement d'un algorithme d'apprentissage en classification de textes contient trois phases: la préparation initiale des données, l'apprentissage et la validation. La première phase consiste à trouver un échantillon de documents types. Cet échantillon doit être représentatif de la sorte de documents à classer. Les documents servant à l'entraînement doivent ressembler aux documents que le classifieur aura à traiter. Il faut ensuite identifier à quelle(s) classe(s) appartient chaque document. Puis, selon la méthode d'estimation de l'erreur de généralisation choisie, il faut séparer de façon aléatoire les ensembles d'entraînement et de validation.

La deuxième étape est l'entraînement de l'algorithme. À l'aide de l'ensemble d'entraînement, l'algorithme apprend quelles sortes de documents se retrouvent dans quelles classes. Un prétraitement des documents est habituellement requis pour transformer le texte libre en un ensemble de caractéristiques faisant bien ressortir les différences entre les classes. Pour n classes $C_1, C_2, \ldots, C_n$, une fonction de coût possible peut être:

$$L(f_{C_j}, z) = \begin{array}{ll} 0 & si\ z \in C_j \\ 1 & sinon \end{array}$$

Cette fonction simple attribue toujours le même coût à une erreur, ce qui donne l'erreur empirique suivante:

$$\hat{R}(f^*(D_e), D_v) \approx \frac{nombre\ de\ documents\ mal\ classés}{nombre\ de\ documents\ traités}$$

Dans certains cas, il peut être avantageux d'attribuer un coût plus élevé à certaines erreurs pour influencer la classification. Vient ensuite la phase de validation. Le classifieur tente de bien étiqueter les documents appartenant à l'ensemble de validation. La classification obtenue est comparée à celle préalablement établie pour en déduire l'erreur empirique Si les résultats sont concluants, le classifieur est utilisé pour traiter de nouveaux cas dont la classification est inconnue.

Le choix de l'algorithme d'apprentissage est important pour la classification de textes et dépend de la tâche à accomplir. Le temps d'apprentissage, le nombre de documents dans l'ensemble d'entraînement, la possibilité de modifier la classification, le type de corpus ainsi que le nombre de classes sont tous des facteurs à prendre en considération. Il est aussi possible de combiner de différentes manières plusieurs algorithmes dans l'espoir d'améliorer les résultats obtenus avec un seul. Nous décrivons dans les sections suivantes quelques types de classifieurs.

2.1.1 Les k-plus-proches-voisins

L'algorithme des k-plus-proches-voisins (kppv) est un algorithme paresseux [9]. Chaque document est représenté par un vecteur de longueur n, soit $\vec{d} = (d_1, d_2, \ldots, d_n)$ où chaque élément d_i est l'une des caractéristiques d'apprentissage. Il n'y a pas vraiment d'apprentissage avec cet algorithme. Pour identifier un nouveau document, le kppv calcule sa similarité avec tous les documents déjà classés. Puis, il place les valeurs de similarité par ordre décroissant et ne garde que les k premières. La classe revenant le plus souvent parmi ces k documents est celle qui est attribuée au nouveau document. Donc, comme son nom l'indique, la classification d'un nouveau document dépend de ses k voisins les plus proches. Cette notion de distance est plus facilement compréhensible avec l'exemple en deux dimensions de la figure 2.1. Dans cet exemple, le nouveau document aura de fortes chances de faire partie de la classe noire puisqu'il côtoit de plus près les documents de cette catégorie.

La similarité entre deux documents se calcule à l'aide d'une corrélation quelconque entre les deux vecteurs. Une telle corrélation peut être la distance euclidienne ou encore le cosinus de l'angle formé par les deux vecteurs. Le choix de la constante k

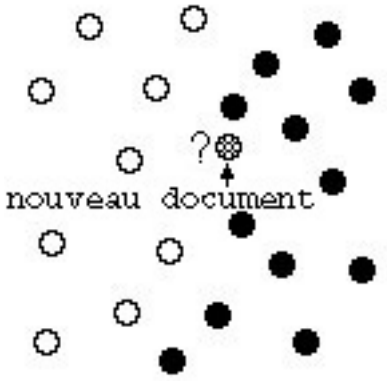

Fig. 2.1: Exemple en deux dimensions du k-plus-proches-voisins

est dépendant de la taille de l'échantillon et des classes, et influence les résultats de la classification. Lorsque k est petit, la classification est plus sensible à cause des documents appartenant à une classe mais dont leur vecteur de représentation ressemble beaucoup plus à une autre. Par contre, lorsque k est trop grand, les catégories ayant peu d'exemples peuvent être désavantagées par rapport à celles qui en ont plus.

L'absence d'apprentissage fait du kvvp un parfait candidat pour les classifications qui doivent être constamment ajustées ou révisées. En fait, la mise-à-jour de la classification ne peut pas être plus simple. Il n'y a qu'à étiqueter et indexer les nouveaux exemples, et enlever les documents nuisibles. En contre-partie, le temps de classement d'un nouveau document est proportionnel au nombre d'exemples et au nombre de caractéristiques utilisées. Mais en raison de la vitesse des ordinateurs actuels, ce temps d'exécution est minime, même lorsqu'il y a plusieurs centaines de documents et plusieurs milliers de caractéristiques. Malgré sa simplicité, le kvvp est l'objet de plusieurs études [15, 16, 46].

2.1.2 Le Bayes naïf

Le Bayes naïf est un classifieur probabiliste [20]. Il classe les nouveaux documents selon la probabilité de ces documents à appartenir à chaque classe. En utilisant le théorème de Bayes (d'où le nom), la probabilité que le document $\vec{d} = (d_1, d_2, \ldots, d_n)$ appartienne

à la classe c_j est:

$$P(c_j|\vec{d}) = \frac{P(c_j)P(\vec{d}|c_j)}{P(\vec{d})}$$

Puisque $P(\vec{d})$ est identique pour chaque classe, il est possible de l'enlever et cette équation devient alors:

$$P(c_j|\vec{d}) = P(c_j)P(\vec{d}|c_j)$$

En supposant l'indépendance des caractéristiques d'apprentissage (d'où l'adjectif naïf), la probabilité de trouver le document à l'intérieur d'une classe est égale au produit des probabilités de trouver ses caractéristiques:

$$P(c_j|\vec{d}) = P(c_j)\prod_{i=1}^{n} P(d_i|c_j)$$

où n est le nombre de caractéristiques. Pour classifier un nouveau document, la probabilité de chaque classe est calculée et la plus haute l'emporte. Avec ce classifieur, l'apprentissage se résume à établir la probabilité des caractéristiques à appartenir à chacune des classes. Selon l'initialisation des probabilités, il est possible de mettre à jour un tel classifieur régulièrement et sans trop de complications. Et malgré plusieurs détracteurs, le Bayes naïf est largement utilisé [5, 27, 37].

2.1.3 L'arbre de décision

Il existe plusieurs variantes de cet algorithme, dont deux des plus connues sont CART [3] et ID3 [35]. L'arbre de décision, comme son nom l'indique, est un regroupement de fonctions locales structurées en forme d'arbre. En fait, les fonctions sont habituellement suffisament localisées pour être de simples règles condition-action. Le noeud racine est associé à l'ensemble de données D, et chacun des autres noeuds est associé à un sous-ensemble D_n. Un noeud est soit une feuille, soit un noeud interne utilisant une fonction f pour séparer son ensemble associé D_n entre au moins deux enfants. Comme le montre la figure 2.2, chaque noeud de l'arbre représente un classifieur simplifié qui prend sa décision à partir d'un minimum de caractéristiques. En outre, il n'est pas rare qu'un noeud ne regarde que l'absence ou la présence d'une seule caractéristique.

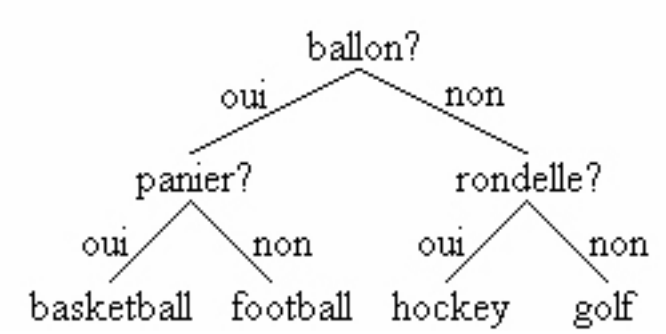

Fig. 2.2: Exemple d'un arbre de décision

L'apprentissage d'un arbre de décision est une procédure assez simple appliquée récursivement à chaque noeud. L'algorithme vérifie premièrement si le noeud doit rester une feuille. C'est le cas lorsque la grande majorité des documents appartenant à l'ensemble D_n associé à ce noeud font partie de la même catégorie. Dans le cas contraire, l'ensemble D_n est séparé en deux ou plusieurs sous-ensembles à l'aide d'une fonction f qui minimise l'erreur empirique de cette classification locale. Il existe plusieurs méthodes pour trouver cette fonction, dont l'une est d'utiliser l'entropie [25]. La même procédure est ensuite appliquée à chacun des sous-ensembles, jusqu'à ce que chaque feuille corresponde majoritairement à une classe. Habituellement, l'arbre complété contient trop de noeuds, ce qui résulte en une mauvaise généralisation et une difficulté à classer de nouveaux documents. Il faut donc l'élaguer, c'est-à-dire enlever les noeuds représentant les règles trop spécifiques. Les multiples manières de produire les noeuds et d'élaguer l'arbre ont résulté en classifieurs tels C4.5 [36] et RIPPER [6].

2.1.4 Les cooccurences de mots

Au lieu de se limiter aux simples mots, il peut être avantageux d'utiliser des cooccurences de mots [39]. Si les mots décrivent bien le contenu d'un document, les groupes de mots le font encore mieux. Selon ce principe, l'algorithme cherche à identifier des groupes de mots qui représentent bien le document et leur attribue un poids selon les même principes que pour des mots simples. Un groupe peut contenir deux mots ou plus, et les mots d'un groupe ne sont pas tenus d'être adajcents, en raison des caprices

des langues. Dans ce cas, on utilise généralement une fenêtre, c'est-à-dire que les mots ne doivent pas être séparés par plus d'un certain nombre de mots. Avec cette méthode, il est préférable d'enlever les mots vides de sens avant l'apprentissage, pour éviter de se retrouver avec des groupes de mots formés d'un article et d'un nom, ce qui équivaut au nom seul.

2.1.5 La combinaison de plusieurs classifieurs

Chaque algorithme possède ses forces et ses faiblesses pour classifier de nouveaux documents. En outre, certaines de ces forces sont complémentaires et peuvent améliorer les résultats lorsqu'elles sont combinées. Cette combinaison est surtout utile lorsque les classifieurs se trompent sur des documents différents. Si les algorithmes font les mêmes erreurs, cette technique perd tout son lustre et ne change presque rien aux résultats. Il y a de nombreuses façons de conjuguer les efforts de plusieurs classifieurs en un comité. Un comité est un ensemble de classifieurs dont on combine les prédictions d'une façon quelconque [8].

La manière la plus simple est sans contredit d'entraîner les classifieurs le plus indépendamment possible, puis de les faire voter. La prédiction du comité est la classe qui revient le plus souvent. Il a été démontré que l'efficacité moyenne d'un comité est supérieure à la moyenne d'efficacité des classifieurs de ce comité [22].

Une méthode un peu plus complexe, mais aussi plus susceptible d'apporter une amélioration à l'efficacité, se nomme "stacking" [49]. Il s'agit d'entraîner séquentiellement plusieurs algorithmes, où le $n^{ième}$ classifieur, dans son apprentissage, tient compte de l'erreur de généralisation de ceux qui le précéde. Par exemple, le comité peut contenir huit classifieurs, dont sept sont entraînés indépendamment sur un même échantillon. L'apprentissage du huitième consiste à associer le meilleur des sept classifieurs à utiliser selon les caractéristiques du document à classer.

Une autre alternative est le "boosting" [42]. Il s'agit d'entraîner séquentiellement plusieurs algorithmes en mettant de plus en plus d'emphase sur les cas difficiles à classer. L'apprentissage devient donc de plus en plus spécifique à chaque fois qu'un nouvel algorithme est ajouté au comité. Un méta-classifieur s'occupe d'apprendre les

forces et les faiblesses des autres classifieurs selon les caractéristiques d'apprentissage.

Puisqu'il n'existe pas une quantité infinie de classifieurs, de plus en plus de travaux se concentrent sur la combinaison de plusieurs techniques différentes. Une telle approche nécessite généralement plus de temps pour l'apprentissage et la classification de nouveaux documents, mais les pourcentages d'efficacité sont aussi plus élevés [23, 28, 30].

2.2 Les caractéristiques d'apprentissage

Les caractéristiques servant à l'apprentissage, et leur représentation, sont aussi importantes que l'algorithme d'apprentissage. Elles lui sont aussi reliées de près car tous les classifieurs ne prennent pas les mêmes formes d'entrée. Quelques possibilités sont les valeurs booléennes, discrètes, et continues.

Les caractéristiques d'apprentissage désignent les aspects sur lesquels porte la classification. Pour la classification de textes, il s'agit habituellement des mots apparaissant dans les documents. Peu importe le document, il est fort probable que les mots qu'il contient représentent bien son contenu. Cependant, selon le type de classification, il peut être utile d'ajouter d'autres caractéristiques comme le nombre de pronoms à la première personne ou encore la quantité de numéraux. En ce qui concerne les mots, tous ne sont pas utilisés. Habituellement, les numéraux et les mots vides de sens sont enlevés car ils n'apportent rien de plus au niveau du contenu. Les mots vides de sens sont surtout des articles, des prépositions et des adverbes. Aussi, il n'est pas inhabituel d'enlever les mots très fréquents et les mots rares. Tous ces mots ont une influence plus ou moins limitée selon l'attribution des poids aux mots [2]. Une chose est cependant sûre: garder les mots qui n'influencent que très peu les résultats rallongent inutilement le temps d'apprentissage et de classification. De plus, les mots conservés sont souvent modifiés pour ne conserver que leur racine, ou une version tronquée du mot. Cela permet de réduire encore le nombre de caractéristiques, mais surtout de faire abstraction des pluriels et des multiples façons d'exprimer la même idée. Habituellement, les caractéristiques à utiliser sont largement dépendantes du type du corpus et de la classification désirée [19, 40].

En classification de textes, comme en recherche d'information en général, la mesure de poids la plus rencontrée est connue sous le nom de *tf*idf*. Il s'agit de l'abréviation pour "term frequency * inverse document frequency". La première partie mesure l'importance du mot dans le document en normalisant le nombre d'occurences du terme à l'intérieur du document. Plus le mot est utilisé, plus il est représentatif du document. La fréquence du terme d_i à l'intérieur du document $\vec{d_k} = (d_1, d_2, \ldots, d_n)$ est:

$$tf(d_i) = \frac{freq(d_i, \vec{d_k})}{max_{1 \leq l \leq n} freq(t_l, \vec{d_k})}$$

où n est le nombre de caractéristiques. La normalisation s'effectue en divisant la fréquence du mot par la plus haute fréquence d'un mot apparaissant dans ce document. Elle a pour but de mettre tous les documents sur un pied d'égalité en terme de longueur. Dans le cas contraire, cette mesure ne serait pas fiable car le poids d'un mot serait proportionnel à la longueur du document dans lequel il apparaîtrait.

La deuxième partie représente l'importance du mot dans l'échantillon. Plus il y a de documents qui contiennent un mot, moins il y a de chance qu'il soit représentatif d'un sujet précis. La fréquence inverse du terme d_i se calcule par:

$$idf(d_i) = log \frac{|D|}{|D_{d_i}|}$$

où $|D|$ est le nombre de documents dans l'échantillon et $|D_{d_i}|$ est le nombre de documents dans l'échantillon qui contiennent le mot d_i. En combinant les deux parties, le poids w_i associé au terme d_i est:

$$w_i = \frac{freq(d_i, \vec{d_k})}{max_{1 \leq l \leq n} freq(t_l, \vec{d_k})} \times log \frac{|D|}{|D_{d_i}|}$$

Cette équation n'est cependant pas unique. Il existe plusieurs autres façons équivalentes de combiner les deux parties ensemble [41].

2.3 Les caractéristiques du courrier électronique

2.3.1 Le langage électronique

À première vue, le courrier électronique peut être perçu comme une façon supplémentaire de transmettre un document écrit. En fait, plusieurs nouveaux concepts reliés à

l'internet sont associés aux documents écrits [33]. Par exemple, le nom donné au système de messagerie de l'internet est le *courrier* électronique, et ses principaux icônes sont l'*enveloppe* et la *boîte aux lettres*. Cependant, en regardant de plus près les caractéristiques propres aux courriels, il devient vite apparent qu'il s'agit d'une nouvelle forme de communication à part entière. Celle-ci forme un juste milieu entre l'écrit et l'oral, tout en étant fortement influencée par l'omniprésente rapidité rattachée à l'informatique.

Les messages électroniques ne sont ni "verbaux" ni "écrits" au sens conventionnel de ces deux termes [7]. Il est difficile de dire qu'ils sont verbaux parce que les gens impliqués ne se parlent pas de vive voix. D'un autre côté, on ne peut les considérer comme strictement écrits parce qu'ils sont souvent composés sur-le-champ, sans planification préalable et en ne suivant aucunement les règles de base de la rédaction. Il ne s'agit pas d'un phénomène unique au courrier électronique mais partagé par la majorité des communications électroniques comme les groupes de nouvelles, les forums de discussion, le bavardage en ligne (Internet Relay Chat) et les programmes de contacts (ICQ[1], Instant Messenger[2], etc.).

Les sections suivantes décrivent les principales caractéristiques des langages électroniques qui diffèrent du texte conventionnel.

2.3.1.1 L'organisation

L'organisation d'un document écrit est généralement linéaire. Plusieurs séquences de lettres forment des séquences de mots qui, à leur tour, mènent à des séquences de phrases, de paragraphes, de sections et de chapitres [33]. Le fil du discours aussi est linéaire. Le texte commence par une introduction qui en annonce le sujet. Suit le corps du document, qui peut grandement varier selon le type de document, mais qui est habituellement divisé en sections distinctes pour bien encadrer les différents points. Et finalement, une conclusion énumère les points les plus importants. Tout est bien structuré. Le lecteur voit clairement le fil directeur tout au long du document et

[1] web.icq.com

[2] www.aim.com

les différentes parties sont séparées par des marqueurs de relation. À l'opposé, une communication orale est plutôt contextuelle. Les personnes parlent en alternance, de façon spontanée, en suivant le cours de leurs pensées. Elles ne se soucient guère de l'organisation de leurs phrases, tant que les autres comprennent ce qu'elles veulent dire. À cause de cela, la conversation peut subitement prendre une toute autre direction sans que les interlocuteurs ne s'en aperçoivent.

À mi-chemin entre les deux, l'organisation des courriels représente bien le caractère hybride des communications électroniques et la sensation de vitesse qui se dégage de l'informatique. Au lieu de reprendre les mêmes concepts d'introduction, de corps et de conclusion typiques des autres documents écrits, les courriels vont directement au but. Grâce aux logiciels qui incluent automatiquement le texte du message auquel on répond, il n'est plus nécessaire d'introduire le sujet. L'individu peut tout simplement ajouter la réponse à la suite du passage auquel il fait référence. En outre, il n'est pas rare qu'un courriel contienne plus d'un sujet de discussion en même temps. Habituellement, chaque sujet est contenu dans un paragraphe différent pour bien les délimiter. Il est alors facile de répondre à chaque partie en insérant les ajouts en-dessous des sections correspondantes, tel qu'illustré à la figure 2.3.

Cet exemple illustre bien l'utilisation de citations, typique de l'écrit, pour promouvoir le contexte comme dans une communication orale [11]. Combiné avec les réponses courtes et concises, il se crée un sentiment de "présence" lors de la lecture. Au lieu de devoir lire un long texte froid et impersonnel, le récepteur se retrouve devant un bout de texte si court et direct qu'il ressemble plus à une réponse verbale pour une question qui vient juste d'être posée. En plus, les différents sujets n'ont pas nécessairement la même durée de vie, et d'autres sujets peuvent se greffer au courriel, même s'ils ne sont pas reliés aux précédents. Cela augmente encore plus la sensation de conversation verbale transposée en document écrit.

2.3.1.2 La ponctuation

L'utilisation de la ponctuation est un autre indice que les courriels sont souvent rédigés différemment des autres documents écrits. En suivant le modèle de l'oral, les phrases

```
>>> Nous devrions nous réunir pour parler de la présentation d'hier.
>>> Quelles sont tes disponibilités pour la semaine prochaine?
>>
>> Je suis disponible lundi, mardi, et peut-être mercredi.
>
> Est-ce que mardi a 14h te convient?

Oui.

>>> Comment avance votre projet depuis le départ du responsable?
>>
>> Assez bien.  Nous pensons être capable de respecter l'échéance.
>
> Ton patron sera surement heureux d'apprendre cela!
>

Penses-tu que j'ai une chance d'obtenir le poste?
```

Fig. 2.3: Exemple de courriel ressemblant à une conversation orale

sont courtes et simples. Cela a mené à la quasi-disparition de certains signes de ponctuation dont le point-virgule (;), le deux-points (:) et le tiret (—). Surtout présents pour bien délimiter de longues phrases ayant une structure complexe, ils deviennent inutiles pour les phrases simples de type sujet-verbe-complément qui forment la majorité des courriels.

À l'opposé, d'autres signes de ponctuation sont beaucoup plus présents, frôlant même la surutilisation. Le meilleur exemple est le point d'exclamation (!). Dans un document écrit, ce signe de ponctuation sert à mettre une grande emphase sur la phrase et peut même aller jusqu'à en changer le sens [47]. Par exemple, *Es-tu fou?* est une question comme tant d'autres. Si le point d'interrogation est remplacé par un point d'exclamation, il ne s'agit plus d'une question mais d'une interjection prononcée avec force. Dans les communications électroniques, le point d'exclamation n'apporte pas autant d'emphase. Par contre, il est fréquent de l'utiliser à répétition pour en augmenter la force. Ainsi, *Heille le malade!!!!!* devrait être prononcé avec plus de vigueur que *Heille le malade!*. Un autre signe de ponctuation qui a vu son utilisation grimper en flèche est

les points de suspension (...). Comme ils servent à marquer l'hésitation ou l'incertitude, ou encore à créer une interruption, une attente ou un suspense, ils comblent une partie du vide laissé par le passage de l'oral à l'écrit. Les points de suspension sont l'équivalent le plus juste de plusieurs formes d'intonation difficilement reproductibles et sont donc employés à profusion lorsqu'il s'agit d'écrire ce qui devrait être prononcé: *J'ai compilé les votes, et le ou la gagnante est... Alice!*

2.3.1.3 Le vocabulaire

Le vocabulaire aussi n'est pas tout-à-fait le même pour le courrier électronique. Premièrement, l'étendue du vocabulaire est assez réduite. Dans un courriel, les gens écrivent ce qu'ils diraient verbalement, et de la manière dont ils le diraient. Ils ne se soucient que très peu d'utiliser un large éventail de mots, en autant que le message soit compréhensible. Cependant, cette diminution marquée du nombre de mots différents a peu d'influence en comparaison avec le très grand nombre d'abréviations qui ont fait leur apparition. Dans toutes les formes de communication électronique où la vitesse est omniprésente, les mots fréquemment utilisés sont abrégés pour sauver du temps. Par exemple, il est normal d'écrire *pkoi* au lieu de *pourquoi*, et *p-e* à la place de *peut-être*. Toujours selon la même idée de vouloir raccourcir le temps de rédaction, des expressions entières peuvent être remplacées par quelques lettres. Par exemple, *mdr* veut dire *mort de rire* et *alp* est un raccourci pour *à la prochaine*. Puis, il y a les abréviations phonétiques dont le but est d'écrire un mot comme il se prononce. De bons exemples sont *a +*, qui signifie *à plus tard*, et *c*, qui équivaut à *c'est*. Toutes ces abréviations ont l'avantage de s'écrire beaucoup plus rapidement que ce qu'elles remplacent, et la compréhension ne s'en trouve aucunement affectée. De plus, il est souvent possible d'en déduire le sens sans grand effort même lors d'une première lecture. Ce décodage est facilité par le fait que le lecteur ne "lit" pas vraiment un courriel, mais l'"écoute visuellement". Il est donc normal d'automatiquement faire le lien entre *tk* et *en tout cas*, ou encore *oqp* et *occupé*, sans même s'en rendre compte puisque la phonétique des deux termes est très similaire.

2.3.1.4 Les majuscules

Contrairement aux autres modes de communication électronique, le courrier électronique se distingue en étant le seul à garder la même utilisation des lettres majuscules que pour les documents écrits. Alors qu'elles ont presque complètement disparu des forums et du bavardage en ligne, elles sont encore très présentes à l'intérieur des courriels. En plus d'identifier le début d'une phrase, elles servent à mettre de l'emphase sur un mot ou une expression à l'intérieure d'une phrase. Par exemple: *N'oubliez pas que la réunion n'est plus lundi mais bien MARDI.* Dans ce cas, les majuscules servent à identifier la portion la plus importante de la phrase. Il est aussi possible d'utiliser les majuscules dans le but d'obtenir une emphase verbale, pour identifier une phrase ou une expression que l'émetteur prononcerait plus fort que normalement. Par exemple, en réponse à *Es-tu encore fâché?*, un *LAISSE MOI TRANQUILLE!!!* tout en majuscule ne laisse place à aucune équivoque.

2.3.1.5 Les émoticons

Les émoticons ne font pas partie intégrante du texte, mais sont de petits dessins formés de caractères alphanumériques qui servent à indiquer les émotions de l'émetteur. La majorité de ces émoticons ont une forme de base représentant un visage couché sur le côté avec 2 yeux et une bouche, et parfois un nez. Par exemple, `:-)` et `:  )` représentent une petite joie, ou encore un petit rire discret. À partir de ce symbole de base, plusieurs autres informations non transmissibles textuellement peuvent être communiquées en changeant les caractères utilisés. Quelques exemples sont disponibles à la figure 2.4.

`; )`	un clin d'oeil	`:-(`	la tristesse
`:-P`	une grimace	`:-o`	la surprise
`|-O`	un baillement	`>:-(`	la colère
`:~(`	une larme	`:-7`	le sarcasme

Fig. 2.4: Exemples d'émoticon

Les émoticons sont loin de fournir toute l'information non verbale transmise lors des communications orales et face-à-face. Néanmoins, ils peuvent fournir suffisamment

d'informations pour mieux interpréter certains passages qui pourraient facilement être ambigus. Par exemple, la phrase *Tout le monde sait que tu n'es pas très bon dans ce domaine.* n'est pas très positive et peut mener à une escalade d'injures, même si l'émetteur ne disait cela que pour rire. Par contre, en ajoutant `;-P` à la fin de la phrase, l'émetteur s'assure que son message ne sera pas pris au pied de la lettre. En voyant l'émoticon suivant la phrase, le récepteur se rend compte qu'il n'y a pas vraiment de méchanceté de la part de l'émetteur et qu'il s'agit plutôt d'une blague. Évidemment, l'utilisation des émoticons ne règle pas tous les problèmes et peut sûrement en causer d'autres. Tout comme pour le texte libre, il est possible de mal interpréter un petit symbole de trois caractères et d'arriver à une conclusion qui n'est pas la bonne.

2.3.2 Les spécifications techniques

L'évolution du courrier électronique est géré par le Internet Mail Consortium (IMC)[3] et par l'Internet Engineering Task Force (IETF)[4]. L'IMC est une organisation internationale vouée au développement, à la promotion et à la facilité d'utilisation du courrier électronique, en particulier pour les novices. Elle est composée des entreprises reliées au courrier électronique, comme les fabriquants et les vendeurs de logiciels, les fournisseurs de services de courrier, etc. L'IETF est une organisation mondiale regroupant tous les gens concernés par l'évolution et le bon fonctionnement de l'Internet, comme les chercheurs, les programmeurs, les entreprises, etc. Elle s'occupe principalement d'analyser et de réviser les "Internet Drafts" (I-D) que ses membres soumettent. Les documents acceptés deviennent des "Request For Comments" (RFC) et servent de standards pour l'industrie. Comme il s'agit d'un processus en constante évolution, beaucoup de spécifications sont laissées à la discrétion du programmeur. Cela permet d'ajouter facilement des fonctionnalités qui n'étaient pas prévues initialement, ou encore de ne pas forcer l'intégration immédiate des additions récentes. Le résultat est un environnement souple et robuste, mais qui peut aussi donner bien des mots de tête lors d'un traitement plus complexe. Plusieurs RFC régissent présentement la structure de base

[3] www.imc.org

[4] www.ietf.org

d'un courriel [32, 38], les extensions au courrier électronique possibles grâce au format MIME[12, 13, 14, 29] et plusieurs autres aspects peu connus comme l'utilisation de caractères internationaux.

2.3.2.1 Les lignes et les caractères

Un courriel ne doit contenir que des caractères US-ASCII (codes 1 à 127), regroupés en lignes. Cela interdit en principe l'utilisation des caractères accentués, mais nous verrons plus loin comment contourner cette restriction. Chaque ligne se termine par un retour de chariot (`CR`, code 13) accompagné d'un saut de ligne (`LF`, code 10). Pour des raisons de sécurité et de robustesse lors du transport, une ligne ne devrait pas dépasser 78 caractères, et ne doit en aucun cas en avoir plus de 998. Ces deux limites ne tiennent pas compte du `CRLF`[5] à la fin de chaque ligne. Pour contourner ces limites, il est possible de séparer une ligne en insérant un `CRLF` juste avant un espace (code 32) ou une tabulation (code 9). Ces `CRLF` doivent être enlevés avant le traitement et remis si le courriel doit poursuivre son chemin. Cette technique, appelée "line folding", permet de traiter des lignes plus longues que les limites, tout en respectant les restrictions lors du transport. Dans l'exemple de la figure 2.5, les trois phrases sont équivalentes, et deviennent identiques lorsque le "line folding" est enlevé (les `CRLF` ne sont visibles que pour indiquer leur emplacement).

```
Subject:  retour sur la présentation et avancement du projetCRLF

Subject:  retour sur laCRLF
 présentation et avancement du projetCRLF

Subject:CRLF
 retour sur la présentation etCRLF
 avancement du projetCRLF
```

Fig. 2.5: Exemple de "line folding"

[5] Nous utilisons le terme `CRLF` pour désigner un `CR` immédiatement suivi d'un `LF`.

2.3.2.2 L'en-tête

Un courriel est divisé en deux sections: l'en-tête et le corps. L'en-tête vient en premier, suivi d'une ligne vide qui ne contient qu'un CRLF, et optionnellement du corps. L'en-tête contient les informations reliées au transport du courriel, telles la date d'envoi, l'émetteur, le ou les récepteurs, les adresses IP rencontrées lors du trajet, etc. Elle est constituée de plusieurs champs ayant tous la même syntaxe de base: le nom du champ, un : et la valeur du champ, qui peut être structurée ou non.

2.3.2.3 Le corps

Si le format de l'en-tête est défini par un format rigide, il en est tout autrement pour le corps d'un courriel. Celui-ci est seulement une suite de caractères formant le message transmis. Les seules restrictions qu'il faut observer sont celles concernant les caractères et les lignes, décrites précédemment. Il ne s'agit pas là d'un problème en soi, mais cette souplesse peut grandement compliquer le traitement automatique de texte libre sur le corps d'un courriel.

Pour faciliter les conversations, les logiciels de courrier électronique sont pour la plupart munis d'une option pour inclure le texte du courriel original dans la réponse. Cette tendance s'est même transformée en une norme de facto et il est rare de répondre à un courriel sans y inclure au moins une partie du message original. Mais comme il n'existe pas de règles pour clairement identifier l'ancienne portion de texte de la nouvelle, chaque logiciel a une manière par défaut qui lui est propre. Par exemple, Pine[6] précède le texte original d'une ligne indiquant le jour, la date et l'émetteur. En plus, il ajoute un > devant chacune des lignes du message antérieur, tel que montré à la figure 2.6.

Selon le même principe, Netscape[7] mentionne aussi l'auteur du message original, en plus d'ajouter un |, au lieu d'un >, devant chaque ligne de texte du courriel antérieur, comme la figure 2.7 l'indique.

[6] www.washington.edu/pine

[7] browsers.netscape.com

```
On Tue, 15 Jan 2002, Alice wrote:

> Bob,
> Peux-tu m'envoyer une copie du rapport?

Je ne l'ai pas.  Demande a Claude.
```

Fig. 2.6: Exemple d'indention d'un courriel de réponse avec Pine

```
Bob wrote:

| On Tue, 15 Jan 2002, Alice wrote:
|
|> Bob,
|> Peux-tu m'envoyer une copie du rapport?
|
| Je ne l'ai pas.  Demande a Claude.

Claude,
Peux-tu m'envoyer ça le plus vite possible.
```

Fig. 2.7: Exemple d'indention d'un courriel de réponse avec Netscape

Outlook[8] n'ajoute aucun indicateur d'indentation pour identifier les lignes originales de celles qui ont été ajoutées. Il les sépare toutefois avec une ligne spéciale, comme dans l'exemple de la figure 2.8.

Pour les exemples précédents, l'identification du nouveau texte par un être humain se fait presque instantanément. Même pour un ordinateur, quelques mécanismes pour traiter les expressions régulières sont suffisants pour correctement repérer les ajouts. Il suffit simplement de répertorier les variantes utilisées par les programmes les plus populaires, pour autant que les caractères et expressions identifiant les ajouts ne changent pas constamment.

Toutefois, comme il est possible de modifier le texte contenu dans le message original et d'y insérer de nouveaux passages, un courriel peut se transformer en véritable casse-

[8] www.microsoft.com/office/outlook

```
Bob wrote:

| On Tue, 15 Jan 2002, Alice wrote:
|
|> Bob,
|> Peux-tu m'envoyer une copie du rapport?
|
| Je ne l'ai pas.  Demande a Claude.

Claude,
Peux-tu m'envoyer ça le plus vite possible.

//=======================================

Le voici en attachement.
```

Fig. 2.8: Exemple d'indention d'un courriel de réponse avec Outlook

tête. Par exemple, après plusieurs échanges entre trois personnes utilisant des logiciels différents, un courriel peut facilement ressembler à l'exemple de la figure 2.9.

Il n'est pas trop difficile pour un être humain de comprendre la progression de cette conversation, mais il s'agit d'un exemple plutôt simple. De plus, il est important de remarquer que Bob (dont le logiciel n'indente pas le texte original) répond toujours à la fin du courriel après une ligne de démarcation. En comparaison, le cas de la figure 2.10 est un peu plus confus.

Encore une fois, il est relativement simple pour un être humain de suivre la discussion. Par contre, si le courriel doit être analysé par un logiciel traitant du texte libre, les problèmes commencent à s'accumuler. S'il n'y a pas de balises pour identifier ce qui est nouveau, il y a trop de possibilités pour qu'un programme puisse déterminer correctement les ajouts. Pour avoir une idée de la complexité impliquée, il faut premièrement regarder combien de logiciels différents (avec la possibilité d'avoir autant de manières différentes de séparer les ajouts) sont disponibles. En plus des 3 logiciels mentionnés précédemment, les plus populaires (selon les statistiques de téléchargement sur C|Net[9])

[9] www.cnet.com

```
Claude wrote:

| On Thu, 17 Jan 2002, Bob wrote:
|
|> Bob wrote:
|>
|>| On Tue, 15 Jan 2002, Alice wrote:
|>|
|>|> J'ai quelques questions:
|>|>
|>|> 1) Est-ce que la rapport est complété?
|>|
|>| Ma partie est complétée.  J'attends celle de Bob.
|>|
|>|> 2) Si oui, est-ce que tu peux m'en envoyer une copie?
|>|>
|>|> 3) Sinon, quand pensez-vous avoir fini?
|>|>
|>|> 4) Où puis-je trouver les sources de votre programme?
|>|
|>| Je ne me rappelle plus.  Bob?
|>|
|>| //=======================================
|>|
|>| 3) Je viens de terminer la mienne.  Je l'envoie a Claude.
|>
|> Ok.  Faîtes moi signe quand vous les aurez jumelées.
|>
|>| 4) Les sources sont disponibles dans mon compte.
|>
|> Ça ne compile pas et je ne comprends pas pourquoi.
|>
|> //=======================================
|>
|> Oups!  Ce n'etait pas les bons fichiers.  C'est correct.
|
| J'ai finalisé le rapport.  Le voici en attachement.

Merci.
```

Fig. 2.9: Exemple d'indentation de plusieurs réponses dans un courriel

```
Merci.

Claude wrote:

| On Thu, 17 Jan 2002, Bob wrote:
|
|> Bob wrote:
|>
|>| On Tue, 15 Jan 2002, Alice wrote:
|>|
|>|> J'ai quelques questions:
|>|>
|>|> 1) Est-ce que la rapport est complété?
|>|
|>| Ma partie est complétée.  J'attends celle de Bob.
|>|
|>|> 2) Si oui, est-ce que tu peux m'en envoyer une copie?
|>|>
|>|> 3) Sinon, quand pensez-vous avoir fini?
|>|
|>| Je viens de terminer la mienne.  Je l'envoie a Claude.
|>
|> Ok.  Faites moi signe quand vous les aurez jumelées.
|>
|>|> 4) Où puis-je trouver les sources de votre programme?
|>|
|>| Je ne me rappelle plus.  Bob?
|>|
|>| Les sources sont disponibles dans mon compte.
|>
|> Ça ne compile pas et je ne comprends pas pourquoi.
|>
|> Oups!  Ce n'etait pas les bons fichiers.  C'est correct.
|
| J'ai finalisé le rapport.  Le voici en attachement.
```

Fig. 2.10: Exemple d'indentation obscure de plusieurs réponses dans un courriel

sont IncrediMail Xe[10], ePrompter[11], Eudora[12], Pegasus Mail[13] et AllegroMail[14]. Cela

[10] www.incredimail.com

[11] www.eprompter.com

[12] www.eudora.com

[13] www.pmail.com

fait donc huit logiciels distincts, sans compter tous les autres qui n'ont pas une grande part du marché. Ensuite, il faut tenir compte que les démarcations indiquant l'émetteur, le jour, la date, etc. peuvent être en anglais, en français ou en n'importe quelle autre langue ou dialecte. Puis, il existe la possibilité que les gens enlèvent ces indications pour alléger le texte, ou encore que le message contienne l'une de ces expressions, comme dans l'exemple de la figure 2.11.

```
> I had an argument with Alice yesterday.
> I wrote her an email to ask why she acted so strange.
> Then, Alice wrote:
> "It's none of your business!"

That's too bad...
```

Fig. 2.11: Exemple de problème avec les balises de réponse dans un courriel

Il y a aussi l'emplacement de la réponse, qui peut être avant, après ou même à l'intérieur du message original. En combinant tous ces facteurs, il est impossible pour un logiciel de prévoir tous ces cas, et donc d'extraire les nouveaux passages avec une efficacité acceptable. Un logiciel traitant du texte libre sur du courrier électronique ne peut donc pas cibler uniquement la partie la plus récente d'un courriel. Cela peut fournir une quantité non négligeable de bruit et peut réduire les performances des logiciels.

2.3.2.4 Le format MIME

Avec l'importance croissante du courrier électronique, les limitations dues aux caractères permis sont devenues plus évidentes. Le choix de la langue est restreint à l'anglais et aux autres langues exprimables avec l'ensemble de caractères US-ASCII. Cela ferme la porte à plusieurs langues non anglaises (dont le français), qui sont utilisées par la majorité de la population mondiale. Cet ensemble restreint de caractères cause aussi des problèmes pour le transport de documents en attachement. Si l'encodage natif d'un document dépasse l'ensemble US-ASCII, il faut d'abord convertir son format d'encodage pour respecter les spécifications du courrier électronique. À l'arrivée,

[14] www.allegromail.com

il faut le reconvertir au format normal du document pour y avoir accès. Pour contourner ces limitations, les “Multipurpose Internet Mail Extensions” (MIME) ont fait leur apparition.

Pour rester conforme avec les spécifications antérieures, un courriel utilisant un format MIME doit contenir quelques champs d'en-tête supplémentaires. Le premier de ces champs doit être `MIME-Version`, suivi du numéro de version. Ensuite, un champ `Content-Type` identifie le type de document que contient le courriel. Si ce document n'est pas encodé avec l'ensemble US-ASCII, le champ `Content-Transfer-Encoding` doit être inséré pour spécifier l'encodage. Le logiciel de courrier se sert de ces informations pour interpréter le corps du courriel, qui ne représente plus uniquement du texte. L'exemple de la figure 2.12 est un courriel envoyé en format HTML.

```
From:  Alice <alice@iro.umontreal.ca
Date:  Tue, 22 Jan 2002, 23:28:11 -0500
To:  Bob <bob@iro.umontreal.ca>
...
MIME-Version:  1.0
Content-Type:  text/html; charset=us-ascii
Content-Transfer-Encoding:  7bit

<html>
<head></head>
<body>
<h1><Rapport</h1>
Ceci est un rapport préliminaire
...
</body>
</html>
```

Fig. 2.12: Exemple de courriel en HTML

Il est aussi possible pour un courriel de contenir plusieurs formats MIME différents. Le champ `Content-Type` contient alors la mention `multipart/mixed` ainsi qu'une expression spéciale pour l'attribut `boundary`. Cette expression est utilisée pour délimiter les sections, chacune étant associée à l'utilisation de l'un des formats MIME du courriel. Chaque section doit être structurée comme un courriel de base, sauf que l'en-tête

n'est pas obligatoire et seuls les champs relatifs au format MIME sont considérés. Donc, chaque section est tenue d'avoir les champs d'en-tête qui conviennent au type de format MIME contenu dans la section. Si aucun champ d'en-tête n'est présent, le logiciel doit considérer la section comme un courriel normal, c'est-à-dire du texte limité à l'ensemble de caractères US-ASCII. Avant chaque section, une ligne en indique le début. Elle doit contenir `--`, suivi de l'expression spéciale définie dans l'en-tête du courriel. À la fin de la dernière section, cette ligne doit aussi apparaître, mais terminée par `--`. Aussi, le texte apparaissant avant la première section peut être ignoré. L'exemple de la figure 2.13 est un courriel en HTML qui contient un fichier zip en attachement.

Comme chaque section est l'équivalent d'un courriel, il est possible d'avoir une section contenant plusieurs formats MIME. La situation est la même que s'il s'agissait d'un courriel, mais il faut s'assurer que les expressions marquant les sections ne sont pas les mêmes.

Pour un logiciel traitant du texte libre, il n'est pas trop compliqué de repérer les sections qui comportent du texte. Repérer ces sections et ne travailler qu'avec elles peut réduire considérablement le temps de traitement, surtout lorsque le document attaché fait plusieurs Mo.

2.3.3 La classification du courrier électronique

Selon le type de courrier électronique (corporatif, personnel, etc.) à traiter, toutes les particularités décrites dans la section précédente ne s'appliquent pas. Il est peu probable qu'un client emploie des émoticons dans un courriel envoyé au département des relations aux investisseurs. Dans le même ordre d'idée, cette sorte de courriels à nature sérieuse ne devrait pas contenir beaucoup de fautes d'orthographe. Par contre, il risque fort d'y avoir de nombreux courriels en HTML, et avec des attachements. Cela implique du traitement supplémentaire et dans certains cas, des problèmes difficilement résolubles. Par exemple, la détection de la dernière partie ajoutée au courriel lors d'une réponse. Nous analyserons les caractéristiques des courriels de BCE dans le chapitre 4 pour déterminer lesquelles nous devrons tenir compte dans nos expériences. Mais avant, nous expliquons dans le prochain chapitre l'environnement technique de BCE

```
From:  Alice <alice@iro.umontreal.ca
Date:  Tue, 22 Jan 2002, 23:28:11 -0500
To:  Bob <bob@iro.umontreal.ca>
...
MIME-Version:  1.0
Content-Type:  multipart/mixed;
boundary="040006000601080907050203"

This is a multi-part message in MIME format.
--040006000601080907050203
Content-Type:  text/html; charset=us-ascii
Content-Transfer-Encoding:  7bit

<html>
<head></head>
<body>
<h1><Rapport</h1>
Ceci est un rapport préliminaire
...
</body>
</html>

--040006000601080907050203
Content-Type:  application/x-zip-compressed;
name="rapport.zip"
Content-Transfer-Encoding:  base64
Content-Disposition:  inline;
filename="rapport.zip"

UEsDBBQAAAAIAACEsSoYb5jcoIMEAOKGBAAMAAAAX2luc3QzMmkuZXhfbPxlV
...
8OAOwbW4a3B3iru7lKBFgrtTihPcW5xCaXFrgQYtxaXQAC3QAi1QCu/vf9/rf
--040006000601080907050203--
```

Fig. 2.13: Exemple de courriel en HTML contenant un fichier zip en attachement

afin de trouver s'il y a d'autres détails à prendre en considération pour la conception de Merkure.

Chapitre 3

L'environnement de BCE

Dans le chapitre précédent, nous avons présenté plusieurs travaux connexes et expliqué les différences entre la classification de textes et la classification du courrier électronique. Il s'agit surtout de considérations théoriques. Cependant, comme Merkure est conçu pour l'entreprise BCE, nous devons aussi tenir compte de l'implantation. Dans ce chapitre, nous décrivons l'environnement de BCE, les considérations pratiques qui lui sont reliées, ainsi que l'élaboration et l'implantation d'un prototype.

3.1 L'architecture en place

Plusieurs choix s'offrent aux clients qui veulent communiquer avec BCE. En plus des traditionnels moyens de communication, BCE maintient un site web[1] et des adresses de courrier électronique pour ses clients branchés. Cet arrangement informatique a subi quelques modifications au cours de l'avancement de Merkure. Ces changements ont eu lieu suite à l'acquisition de Montreal Trust, l'agent de transfert de BCE, par Computershare[2] en février 2001. Cette société de fiducie offre de nombreux services reliés à l'investissement, autant pour les entreprises que pour les investisseurs:

- gestion d'actions corporatives, incluant l'émission de nouveaux capitaux et le rachat d'actions

[1] www.bce.ca

[2] www.computershare.com

- gestion de plans d'achat direct et de plans de réinvestissement des dividendes
- gestion de listes de distribution
- gestion et analyse des registres
- services en-ligne pour les entreprises et les investisseurs

3.1.1 Le point de vue du client

BCE possède un site web bilingue bien organisé où il est facile de s'y retrouver. La version anglaise affiche exactement le même contenu que celle en français. Le site est séparé en sections et sous-sections bien identifiées, et le passage d'une section à une autre s'effectue à l'aide d'une barre de navigation simple et concise. Pour les clients de BCE, deux sections offrent des façons pour communiquer avec l'entreprise. La première est entièrement consacrée aux investisseurs, enregistrés ou potentiels. Pour des raisons pratiques, toutes les informations relatives aux investisseurs sont regroupées dans cette section, incluant les moyens pour rejoindre les préposés de BCE. La deuxième section est celle des informations rapides, contenant une page avec les contacts de BCE.

Dans les deux sections, outre les numéros de téléphones et les adresses civiques, des adresses électroniques et des formulaires sont disponibles pour rejoindre les préposés de BCE. Initialement, deux adresses électroniques et un formulaire de commentaire étaient en fonction. Ces options sont toujours présentes aujourd'hui. Les deux adresses originelles pour contacter BCE sont:

investor.relations@bce.ca Service des relations aux investisseurs. Pour obtenir des informations financières au sujet de BCE, inluant les demandes de rapports, les régimes de réinvestissement des dividendes et d'achat direct, etc.

bcecomms@bce.ca Pour obtenir des renseignements généraux sur BCE.

Le formulaire simple, montré à la figure 3.1, permet d'écrire un commentaire ou une question. Pour obtenir une réponse, il existe un champ pour inscrire son adresse de courrier électronique. Un message à l'écran confirme la soumission du formulaire,

même lorsque le message est vide. Ce formulaire n'est disponible que dans la section des informations rapides, comparativement aux adresses qui sont fournies dans les deux sections.

Fig. 3.1: Formulaire de commentaire

Pendant les mois de juin et juillet 2001, plusieurs modifications ont été apportées au site web de BCE. Ces changements ont contribué à l'apparence et aux fonctionnalités du site web tel qu'il est en ce moment. Pour ce qui est des moyens de communiquer avec BCE, quelques ajouts ont été effectués, et ce qui existait est resté inchangé. L'adresse investor.relations@bce.ca a maintenant une adresse jumelle: relations.investisseurs@bce.ca. Cette dernière est visible sur la version française du site tandis que la première est toujours disponible sur la version anglaise. Aussi, une nouvelle adresse électronique, bce@computershare.com, permet de rejoindre l'agent de transfert de BCE. Il est suggéré d'utiliser cette adresse pour toutes les demandes à titre d'actionnaire:

- changement d'adresse
- changement sur les comptes
- paiements de dividendes

- perte de certificats
- régime de réinvestissement des dividendes et d'achat d'actions

En plus, un nouveau formulaire de demande de documents, illustré à la figure 3.2, permet d'effectuer une requête des principaux documents corporatifs tels que les rapports annuels et trimestriels. Il est présent dans les deux sections. Une procédure vérifie que toutes les informations sont complètes lors de la soumission du formulaire et le cas échéant, signale celles qui sont incorrectes. Lorsque la soumission est acceptée, un message à l'écran confirme que les informations ont été envoyées.

3.1.2 Le fonctionnement interne

BCE fait présentement appel aux services de Stylus[3], une entreprise canadienne spécialisée dans les communications corporatives. Quelques-uns des services offerts par Stylus sont:

- gestion d'un programme de relations avec les investisseurs
- rédaction des contenus (rapports annuels et trimestriels, communiqués de presse, etc.)
- préparation de réunions et de présentations
- gestion d'un site web

Parmi ces services, seule la gestion du site web et du courrier électronique nous concerne dans notre travail. Pour des raisons pratiques, les serveurs de Stylus servent de pont entre les ordinateurs de BCE et le reste de l'internet. Ces serveurs hébergent le site web de BCE et effectuent la redirection de leur courrier électronique. La figure 3.3 illustre le cheminement des courriels envoyés à BCE.

Tous ces courriels se dirigent dans un premier temps vers les serveurs de Stylus. À cet endroit, ils sont redirigés vers les serveurs de BCE à l'aide du logiciel qmail[4]. Il

[3] www.stylus.ca

[4] www.qmail.org

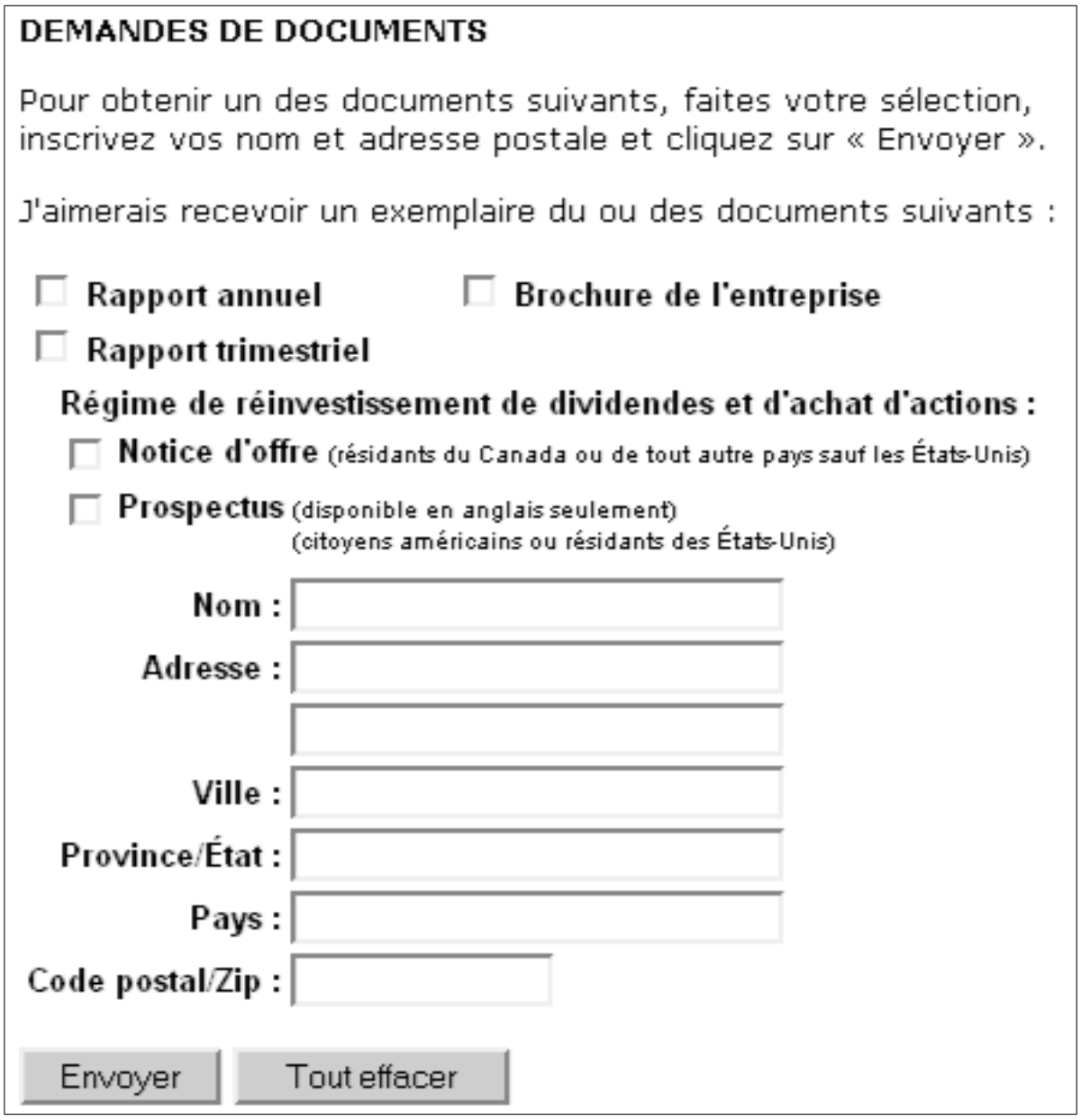

Fig. 3.2: Formulaire de demande de documents

s'agit d'un "Mail Transfer Agent", c'est-à-dire un logiciel gérant le transfert du courrier électronique. Ce type de programme est surtout utilisé pour les serveurs intermédiaires comme ceux de Stylus. Qmail offre plusieurs fonctionnalités:

- redirection de courriels
- gestion de listes de distribution

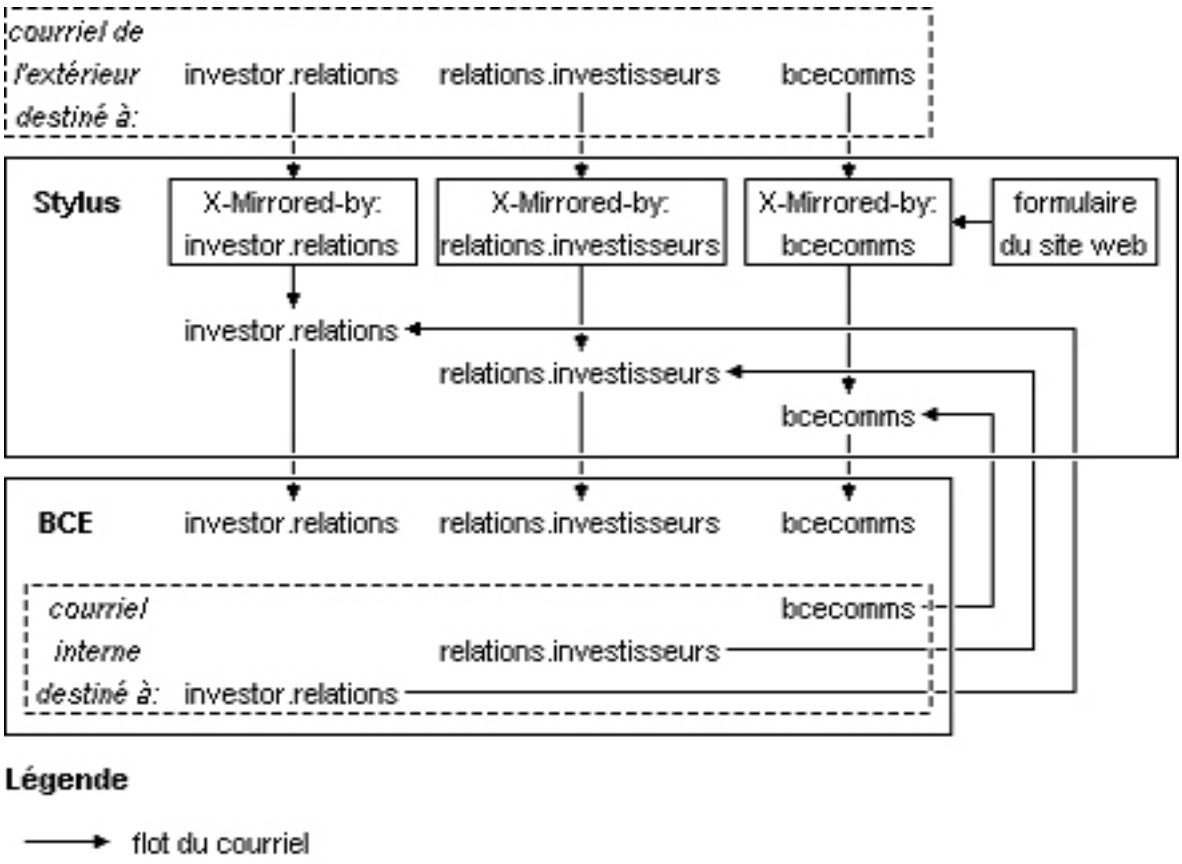

Fig. 3.3: Cheminement des courriels envoyés à BCE

- service de livraison locale des courriels
- service SMTP
- service POP3, servant à héberger des courriels sur un serveur en attentant qu'un client extérieur les retrouve

Si le courriel provient de l'extérieur, c'est-à-dire qu'il ne vient pas du domaine bce.ca, la copie envoyée à BCE contiendra un champ d'en-tête supplémentaire. Le champ `X-Mirrored-by`, bien que non officiel, est fréquemment utilisé pour indiquer le nom du compte usager qui a redirigé le message. Dans le cas de BCE, il existe un compte pour chaque adresse électronique, et la valeur du champ `X-Mirrored-by` correspond à l'adresse de destination. Un courriel envoyé à relations.investisseurs@bce.ca contiendra donc la mention `X-Mirrored-by:` relations.investisseurs@smtp.stylus.ca.

Les informations entrées dans les formulaires sont aussi transmises par courrier électronique. Lorsque la soumission d'un formulaire est acceptée, un courriel est généré automatiquement et envoyé à bcecomms@bce.ca. Tout comme ceux venant de l'extérieur, ces courriels contiennent la mention `X-Mirrored-by:` bcecomms@smtp.stylus.ca. Puisqu'ils sont générés automatiquement, ils sont facilement identifiables et suivent tous le même format. Le sujet du courriel indique le type de formulaire, la date et l'heure de la soumission. Le formulaire de commentaire est défini par *bce-comments*, et le formulaire de demande de documents est identifié par *bce-documents*. Le corps du courriel est structuré comme une liste de champs accompagnés de leurs valeurs. Les deux formulaires ont sept champs en commun: `request`, `url`, `referer`, `client`, `domain`, `language` et `e-mail`. Le premier champ possède la même valeur que le sujet du message, soit *bce-comments* ou *bce-documents*. Le champ `referer` contient l'adresse de la page web où se situe le formulaire et `url` identifie l'adresse complète du formulaire. Le champ `client` donne des informations sur le fureteur utilisé par le client pour visionner la page web et `domain` identifie l'adresse IP du client. Les deux derniers champs fournissent respectivement la version (française ou anglaise) du site web visitée et l'adresse électronique du client. Dans le cas du formulaire de commentaire, le message en format libre suit la liste de champs tels qu'illustré à la figure 3.4.

Quant à lui, le formulaire de demande de documents étire la liste de champs avec les choix de documents et les informations personnelles du client comme l'indique la figure 3.5.

Les deux options concernant le régime de réinvestissement des dividendes et d'achat d'actions ne se retrouvent pas dans le courriel. Selon le même principe que pour l'adresse bce@computershare.com, ce qui doit être traité par Computershare ne transite pas par BCE et leur est directement acheminé. Un second courriel, similaire à celui destiné à bcecomms@bce.ca, est donc envoyé à bce@computershare.com. La seule différence est que les 3 champs `annual`, `quaterly` et `brochure` sont remplacés par `circular` et `prospectus`.

```
Date:  16 Feb 2002 19:41:16 -0000
From:  duboisju@iro.umontreal.ca
To:  bcecomms@bce.ca
Subject:  bce-comments,16/2/2002,14:41:16

 request:  bce-comments
     url:  http://www.bce.ca/fr/quickinfo/contacts/bce/...
 referer:  http://www.bce.ca/fr/quickinfo/contacts/bce/
  client:  Mozilla/4.0 (compatible; MSIE 5.5; Windows 98; Win 9x)
  domain:  193.194.79.65
language:  French

email:  duboisju@iro.umontreal.ca

Je trouve votre site web très bien fait, mais il contient trop d'ima-
ges.  Ceux qui ont une connection plus lente doivent être patients.
```

Fig. 3.4: Exemple de courriel généré par le formulaire de commentaire

3.2 L'intégration du système

Il faut considérer avec soin les choix pour l'intégration du système dans l'environnement de BCE. Que ce soit pour le type et le déclenchement du traitement, la supervision des suivis ou l'emplacement du système, les choix que nous ferons influenceront la conception et l'implantation de presque tous les modules de Merkure. Une erreur d'adaptation à l'environnement de BCE pourrait dans le pire des cas causer la perte de nombreux courriels.

3.2.1 Le type de traitement

Le traitement des courriels peut se faire un à un, dès qu'ils arrivent à destination, ou en série. Cette deuxième façon semble à priori plus simple à implanter, en partie parce qu'il n'est pas nécessaire d'installer et de configurer un logiciel de gestion de courrier électronique avancé. Pour traiter un courriel lors de son arrivée, un tel programme doit pouvoir gérer des actions conditionnelles. En premier lieu, il doit être possible de configurer des filtres basés sur les champs des courriels. Par exemple, il peut s'avérer

```
Date:  16 Feb 2002 19:39:49 -0000
From:  No email address provided <webmaster@bce.ca>
To:  bcecomms@bce.ca
Subject:  bce-documents,16/2/2002,14:39:49

   request:  bce-documents
       url:  http://www.bce.ca/fr/quickinfo/contacts/bce/...
   referer:  http://www.bce.ca/fr/quickinfo/contacts/bce/
    client:  Mozilla/4.0 (compatible; MSIE 5.5; Windows 98; Win 9x)
    domain:  193.194.79.65
  language:  French

    annual:  yes
 quarterly:
  brochure:

      name:  Julien Dubois
  address1:  2920 chemin de la Tour
  address2:
      city:  Montreal
prov_state:  Quebec
   country:  Canada
 post_code:  H3T 1J4
```

Fig. 3.5: Exemple de courriel généré par le formulaire de demande de documents

utile d'ignorer les courriels provenant d'une certaine adresse, que ce soit une adresse interne ou une adresse externe qui n'envoie que des courriels non sollicités. Dans ce cas, un simple filtre tel qu'à la figure 3.6 suffit.

```
SI
  le champ "To" contient l'expression @bce.ca
ALORS
  sauvegarder le courriel dans le répertoire "Courrier interne"
```

Fig. 3.6: Exemple de pseudo-code pour un filtre basé sur le champ d'un courriel

Puisque ce genre de filtre existe dans la plupart des logiciels de courrier électronique, il ne s'agit pas là d'un problème. Par contre, le traitement automatique ne peut pas

reposer uniquement sur des filtres restreints aux champs des courriels. De plus, les actions possibles ne doivent pas être limitées à la sauvegarde, à la suppression et à la redirection des courriels. Il faut aussi être en mesure de lancer l'exécution d'un autre programme avec le courriel reçu en entrée.

Avec un logiciel capable d'interpréter une structure condition-action et de démarrer l'exécution d'autres programmes, le besoin de supervision du système devient presque inexistant. Et si le niveau de confiance du système est jugé assez élevé, il est même possible de traiter certains courriels automatiquement sans aucune intervention humaine.

Cependant, il faut être prudent avec un tel arrangement. L'inconvénient le plus susceptible d'apparaître est un embouteillage au niveau des disques parce que de toutes les opérations gérées par le processeur, les accès aux disques sont les plus lentes. Si un courriel est traité dès son arrivée et que son traitement s'éternise suite à de trop nombreux accès aux disques, il est possible que d'autres courriels arrivent avant que le traitement du premier ne soit terminé. Mais comme le logiciel de gestion est mobilisé par le premier courriel reçu, il ne peut s'occuper des autres immédiatement, ce qui peut causer une ralentissement général du système et en pire cas, la perte de courriels. Même s'il s'agit de conditions extrêmes, il est important de considérer ce problème potentiel qui peut survenir si un très grand nombre de courriels sont reçus quotidiennement. En outre, plus le traitement est complexe, plus il y a des accès aux disques. Lorsqu'un courriel arrive et qu'il est simplement sauvegardé, les chances d'avoir des problèmes sont presques nulles. Par contre, si le logiciel de gestion doit envoyer le courriel vers un script effectuant un prétraitement, rediriger la sortie vers un second programme qui doit parcourir une énorme base de données pour fournir un suivi, et reprendre ce suivi pour finalement le sauvegarder, les risques sont plus grands, sans toutefois être énormes.

Pour régler cette faille potentielle, il est possible de simplement sauvegarder les courriels lors de leur arrivée, et de partir un traitement en série lorsque la situation le demande. Cette façon de procéder nous assure qu'aucun courriel ne sera perdu à cause de temps d'exécution trop longs. Toutefois, le traitement en série peut entraîner une baisse de la qualité du service car les courriels ne sont pas répondus immédiatement mais à certains intervalles. Cela peut nuire lorsqu'un courriel est urgent, mais pas

nécessairement plus que si un préposé s'en occupe. En effet, il serait surprenant qu'un préposé réponde sans cesse aux courriels au fur et à mesure qu'ils arrivent. Aussi, si le niveau de confiance du système n'est pas assez élevé et qu'un préposé doivent vérifier les réponses avant de les envoyer, le traitement en série n'apporte pas plus d'inconvénients que le traitement immédiat des courriels.

3.2.2 Le déclenchement du traitement

Le déclenchement du traitement, qu'il soit un à un ou en série, peut être manuel ou automatique. Il peut paraître étrange de suggérer qu'il faille démarrer manuellement le traitement d'un projet de réponse *automatique* au courrier électronique, mais cela peut s'avérer utile. Si très peu de ressources informatiques sont disponibles pour faire fonctionner le système et que son utilisation interfère avec d'autres applications plus importantes, il peut être souhaitable de pouvoir lancer son exécution seulement aux moments opportuns. Cependant, un déclenchement manuel implique une plus grande connaissance du système de la part des préposés. Habituellement, cela résulte en problèmes supplémentaires car les préposés ne sont formés qu'au strict minimum pour utiliser leurs logiciels.

3.2.3 La supervision des suivis

Bien que Merkure soit un projet de réponse automatique au courrier électronique, une supervision humaine est fortement conseillée. Le traitement des langues naturelles n'est pas encore assez perfectionné pour laisser un programme répondre automatiquement à tous les courriels qu'une entreprise peut recevoir de ses clients. Le logiciel de réponse automatique aux courriels ne doit pas actuellement être considéré comme un remplaçant aux préposés mais plutôt comme une aide pour réduire leur tâche. En outre, il n'y a rien de plus frustrant pour un client que recevoir une réponse qui ne correspond pas à la question qu'il a envoyée. Il est donc important que les courriels traités automatiquement par le système soient accompagnés d'un niveau de confiance extrêmement élevé. Et même en utilisant les meilleures techniques disponibles, il y a toujours un risque d'erreur

associé au traitement du texte libre. Pour ces raisons, il est préférable de soumettre les suivis suggérés par le système à un préposé.

Une première manière est que le système propose un suivi au courriel et le place dans un répertoire “à envoyer”. Le préposé n'a donc qu'à ouvrir son logiciel de courrier électronique, lire le courriel, apporter des modifications au suivi s'il y a lieu et l'envoyer. Les modifications peuvent être au niveau de la réponse comme telle, qui peut être inexacte. Elles peuvent aussi se situer au niveau du type de suivi. Par exemple, si le courriel contient un message de remerciement et que le suivi suggéré est de retourner une réponse au client, le type de suivi est incorrect. La correction est donc d'enlever la réponse textuelle et de remplacer l'envoi du courriel à l'adresse électronique du client par la sauvegarde du courriel sans action supplémentaire. Lorsque le préposé approuve le suivi final, le suivi (et la réponse s'il y a lieu) est introduit dans la base de données pour le raisonnement à base de cas. Le problème avec cette approche vient des complications dues au départage du message original et de la réponse que nous avons vues dans le chapitre précédent. Une solution plus adéquate pour la mise à jour de la base de données est l'utilisation d'un programme de courrier électronique qui affiche le courriel original, le suivi suggéré et la réponse proposée s'il y a lieu. Comme pour la première approche, le préposé peut corriger le suivi et la réponse s'il le faut, avant d'approuver le tout. À ce moment, le programme effectue le suivi approprié et s'il s'agit d'un courriel à retourner au client, inclut la réponse dans le courriel. Cette approche facilite l'intégration des messages et des suivis dans la base de données, mais implique la conception d'un nouveau logiciel ou l'adaptation d'un produit existant.

Une autre source de préoccupation est le nombre de préposés à pouvoir approuver des suivis en même temps. Si plus d'un préposé peut le faire, il faut prévoir des mécanismes d'exclusion afin d'éviter des réponses doubles ou triples.

3.2.4 L'emplacement du système

Puisque qu'il y a présentement deux serveurs de courrier électronique, l'un chez Stylus et l'autre chez BCE, nous devons décider où installer notre système. Dans le premier cas, la gestion se déroule sur le serveur de Stylus. Comme il s'agit d'une petite entreprise (ou

du moins, beaucoup plus petite que BCE), il y a moins de niveaux de décision impliqués et par conséquent, nos échanges avec leurs administrateurs se font plus rapidement et les délais d'implantation sont moins longs. Cependant, comme Merkure est un projet pour BCE et que Stylus a plusieurs autres clients à s'occuper, il n'est pas clair s'il y a vraiment un gain à faire au niveau des communications et des délais. Aussi, il n'est pas impossible que BCE change son fournisseur de services de relations aux investisseurs. Devant une telle situation, il faudrait recommencer une bonne partie de l'implantation chez le nouveau consultant ou directement chez BCE. De plus, rien ne garantit que les serveurs et logiciels utilisés à ce nouvel emplacement seront les mêmes que chez Stylus. Il faudrait donc adapter le système au nouvel environnement, et dans ce genre de situation, il est très rare qu'aucune complication ne survienne.

Il ne faut pas non plus négliger l'interaction entre le système et les préposés. Puisqu'une base de données doit être mise à jour régulièrement, cela ne doit pas devenir une source d'ennuis. Un système installé chez BCE comporte très peu de problèmes car tout est local. Chez Stylus, le système doit prévoir des mécanismes pour transférer les messages et les suivis proposés chez BCE et pour retourner les suivis approuvés chez Stylus. Il n'y a rien de bien compliqué ici non plus, mais il faut tout de même considérer le temps de conception et d'implantation supplémentaire requis.

3.3 Implantation d'un prototype

3.3.1 Le serveur du RALI

Nous avons décidé de bâtir un prototype simplifié de Merkure dès les premiers mois du projet afin de régler le plus rapidement possible une bonne partie des problèmes reliés à l'implantation. Nous avons choisi de monter cette coquille du système en vue de l'installer chez Stylus. Nous appelons ça une coquille puisqu'il s'agit seulement de mettre sur pied un programme effectuant un traitement automatique sur le courrier électronique. Le traitement comme tel n'est pas si important, en autant que le prototype soit en mesure de recevoir les courriels, de les passer en entrée à un module, et de récupérer la réponse pour un usage futur. Lorsque cela fonctionne parfaitement, il n'est

pas difficile de remplacer un module par un autre s'ils observent les mêmes conditions d'entrée et de sortie.

Dans un premier temps, nous avons modifié un ordinateur du RALI pour le transformer en une copie du serveur de courrier électronique de Stylus. Comme nous le verrons plus en détail dans le prochain chapitre, Stylus nous envoie les courriels destinés aux adresses investor.relations@bce.ca, relations.investisseurs@bce.ca et bcecomms@bce.ca, et ceux en provenance de bcecomms@bell.ca. Il s'agit du corpus BCE-4, soit le quatrième que BCE nous a fournit. En plus d'augmenter notre banque de messages, ce flot continu des messages quotidiens de BCE nous permet de monter notre prototype au RALI en toute quiétude, sans avoir à impliquer Stylus à chacune de nos modifications. Aussi, il n'y a aucune chance de perturber les communications entre BCE et ses clients en travaillant avec une copie des messages.

Nous avons tenté de recréer leur serveur le plus fidélement possible, mais dû à certaines contraintes, les deux machines ne sont pas identiques. Les deux ordinateurs de marque Sun[5] ont des versions différentes du système d'exploitation SunOS. Le nôtre correspond à la version 5.8 tandis que celui installé sur le serveur de Stylus affiche la version 5.7. Une deuxième différence se trouve au niveau du compilateur gcc[6] et de ses librairies. Puisque les versions ne différaient que très peu, nous n'y avons pas apporté une très grande attention.

3.3.2 Le logiciel procmail

Pour gérer les messages entrant, nous utilisons le logiciel procmail[7]. Ce programme a été originellement conçu et développé par Stephen R. van den Berg. À l'automne 1988, se rendant compte qu'il n'avait plus le temps de s'en occuper personnellement, Stephen a créé une liste de distribution[8] pour discuter des développements futurs et a désigné Philip Guenther pour la maintenance du logiciel. Par la même occasion, une licence axée sur la publication du code source a été adoptée. En bref, cette license permet à

[5] www.sun.com

[6] gcc.gnu.org

[7] www.procmail.org

[8] procmail-dev-request@procmail.org

quiconque de changer le code source et de distribuer le logiciel modifié, en autant que le code source modifié accompagne le programme et que les modifications soient décrites. À cause de cela, plusieurs versions différentes ont été créées, modifiées et fusionnées, de sorte que la documentation n'est plus toujours précise ni adéquate. Cela nous a causé passablement de problèmes.

Procmail est un "Mail Delivery Agent", c'est-à-dire un logiciel chargé d'effectuer la gestion des messages lorsqu'ils sont rendus à leur destination finale. Contrairement à qmail, procmail est axé sur la gestion finale du courrier électronique. Par conséquent, il offre de nombreuses possibilités pour traiter automatiquement les courriels à leur arrivée:

- envoyer des accusés de réception
- faire suivre les courriels vers une autre adresse
- trier les courriels à l'aide d'expressions régulières contenues dans l'un ou plusieurs des champs (`to`, `from`, `subject`, etc.)
- séparer le corps de l'en-tête du courriel
- exécuter d'autres programmes et leur fournir le courriel en entrée

Promail fonctionne avec des règles, appelées *recettes*, contenues dans un fichier de configuration. Une recette est constituée de l'initialisation, suivie d'aucune, une ou plusieurs conditions, et d'une action. Lorsqu'un courriel arrive au serveur, procmail applique les recettes une à une, selon leur ordre d'apparition dans le fichier de configuration. Dès que toutes les conditions d'une recette sont remplies, l'action de cette recette est appliquée et le traitement de procmail s'arrête. Dans le cas contraire, procmail passe à la recette suivante. Les recettes possèdent toutes la structure montrée à la figure 3.7.

L'initialisation d'une recette débute par `:0` et se termine par les options applicables à la recette. Ces options dictent l'utilisation du corps et de l'en-tête pour les conditions de la recette. Une autre option intéressante est de pouvoir poursuivre le traitement

```
initialisation
* condition 1
* condition 2
...
* condition n
action
```

Fig. 3.7: Structure d'une recette de procmail

avec une copie carbone du courriel même si les conditions de la recette sont remplies. Si le courriel satisfait les conditions, procmail exécute l'action et continue le traitement à partir de la prochaine recette avec une copie conforme du courriel. Sinon, une seule copie du courriel poursuit son chemin.

Les conditions sont très générales. Il est possible de rechercher des expressions régulières à l'intérieur des champs d'en-tête, ou parmi tout le corps du courriel. Il est aussi possible de vérifier la longueur d'un courriel. Cependant, la véritable force de procmail est sa capacité de lancer l'exécution d'un programme quelconque et de lui fournir le courriel (ou une partie) en entrée. Lorsque le programme se termine, sa valeur de sortie (0 pour un fin normale et un entier différent de 0 pour une erreur) est retournée à procmail pour savoir s'il doit continuer le traitement.

L'action peut être une redirection du courriel vers une ou plusieurs autres adresses électroniques, ou une sauvegarde du courriel dans un répertoire local. En plus, il est possible de fournir le courriel en entrée pour un autre logiciel. La dernière action, qui n'en est pas vraiment une, est l'utilisation d'un bloc d'exécution. Cela permet d'introduire d'autres recettes qu'il ne faut appliquer qu'aux courriels qui remplissent les conditions de la première recette, comme dans l'exemple de la figure 3.8.

Le première recette recherche les courriels dont le champ `to` contient le terme *duboisju*. Lorsque c'est le cas, le courriel devient l'entrée du programme "en_anglais". En supposant que la valeur de sortie de ce logiciel soit positive si l'entrée est en anglais et négative autrement, la deuxième recette sauvegarde le courriel dans le répertoire `Anglais` si le courriel est en anglais. Sinon, la troisième recette entre en vigueur et puisqu'elle ne contient pas de condition, enregistre immédiatement le courriel dans le

```
:0 (recette 1)
* ^To:duboisju
{
  :0 (recette 2)
  ?  en_anglais
  Anglais

  :0 (recette 3)
  Autres
}
```

Fig. 3.8: Exemple d'utilisation d'un bloc d'exécution dans une recette de procmail

répertoire `Autres`. La flexibilité de cette structure permet d'introduire un ensemble de règles complexes sur plusieurs niveaux. Et puisqu'il est possible de lancer l'exécution de n'importe quel programme en tant que condition, il n'y a aucune limite au traitement à effectuer sur le courrier électronique avec procmail.

3.3.3 Le traitement effectué

Le traitement effectué par le prototype est minimal, mais en même temps, représentatif de celui qui sera accompli par Merkure une fois complété. Dès qu'un courriel arrive sur notre serveur, il est pris en charge par procmail. Premièrement, procmail place une copie de tous les courriels provenant de Stylus dans le répertoire `BCE-4`. Ce répertoire agit comme sauvegarde. Ces courriels sont aussi enregistrés dans le répertoire portant le nom de l'adresse électonique qui leur est associée. Par exemple, tous les courriels destinés à investor.relations@bce.ca sont gardés dans le répertoire `To_Investor.relations`. Les courriels ne venant pas de chez Stylus sont sauvegardés dans le répertoire `Varia`. Dans cette catégorie, nous retrouvons principalement des courriels non sollicités envoyés à merkure@iro.umontreal.ca.

Puisque Merkure vise principalement les courriels des relations aux investisseurs en anglais, seuls ceux destinés à investor.relations@bce.ca sont traités. La figure 3.9 illustre le cheminement de ces courriels. Lorsque l'un d'entre eux est reconnu par procmail, il est

envoyé au programme silc[9]. Ce programme conçu au RALI détermine automatiquement la langue d'un document, le jeu de caractère employé ainsi que son niveau de confiance. S'il ne s'agit pas d'un message en anglais, le courriel est enregistré dans le répertoire "Autre_langue". Sinon, il poursuit le traitement en devenant l'entrée du tokeniseur, qui élimine les attachements, les balises HTML et les expressions reliées au courrier électronique, et qui sépare le texte restant en mots.

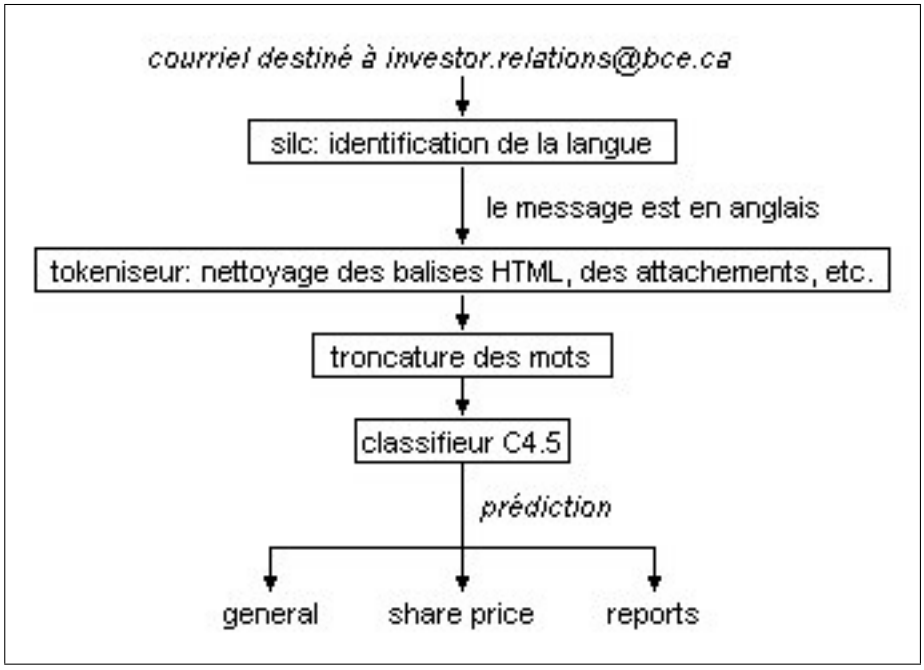

Fig. 3.9: Traitement effectué par notre prototype

Le tokeniseur fait donc un nettoyage du courriel en enlevant les expressions inutiles et en séparant le texte en une liste de mots. Cette liste devient ensuite l'entrée d'un programme de troncature. Selon l'algorithme de Porter [34], ce programme utilise des règles très simples pour extraire la racine des mots. En fait, il ne s'agit pas exactement de la racine mais plutôt d'une partie du mot, généralement commune aux autres mots de même racine. Puis, ces mots tronqués servent de caractéristiques pour un classifieur de type C4.5 (section 2.1.3) préalablemement entraîné. La classification est orientée sur le contenu des messages et comporte trois catégories. Les deux premières sont les deux classes les plus importantes du corpus BCE-3, soit `share price` et `reports`. L'autre

[9] www-rali.iro.umontreal.ca/SILC/

catégorie est tout le reste des courriels. Avec seulement un sous-ensemble restreint du corpus comme ensemble d'entraînement et aucune optimisation notable pour le classifieur, environ 80% des courriels sont bien étiquetés. Pour faciliter la résolution des problèmes, une trace assez précise de chaque opération est gardée pour chaque courriel entrant dans le système. Un exemple de la trace est disponible dans la figure 3.10.

```
Date:  Fri, 28 Sep 2001 21:57:23 -0500
From:  Julien Dubois <duboisju@iro.umontreal.ca>
Subject:  annual report request

LANGUAGE:
en

TOKENIZED MESSAGE:
from sat sep 29 2001 annual report request can you please send
me 2 annual reports

STEMMED WORDS:
from sat sep 29 2001 annual report request can you pleas send
me 2 annual report
```

Fig. 3.10: Exemple de la trace gardée par notre prototype pour un courriel

3.3.4 L'implantation chez Stylus

Après plusieurs jours d'observation, nous avons noté le bon fonctionnement du prototype sur notre serveur au RALI. Nous avons donc entrepris les démarches pour implanter ce même prototype chez Stylus. Suite à une réunion avec les gens de Stylus, un représentant du LUB et le responsable des relations aux investisseurs de BCE, l'administrateur de Stylus a modifié légèrement leur architecture pour accueillir notre prototype. Dans un premier temps, il a installé procmail sur leur serveur et il a ajouté une autre redirection à l'adresse investor.relations@bce.ca. Les courriels dirigés à cette adresse électronique sont maintenant envoyés vers trois destinations différentes: les serveurs de BCE, le serveur du RALI et un serveur local chez Stylus. Cette partie s'est accomplie sans heurt. Ensuite, il a procédé à l'installation des programmes et à

la configuration de procmail.

L'installation des programmes n'a causé aucun accroc majeur, si ce n'est que les versions différentes du compilateur gcc ne contenaient pas les mêmes librairies. Comme nous avions utilisé un "linkage" dynamique, et donc dépendant des librairies installées lors de l'utilisation, certaines erreurs sont survenues. Nous avons vite corrigé la situation en recompilant nos programmes avec un "linkage" statique, qui inclut les librairies dans l'exécutable et qui élimine les problèmes dus aux différences de version des librairies installées sur chaque ordinateur. La configuration de procmail nous a encore une fois causé des mots de tête. Nous avons dû retourner chez Stylus et collaborer avec l'administrateur pour résoudre les problèmes de configuration. Un après-midi a suffi pour finaliser le tout et après quelques jours en opération, l'administrateur nous a confirmé que le prototype installé chez Stylus fonctionnait parfaitement.

Bien que ce prototype n'effectue rien de très poussé, il nous a permis de régler plusieurs problèmes qui pouvaient survenir pendant la construction du système. En particulier, le manque de documentation au sujet de procmail a été une source d'ennuis majeurs pour effectuer un traitement automatique sur le courrier électronique. Aussi, il nous a révélé certains oublis lors du passage de notre serveur au RALI vers celui de Stylus. Même si le but du prototype n'était que monter la coquille du système, nous avons tout de même pris soin d'y inclure plusieurs étapes du traitement, dont l'identification de la langue, la séparation des courriels en une liste de mots, la troncature des mots et une classification partielle. Avec autant de composantes distinctes, il a été possible de mieux comprendre plusieurs mécanismes de gestion de procmail. En plus, nous avons pu mesurer le temps de réaction du système lors de l'arrivée d'un courriel. En moyenne, ce temps est inférieur à une seconde par courriel.

Chapitre 4

Les corpus

Au chapitre 2, nous avons vu plusieurs caractéristiques propres au courrier électronique. Cependant, il n'est pas garanti qu'elles soient toutes importantes dans le cas de Merkure. Nous décrivons donc ici les quatre corpus de courriels de BCE, en plus d'analyser en profondeur BCE-3, le corpus utilisé pour la conception du module de classification.

4.1 La description des corpus

4.1.1 Les corpus de BCE

Depuis le commencement de Merkure, nous avons obtenu trois corpus des messages et des suivis de BCE, et un quatrième est en constante construction. Nous les avons respectivement nommés BCE-1, BCE-2, BCE-3 et BCE-4, selon le rang de leur acquisition.

4.1.1.1 BCE-1

Nous avons reçu le premier corpus à la fin du mois de septembre 2000. En format électronique, il est composé de 141 messages *bce-comments*. Ceux-ci ont été envoyés à l'aide du formulaire de commentaire sur le site web de BCE entre le 26 avril 2000 et le 25 septembre 2000. Ces messages sont de nature générale et portent sur tous les aspects de l'entreprise, incluant le site web, les contacts, le domaine des relations aux investisseurs

et les plaintes sur les services. Ce corpus a servi à observer les caractéristiques des courriels de BCE pour faire une analyse préliminaire de leurs dossiers.

4.1.1.2 BCE-2

En octobre 2000, nous avons obtenu un deuxième corpus, mais en format papier. Il est constitué de 865 paires message-suivi qui ont transité par les adresses bcecomms@bce.ca et bcecomms@bell.ca. Alors que nous nous sommes basés sur BCE-1 pour recueillir des statistiques à cause des avantages de son format électronique, BCE-2 nous a été tout aussi utile pour deux raisons. Premièrement, le plus grand nombre de messages nous a permis de vérifier les résultats tirés de BCE-1. Et deuxièmement, la combinaison des messages avec leurs suivis nous a permis de compléter l'analyse débutée avec BCE-1 et d'établir la proposition de projet de réponse automatique au courrier électronique qu'est devenu Merkure.

4.1.1.3 BCE-3

Au mois de décembre 2000, nous avons reçu un troisième corpus, le deuxième en format électronique. Il regroupe 1568 paires message-suivi reliées à investor.relations@bce.ca. Les dates d'envoi des courriels varient entre le 29 juin 1999 et le 2 novembre 2000. Comme nous concentrons nos efforts sur le département des relations aux investisseurs, il s'agit du premier corpus représentatif des messages et des suivis que nous rencontrerons. Nous analysons en détail ce corpus dans la deuxième partie de ce chapitre.

4.1.1.4 BCE-4

Le dernier corpus, toujours en format électronique, ne cesse de grossir. Comme nous l'avons vu dans le dernier chapitre, Stylus nous envoie depuis le début du mois d'avril 2001 une copie de tous les courriels transitant par certaines adresses électroniques de BCE. Il s'agit bien sûr des courriels dirigés vers les adresses investor.relations@bce.ca, relations.investisseurs@bce.ca et bcecomms@bce.ca, et des courriels en provenance de bcecomms@bell.ca. La table 4.1 résume les statistiques des courriels reçus quotidiennement de BCE. Nous n'avons pas utilisé les courriels du mois de septembre pour les

statistiques de investor.relations@bce.ca parce que nous avons égaré trop de courriels en faisant des tests avec procmail. La première ligne de cette adresse donne les statistiques avant le mois de septembre, la deuxième fournit celles après septembre et la troisième est la combinaison des deux périodes.

adresse électronique	**début**	**fin**	**courriels**	**jours**	**moy.**
investor.relations@bce.ca	2001-05-01	2001-08-31	474	123	3.85
	2001-10-01	2002-01-31	682	123	5.54
	2001-05-01	2002-01-31	1156	246	4.70
relations.investisseurs@bce.ca	2001-08-01	2002-01-31	79	184	0.43
bcecomms@bce.ca	2001-05-01	2002-01-31	3392	276	12.29
bcecomms@bell.ca	2001-05-01	2001-01-31	991	276	3.69
total			6774		5.28

Tab. 4.1: Statisitiques sur les courriels reçus quotidiennement de BCE-4

Nous avons déterminé la période en prenant tous les mois complets disponibles lors du calcul. Fruit du hasard, cette période équivaut aux mois qui ne correspondent pas au temps des impôts. Comme les mois de février, mars et avril sont les plus achalandés, les moyennes sont biaisées vers le bas. Par exemple, au cours des 27 jours d'avril 2001 pendant lesquels nous avons reçu des courriels, 323 sont entrés pour investor.relations@bce.ca pour une moyenne de 11,96 courriels par jour. Cela dépasse le double des courriels reçus quotidiennement pendant les mois de mai à décembre et janvier. Aussi, les moyennes sont calculées en comptant tous les jours, ouvrables ou non. Et nous avons estimé que BCE reçoit environ deux fois moins de courriels les jours de congé que les jours ouvrables.

4.1.2 Les autres corpus

Nous utilisons aussi deux autres corpus à titre de comparaison pour mieux situer ceux de BCE-3. Pour l'analyse de BCE-3, nous tentons de mesurer numériquement certaines caractéristiques et nous nous servons de ces corpus auxiliaires comme points de repère.

4.1.2.1 Assisted Living

Nous avons bâti ce corpus à partir d'un forum de discussion[1] sur la perte d'autonomie des personnes âgées. Évidemment, il y a très peu de liens entre ce sujet et les relations aux investisseurs. Par contre, comme beaucoup de babillards en ligne, ce corpus a la particularité d'être axé sur les expériences personnelles des gens. Il s'agit donc d'un bon exemple de messages personnels portant les caractéristiques des langages électroniques que nous avons vues au chapitre 2. Ce type de message est loin des documents froids et impersonnels publiés officiellement dans les revues et les journaux. Aussi, ils n'ont pas nécessairement été sujets à une révision et une correction adéquates de la part de leurs auteurs. Un ensemble de 342 messages publiés sur le forum entre le 17 mai 2001 et le 29 septembre 2001 composent ce corpus.

4.1.2.2 Reuters

Reuters[2] est le plus gros fournisseur de nouvelles et d'informations financières aux médias, institutions financières, entreprises et individus. Comme service à la communauté scientifique, cette entreprise met à la disposition des chercheurs une grande quantité de documents pour des expériences sur le traitement des langues naturelles, la recherche d'information et les algorithmes d'apprentissage. Par conséquent, la distribution 1.0 de Reuters-21578[3] est un corpus rencontré très souvent en classification de textes.

Ce corpus regroupe 21578 articles publiés par Reuters durant l'année 1987. Il est spécialement conçu pour la classification de texte. Tous les documents sont structurés en NewsML, un langage à balise basé sur le XML adapté pour des documents informatifs. Ce format permet de bien séparer le contenu du document des informations relatives au traitement, comme la date et le numéro d'identification unique. De plus, la majorité des documents sont étiquetés de l'une ou plusieurs des 135 catégories. De ce nombre, une douzaine sont prédominantes tandis que beaucoup ne contiennent que très peu

[1] www.assistedlivingforum.com

[2] www.reuters.com

[3] about.reuters.com/researchandstandards/corpus/

d'exemples, voire un seul.

Il existe officieusement quatre façons de diviser le corpus en ensembles d'entraînement et de validation. Les caractéristiques de ces ensembles sont rapportées dans la table 4.2. La première manière, nommée ModHayes, est sensiblement la même que celle utilisée pour le premier corpus de Reuters, Reuters-22173 [18], et n'a jamais été très utilisée. Les trois autres séparations se basent sur la date du 8 avril 1987 pour diviser le corpus. Tous les documents publiés avant cette date font partie de l'ensemble d'entraînement, et les autres sont dans l'ensemble de validation [24]. Ces trois séparations diffèrent sur l'utilisation des documents selon leur étiquetage:

ModLewis Presque tous les documents sont utilisés, incluant ceux qui ne sont pas classés.

ModApte Seuls les documents catégorisés sont utilisés.

ModTrivial Seuls les documents classés dans une seule catégorie sont utilisés.

Reuters-21578 est le meilleur corpus à utiliser pour analyser les performances d'un classifieur. Étant le corpus le plus mentionné, il est facile de dénicher des travaux qui l'utilisent et de compararer ses résultats avec ceux déjà publiés. Cela donne de bonnes indications des performances relatives d'un classifieur par rapport à un corpus standard. Aussi, les différentes séparations offrent la possibilité de travailler avec certaines caractéristiques qui peuvent nuire aux classifieurs. La séparation ModLewis permet aux documents d'appartenir à plusieurs classes tandis que c'est interdit avec la séparation ModTrivial. Il est aussi possible de ne travailler qu'avec les plus grosses classes ou au contraire, se compliquer la tâche en incluant toutes les classes, y compris celles ne contenant qu'un seul exemple. Dans le cadre de Merkure, nous n'avons utilisé que la

séparation	**entraînement**	**validation**	**inutilisé**
ModHayes	20856	722	0
ModLewis	13625	6188	1765
ModApte	9603	3299	8676
ModTrivial	6555	2583	12440

Tab. 4.2: Répartition des documents selon la séparation de Reuters-21578

séparation ModTrivial, car c'est celle qui se rapproche le plus du type de classification que nous envisageons, c'est-à-dire qu'un message ne peut appartenir qu'à une catégorie.

4.2 L'analyse de BCE-3

4.2.1 La description générale

Comme nous l'avons déjà mentionné, le corpus BCE-3 est le premier à contenir le genre de messages avec lequel nous travaillons pour Merkure. Il donne un bon aperçu des échanges entre les clients de BCE et les préposés du département des relations aux investisseurs. BCE-3 est composé de 1629 paires messages-suivi, dont les dates d'envoi se situent entre le 29 juin 1999 et le 2 novembre 2000. Toutefois, il semble provenir tout droit de plusieurs répertoires de courriels envoyés par les préposés de BCE. Un bon indice qui nous porte à croire cela est les 22 catégories[4] de BCE-3 détaillées dans la table 4.3.

À première vue, les classes semblent bien représenter les différentes sortes de courriels. Par contre, il est évident que certaines d'entres elles sont temporelles, c'est-à-dire qu'elles contiennent des courriels concernant un événement précis dans le temps. De bons exemples de classes temporelles sont les acquisitions de CTV et de Teleglobe, ainsi que la conférence du chef des opérations, qui ont donné lieu à trois catégories supplémentaires. Aussi, la classe `instant reply` est tout ce qu'il y a de plus inhabituel. Alors que le reste des courriels sont classés selon leur contenu, cette catégorie regroupe ceux qui semblent avoir été répondus dès leur arrivée, ce qui n'a aucun rapport avec le contenu.

Un autre indice pour renforcer l'opinion que le corpus ressemble à un assemblage de plusieurs répertoires est la distribution des dates d'envoi des courriels. La très grande majorité de ces courriels ont tous été envoyés par investor.relations@bce.ca à l'intention du client ou d'une autre personne ressource. Et en regardant de plus près les dates d'émission des courriels, nous nous apercevons que BCE-3 ne regroupe pas un ensemble

[4] Cette énumération contient les noms des catégories tels qu'attribués par les préposés de BCE. Nous utilisons ces noms en anglais pour éviter toute forme de confusion.

catégorie	**description**
admin	abréviation d'"administration"; tout ce qui se rapporte à l'administration commme les CV et les demandes de stage
agm	abréviation d'"annual general meeting"; courriels concernant l'assemblée générale annuelle
annual reports	demandes de rapports annuels et trimestriels
appreciation	commentaires positifs
ceo conference 2000	questions sur la conférence du chef des opérations de BCE au début de l'année 2000
complaints	plaintes
CTV	courriels concernant l'achat de CTV par BCE
dividend	tout ce qui concerne les dividendes
drp	abréviation de "dividend reinvestment plan"; demandes d'information sur le régime de réinvestissement des dividendes
esp	abréviation de "employee savings plan"; demandes d'information sur le régime d'épargne des employés de BCE
financial info	demandes d'informations financières absentes du rapport annuel
info on account	questions personnelles d'investisseurs qui ont perdu leur certificat d'action, qui veulent faire un changement d'adresse, etc.
iob	abréviation de "info on BCE's subsiduaries"; courriels concernant les compagnies sous la tutelle de BCE
instant reply	courriels qui ont été répondus dès leur arrivée
other	tout ce qui ne peut pas être placé ailleurs
other documentation	demandes de documents divers, dont la brochure corporative pour les investisseurs potentiels
pnd	abréviation de "prefered notes debentures"; tout ce qui concerne les actions privilégiées de BCE
sales purchases	questions sur l'achat et la vente d'actions
share price	questions sur la valeur des actions de BCE
stock split	tout ce qui concerne le fractionnement des actions de BCE
tax	demandes d'information pour le calcul des taxes
Teleglobe	courriels concernant l'achat de Teleglobe par BCE

Tab. 4.3: Description des catégories originales de BCE-3

de courriels que le département des relations aux investisseurs a reçu pendant la même période de temps pour toutes les classes. La table 4.4 affiche les dates d'envoi du premier et du dernier courriel ainsi que le nombre de courriels pour chaque catégorie.

catégorie	courriels	premier	dernier
admin	25	1999-06-29	2000-06-16
agm	22	2000-02-16	2000-04-18
annual reports	264	2000-01-04	2000-10-23
appreciation	73	2000-02-15	2000-09-13
ceo conference 2000	6	2000-05-17	2000-05-17
complaints	3	2000-05-21	2000-05-25
CTV	6	2000-02-29	2000-08-10
dividend	69	2000-01-10	2000-10-25
drp	75	2000-01-10	2000-10-31
esp	8	2000-02-11	2000-09-11
financial info	103	2000-01-05	2000-11-01
info on account	155	2000-01-17	2000-11-01
iob	124	2000-01-05	2000-10-31
instant reply	20	2000-01-10	2000-06-05
other	194	2000-01-06	2000-11-01
other documentation	32	2000-01-16	2000-10-20
pnd	45	2000-02-15	2000-11-02
sales purchases	15	2000-02-09	2000-10-26
share price	258	2000-01-06	2000-10-20
stock split	56	2000-01-05	2000-09-25
tax	5	2000-05-24	2000-11-01
Teleglobe	71	2000-02-21	2000-10-30
total	1629	1999-06-29	2000-11-02

Tab. 4.4: Dates des premiers et derniers courriels envoyés pour BCE-3

La distribution des dates est très inégale. La majorité des classes contiennent un premier courriel envoyé au mois de janvier et quelques-unes retardent au mois de février, mais certaines doivent attendre jusqu'aux mois de mai, juin et même juillet. Du côté du dernier courriel, la situation se répète et les dates s'étalent du milieu d'avril au début de novembre. Il est possible que différents types de questions reviennent plus souvent à certaines périodes de l'année. Par exemple, les courriels de la classe `agm` sont concentrés de février à avril, et l'assemblée dont il est question se situait le 26 avril 2000. Par contre, il serait surprenant que les plaintes se limitent au mois de mai ou qu'aucune demande concernant les dividendes n'ait été faite avant la fin de février. À partir de ces anomalies, il est n'est pas faux de penser que BCE-3 ne contient pas tous les courriels envoyés au cours d'un certaine période de temps. Il n'y a rien de dramatique avec cette situation, mais nous devons tenir compte du fait que la distribution des courriels dans

chaque classe n'est pas exacte.

4.2.2 Le nettoyage

Comme nous devions séparer tous les courriels en paires message-suivi, nous en avons profité pour entièrement nettoyer BCE-3. La table 4.5 résume les opérations effectuées. La première colonne indique le nombre initial de courriels dans chaque catégorie. Parmi ceux-ci, quelques-uns résultent d'un échange multiple[5]. Nous utilisons ce terme pour désigner une séquence prolongée de courriels entre un client et un préposé des relations aux investisseurs, où chaque suivi devient un message dans la suite de la conversation. Nous avons séparé ces courriels en autant de paires messages-suivi que nécessaire. La deuxième colonne indique le nombre de messages additionnels, excluant le premier de l'échange qui est compté dans le nombre de messages initial. La troisième colonne représente donc le nombre total de messages. À cela, nous avons retranché les courriels en français. Comme nous l'avons mentionné précédemment, nous concentrons nos efforts sur les messages en anglais, qui composent 92.3% du corpus. Puis, les courriels ne formant pas une paire message-suivi valide ont été éliminés. Par exemple, nous n'avons pas les suivis des quelques courriels ne provenant pas de investor.relations@bce.ca. Certains courriels ne contiennent qu'une réponse sans avoir inclu le message initial. Il y a aussi quelques échanges entre deux préposés des relations aux investisseurs. La dernière colonne affiche le nombre de messages valides à la fin du nettoyage.

Le nettoyage nous a donné une première vue d'ensemble du corpus. En le parcourant, nous nous sommes aperçu que la classification fournie par les préposés de BCE manque de rigueur. Par exemple, des demandes de rapport annuel se retrouvent dans la classe `other documentation`, et des courriels concernant l'achat et la vente d'actions font partie de `info on account`. Il est vrai que certaines erreurs peuvent se produire, surtout lorsque deux catégories se recoupent et que le message touche les deux. Par contre, lorsque des demandes de prix sont placées dans `sales purchases` et que la classe `annual reports` contient des courriels concernant un changement d'adresse, il y a lieu de se poser des questions. Ces erreurs de classement sont assez fréquentes pour

[5] Le terme anglais équivalent est "thread".

catégorie	initial	addit.	total	français	invalide	valide
admin	25	0	25	1	8	16
agm	18	4	22	1	5	16
annual report	264	0	264	20	43	201
appreciation	42	31	73	0	1	72
ceo conference	6	0	6	0	0	6
complaints	3	0	3	0	2	1
CTV	6	0	6	1	0	5
dividend	69	0	69	8	6	55
drp	72	3	75	4	5	66
esp	7	1	8	2	1	5
financial info	102	1	103	6	24	73
info on account	153	2	155	13	41	101
iob	122	2	124	17	30	77
instant reply	20	0	20	3	14	3
other	192	2	194	7	104	83
other documentation	31	1	32	2	6	24
pnd	45	0	45	2	21	22
sales purchases	15	0	15	0	1	14
share price	244	14	258	12	11	235
stock split	56	0	56	4	2	50
tax	5	0	5	1	0	4
Teleglobe	71	0	71	21	8	42
total	1568	61	1629	125	333	1171
% des messages	96.3%	3.7%	100.0%	7.7%	20.4%	71.9%

Tab. 4.5: Validité des paires message-suivi de BCE-3

justifier une vérification en profondeur de la catégorisation. Nous avons donc établi une nouvelle classification, comme nous le verrons ultérieurement.

4.2.3 Les caractéristiques des messages

Nous présentons ici plusieurs caractéristiques des messages de BCE-3. Toutes les données n'apportent pas la même influence sur la progression de Merkure. Quelques-unes ont des effets concrets sur l'élaboration des modules de Merkure. D'autres, comme la longueur des messages, n'ont pas une influence directe sur la conception mais nous donnent un aperçu des difficultés qui nous attendent.

4.2.3.1 Les échanges multiples

Environ 9.9% des messages font partie d'un échange multiple, incluant le premier courriel de chaque échange. Si nous considérons uniquement les messages additionnels, qui contiennent le texte d'au moins un autre courriel, le pourcentage chute à 5.2%. C'est ce dernier qui est plus approprié, car le premier courriel d'un échange est comme tout autre courriel normal et ne contient pas de texte surperflu. Plus souvent qu'autrement, les messages additionnels ont plus ou moins rapport avec le premier. Il s'agit surtout de clients qui ont déjà envoyé un courriel aux relations aux investisseurs et qui ont gardé la réponse qu'ils ont reçue. Puis, lorsqu'ils ont une autre question, ils reprennent simplement ce dernier courriel et y répondent sans changer le texte, et en gardant l'ancienne conversation indentée dans leur courriel. En outre, ils ne prennent même pas le temps de changer le sujet du nouveau courriel.

Un pourcentage aussi faible de courriels additionnels est loin de représenter un sous-ensemble distinct significatif. De plus, comme nous l'avons précisé dans le chapitre 2, séparer la nouvelle partie de l'ancienne n'est pas une chose facile. La combinaison de ces deux facteurs fait en sorte qu'il ne vaut pas vraiment la peine de s'inquiéter avec ces courriels plus difficiles à traiter.

4.2.3.2 La signature

Nous considérons que la signature d'un courriel est tout ce qui suit le nom de l'émetteur à la fin du message. L'exception à cette définition est l'utilisation d'un post-scriptum, ce qui est très rare dans un courriel. Habituellement, les gens se contentent d'écrire leur prénom et leur nom de famille. Toutefois, beaucoup de clients ajoutent d'autres informations pouvant nuire au traitement. La table 4.6 illustre les différentes sortes d'informations rencontrées dans les signatures de BCE. Certains clients ne sentent pas le besoin de signer leur courriel. Ils n'ajoutent donc rien après leur message, sauf parfois un remerciement. Cependant, la majorité des clients signent leur courriel avec leur prénom et leur nom de famille ou leurs intiales. De ce second groupe, plusieurs écrivent d'autres informations personnelles comme leur numéro de téléphone, leur adresse civique ou électronique, la compagnie pour laquelle ils travaillent et leur poste, l'adresse d'un

site web personnel, etc. Aussi, quelques-uns utilisent un service de courrier électronique gratuit qui ajoute de la publicité au bas du message. Finalement, d'autres finissent avec une notice légale à l'égard des informations contenues dans le courriel.

type	**messages**	**%**
nom	435	37.1
nom + informations	378	32.3
vide	251	21.5
nom + publicité	45	3.9
nom + information + notice	24	2.0
pub	22	1.9
nom + informations + publicité	12	1.0
nom + notice	4	0.3
total	1171	100.0

Tab. 4.6: Types de signature des messages de BCE-3

Quelques rares clients ne signent pas leur nom mais utilise une expression pour se "décrire". Par exemple, un client demandant des informations au sujet de ses taxes a signé: *un actionnaire confus*. Un autre a préféré terminer son message par: *un fier actionnaire de BCE*. Nous considérons que cela équivaut à aucune signature, même si c'est plus ou moins le cas.

Beaucoup d'informations inutiles se retrouvent au niveau de la signature. Bien que les informations personnelles soient très fréquentes, elles ne sont pas trop nuisibles au traitement car elles contiennent surtout des numéraux, que nous pouvons facilement éliminer. La publicité aussi n'est guère une source de préoccupation majeure car elle se limite habituellement à deux ou trois courtes lignes. Par contre, les notices s'étirent généralement sur plusieurs lignes de texte. Dans BCE-3, l'une d'elle s'étend même sur quinze lignes, ce qui représente beaucoup de texte supplémentaire qui n'a rien à voir avec le message.

4.2.3.3 La longueur des messages

La longueur des messages n'a pas une grande influence sur la conception des modules. Par contre, elle procure une bonne idée de la difficulté à extraire les informations importantes d'un message. À moins de devoir faire une mise en situation détaillée pour

un cas spécial, les informations supplémentaires supportant la question n'apportent habituellement que du bruit, c'est-à-dire des informations nuisibles au traitement. Par exemple, le message de la figure 4.1 est court et direct. Le client veut une information et pose la question la plus simple pour l'obtenir. Il n'y a rien de superflu. En comparaison, le message de la figure 4.2 est beaucoup plus long.

```
Pouvez-vous me donner les informations concernant le fractionne-
ment des actions au cours des 30 dernières années?
```

Fig. 4.1: Exemple de message court et précis

```
J'ai cherché longtemps sur votre site pour les informations con-
cernant le fractionnement des actions au cours des 30 dernières
années.  Votre site ne recule que de quelques années.  J'ai vu
celui du 12 mai 1977, mais je n'en vois pas d'autres.  Je suis
actionnaire depuis fort longtemps et je me demande s'il y a eu
d'autres fractionnements avant 1977.  Mon numéro de certificat
est xxxxxx.
```

Fig. 4.2: Exemple de message contenant beaucoup d'informations inutiles

Le deuxième message contient beaucoup d'informations inutiles par rapport à la question. Premièrement, les informations personnelles du client (son numéro du certificat et son investissement de longue date) ne servent à rien dans ce cas-ci. Deuxièmement, le fait que le client ait cherché longtemps sur le site web pour l'information n'a rien à voir avec la question. Si ce genre de remarque revient fréquemment, un préposé peut noter ce commentaire et en faire part aux personnes concernées. Par contre, un système de traitement automatique n'a aucune chance de repérer cette redondance, à moins d'avoir été conçu pour ça. Troisièmement, que le client connaisse ou non le fractionnement de 1977 ne change presque rien, si ce n'est qu'il en parle et qu'il rallonge son message sans apporter quelque chose de nouveau. Le préposé ne voit pas une grande différence entre les deux messages, à part que le second est plus humain et moins aride. Un traitement automatique ne peut pas repérer aussi bien les informations vraiment importantes et

peut se tromper en considérant le message comme un problème personnel de ce client ou d'un commentaire par rapport au site web, selon les caractéristiques utilisées par le système. Toutefois, comme le traitement des langues naturelles repose principalement sur les mots des documents, un message court implique plus d'emphase sur chacun des mots utilisés, ce qui peut être aussi néfaste. Si le vocabulaire employé est trop différent des autres messages posant la même question, il y a moins de chance qu'il leur soit apparenté et que le contenu soit bien identifié. C'est justement l'une des tâches qui incombent au module de question-réponse, qui fait l'objet d'un travail de thèse de doctorat dans le projet Merkure.

Comme le montre la figure 4.3, la majorité des messages ne dépassent pas 90 mots. Une telle longueur suggère que ceux-ci vont droit au but, sans ajouter de larges quantités d'informations superflues. Il faut noter que cette courbe de distribution est légèrement biaisée vers la droite car certains messages font partie d'un échange multiple. Ceux-ci sont plus longs qu'ils ne devraient parce qu'ils contiennent le texte des courriels échangés précédemment en plus du message actuel. Toutefois, les signatures influencent encore plus ce biais, surtout avec les notices légales de plusieurs lignes.

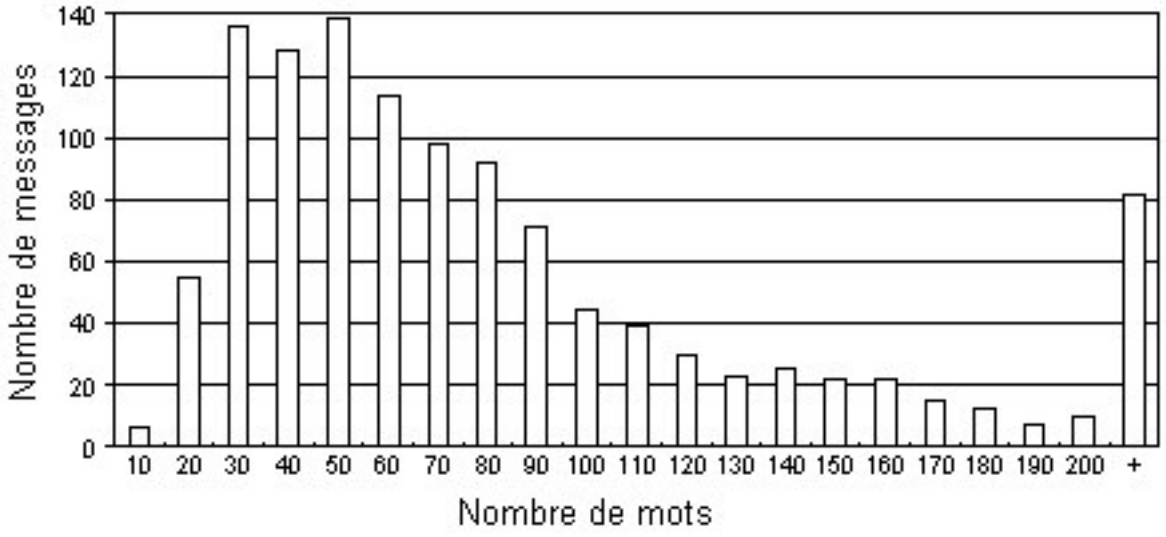

Fig. 4.3: Longueur des messages

4.2.3.4 La généricité du contenu

La généricité du contenu indique jusqu'à quel point ce dernier fait référence à des entités spécifiques. Habituellement, les messages les plus faciles à traiter sont les plus génériques. Les messages spécifiques requièrent plus de connaissances et surtout, des techniques de traitement plus approfondies. En outre, l'identification des passages les plus informatifs revêt une importance capitale. Si le système ne repère pas l'expression qui transforme une question d'ordre général en un cas particulier, il est peu probable que la réponse soit complète. Bien qu'il ne soit pas possible de calculer avec précision la généricité du contenu, une combinaison de quelques aspects en représente une bonne approximation:

pronoms et déterminants à la $1^{ère}$ personne Le nombre de pronoms et de déterminants à la $1^{ère}$ personne (je, nous, mon, nos, etc.) par rapport au total des pronoms et des déterminants donne une bonne indication de la nature personnelle du message. Un texte personnel est habituellement plus spécifique qu'un document purement objectif parce que le contenu est orienté et adapté en fonction de la personne qui l'a écrit et de l'auditoire cible. Cette caractéristique se retrouve à la première colonne de la table 4.7.

noms propres La proportion des noms propres sur le total des noms révèle l'importance des entités nommées. Plus la proportion est grande, plus le texte tourne autour de ces entités et par conséquent, plus il est spécifique. Il s'agit de la deuxième colonne de la table 4.7.

numéraux Le ratio de numéraux par rapport au nombre total de mots donne une bonne indication de la spécificité du message. Les valeurs numériques représentent généralement des entités spécifiques comme la date, les numéros de compte ou de certificat, les numéros de téléphone, les adresses, etc. Un ratio élevé de ces valeurs est relié à un texte spécifique. Cette caractéristique est présente à la troisième colonne de la table 4.7.

Il faut se méfier des deux derniers aspects. Alors qu'il est rare de rencontrer des pronoms et des articles autour d'un texte, ce n'est pas le cas des noms propres et des numéraux. Le meilleur exemple est l'utilisation de longues signatures comprenant des noms d'entreprise, des numéros de téléphone, des adresses civiques, etc. Ces informations additionnelles gonflent les pourcentages à tort. Par conséquent, un message générique peut paraître spécifique à cause d'une simple signature. La table 4.7 compare les propriétés de BCE-3 à BCE-1, Reuters et Assisted Living.

corpus	**1^{ere} personne**	**noms propres**	**numéraux**
BCE-3	25.1%	27.0%	5.2%
BCE-1	31.3%	20.3%	3.3%
Assisted Living	25.2%	7.5%	1.2%
Reuters	2.1%	17.0%	16.2%

Tab. 4.7: Comparaison de la généricité de BCE-3 à d'autres corpus

Comme nous l'avons déjà mentionné, Reuters est un corpus d'articles financiers. Les pronoms personnels sont donc peu présents, tandis que les noms propres et les numéraux sont fréquents. À l'opposé, le corpus Assisted Living est composé d'expériences personnelles, ce qui augmente l'utilisation des pronoms au détriment des noms propres et des numéraux. En comparaison, les corpus de BCE semblent de nature aussi personnelle qu'Assisted Living, avec une importance évidente pour les entités nommées et un ratio de numéraux qui se démarque à peine. Les différences entre les deux corpus de BCE sont expliquables par la spécialisation des courriels envoyés au département des relations aux investisseurs (BCE-3) et aux questions d'ordre général de BCE-1. Les messages de BCE-3 sont souvent reliés à la valeur des actions pour une certaine date, ce qui explique en partie le pourcentage de numéraux plus élevé et la moins grande proportion des pronoms et déterminants à la $1^{ère}$ personne. Un message typique est montré à la figure 4.4.

```
Quelle était la valeur des actions de BCE le 19 janvier 1985?
```

Fig. 4.4: Exemple de message typique sur la valeur des actions de BCE

Le deuxième apport important de numéraux se trouve dans les informations personnelles que les gens fournissent comme les numéros de certificat, les adresses, les numéros de téléphone, etc. D'autre part, le ratio élevé des noms propres est bien sûr relié à la signature des courriels par leur émetteur, mais surtout par l'abondance de références à BCE et Bell, ainsi que les entreprises sous leur tutelle comme Bell Mobilité, Bell Globemedia, Nortel, Teleglobe, etc. En résumé, les messages de BCE-3 sont très spécifiques car ils font référence à des dates et des événements bien précis. De plus, ils sont de nature personnelle parce que les clients demandent des informations ajustées en fonction de leur situation.

4.2.3.5 Le sujet

Le sujet[6] d'un courriel sert à exposer l'objet du message, habituellement à l'aide de quelques mots clés. Dans le cas de BCE-3, la table 4.8 montre les proportions des différents types de sujets:

vide Un sujet vide correspond à une absence de mots, ou à une expression générée par le logiciel de courrier électronique qui signifie l'absence d'un sujet, telle que *no subject*.

bce-comments Ces courriels proviennent du formulaire de commentaire sur le site web de BCE. Ils ont tous un sujet contenant le terme *bce-comments*, la date et l'heure, selon le format suivant: *bce-comments,14/10/2000,19:48:32*.

bce-documents Ces courriels viennent du formulaire de demande de documents sur le site de BCE. Comme ce formulaire a fait son apparition après que nous ayons reçu BCE-3, ce type de sujet n'apparaît pas.

non significatif Ce sujet définit le but du courriel trop vaguement ou seulement en partie.

[6] Pour éviter toute confusion, nous employons le terme *sujet* uniquement pour désigner le champ `Subject` d'un courriel. Nous utilisons les termes *but* et *objet* pour identifier le contenu du courriel.

significatif Ce sujet définit si bien l'objet du courriel qu'il est possible d'avoir une très bonne idée du contenu en ne lisant que le sujet.

type	**messages**	**%**
vide	63	5.4
bce-comments	98	8.4
non significatif	348	29.7
significatif	662	56.5
total	1171	100.0

Tab. 4.8: Types de sujet des messages de BCE-3

À des fins de traitement automatique, les sujets vides ou provenant du site web de BCE ne sont guère informatifs, mais ils ne sont pas nuisibles non plus. Ils sont tout simplement ignorés. Les sujets non significatifs le sont souvent en raison d'un manque de précision. Par exemple, une requête pour obtenir une copie du dernier rapport annuel peut avoir un sujet significatif ressemblant à *rapport annuel*, ou *requête de document financier*, mais *info financière* et *requête de document* ne sont pas assez spécifiques. Aussi, certains clients choisissent un sujet spécifique, mais peu pratique pour un traitement automatique. Un courriel dont le but est d'obtenir un rapport annuel dans le cadre d'un projet scolaire et qui a comme sujet *projet scolaire* est un bon exemple. Il faut aussi noter que nous considérons significatif un sujet qui ne représente qu'une partie du message. Cela se produit lorsque le client fait plusieurs demandes à l'intérieur d'un seul courriel.

Puisque plus de la moitié des courriels ont un sujet significatif, il semble normal de jumeler le sujet au corps du courriel pour le traitement. Lorsque les mots contenus dans les sujets non significatifs sont différents des autres, il est envisageable d'attribuer un poids plus important aux mots contenus dans le sujet des courriels. Puisque les termes formant les sujets significatifs sont souvent des mots clés, il est normal de leur accorder plus d'emphase. Une telle pondération peut améliorer le traitement des courriels ayant un sujet significatif, mais peut aussi nuire à ceux qui ont un sujet non significatif. C'est le cas lorsque les mots employés pour les sujets non significatifs sont similaires à ceux utilisés pour les sujets significatifs. La table 4.9 indique justement que nous faisons face à cette situation. Elle montre les quinze mots les plus fréquents contenus dans les

sujets significatifs, non significatifs et tous les sujets sans distinction. Les pourcentages correspondent au nombre de courriels qui contiennent ce mot par rapport au nombre de courriels appartenant à ce type de sujet. Par exemple, 19.2% des courriels ayant un sujet significatif contiennent le terme *bce*.

significatifs			**non significatifs**			**tous les sujets**		
mot	**fréq.**	**%**	**mot**	**fréq.**	**%**	**mot**	**fréq.**	**%**
bce	127	19.2	bce	75	21.6	bce	202	17.2
price	114	17.2	share	53	15.2	share	153	13.1
report	113	17.1	inform	37	10.6	price	134	11.4
annual	104	15.7	stock	33	9.5	stock	125	10.7
share	100	15.1	nortel	24	6.9	report	125	10.7
stock	92	13.9	request	21	6.0	annual	109	9.3
dividend	66	10.0	bell	21	6.0	of	84	7.2
of	65	9.8	price	20	5.7	dividend	73	6.2
for	38	5.7	question	19	5.4	nortel	57	4.9
split	34	5.1	of	19	5.4	inform	51	4.4
nortel	33	5.0	info	18	5.2	for	51	4.4
investor	30	4.5	sharehold	16	4.6	request	46	3.9
histor	29	4.4	investor	13	3.7	split	44	3.8
and	29	4.4	for	13	3.7	investor	43	3.7
list	27	4.1	report	12	3.4	bell	42	3.6

Tab. 4.9: Mots les plus fréquents des sujets des messages de BCE-3

Plusieurs mots se retrouvent autant dans les sujets significatifs que non significatifs, ce qui n'est pas idéal pour une pondération majorée en faveur des mots contenus dans le sujet. En plus, un courriel abordant plusieurs objets différents a plus de chance d'être mal traité si un seul de ces aspects est présent dans le sujet, comme c'est souvent le cas.

Dans un autre ordre d'idée, revenons au manque de liens entre les courriels faisant partie d'un échange multiple. Dans ce sous-ensemble, 47.4% des messages ont un sujet significatif si nous incluons le premier courriel de chaque échange. En comparaison, le pourcentage descend à 21.3% si le premier courriel de chaque échange n'est pas inclus dans l'ensemble. Ces chiffres confirment ce que nous avions affirmé précédemment par rapport au manque de rigueur des clients lorsqu'ils utilisent une ancienne réponse pour envoyer un nouveau message.

4.2.3.6 Les caractéristiques mineures

Les dernières caractéristiques présentées ici sont beaucoup moins préoccupantes en raison de leur absence presque totale du corpus BCE-3. De plus, même si elles prennent un ampleur modérée, nous pouvons en venir à bout assez facilement. Ces caractéristiques sont au nombre de quatre:

les virus Bien entendu, aucun courriel de BCE-3 ne contient de virus. Cependant, le nombre de virus transmis par le courrier électronique ne cesse de grimper en flèche depuis le moment où nous avons reçu BCE-3. Nous pouvons les filtrer à l'aide d'un anti-virus et ne traiter que les courriels non contaminés.

les attachements Très peu de courriels contiennent des attachements. Ceux-ci ne représentent pas un problème puisque nous les ignorons tout simplement. S'ils contiennent des informations utiles, nous ne pouvons pas les utiliser parce qu'il n'est pas possible de reconnaître le format d'un fichier en attachement sans indice valable.

la publicité non sollicitée Aucun courriel de publicité non sollicitée ne s'est infiltré dans BCE-3 à l'exception de quelques offres d'affaires sérieuses. Il n'aurait pas été nécessairement mauvais d'en avoir plus, car nous aurions pu entraîner le classifieur avec cette catégorie supplémentaire. Par contre, la publicité non sollicitée n'est vraiment pas préoccupante car plusieurs études démontrent qu'il est facile de classer cette sorte de courriel [1, 4].

les demandes multiples Une portion minime des courriels contiennent au moins deux demandes qui n'ont aucun rapport entre elles, comme le prix des actions pour une date et une requête de rapport annuel. Puisque notre système fait vérifier les suivis suggérés par un préposé, la classe où aboutit un tel courriel est peu importante s'il n'est pas égaré. Une seule des demandes n'a qu'à s'apparenter à la catégorie choisie pour que le classement soit jugé correct.

4.3 Le domaine de discours de BCE-3

Il est possible que les messages contenant plusieurs demandes et les échanges multiples ne causent plus de problèmes que nous le pensions. Et en travaillant avec les messages, nous avons cru remarquer que le domaine de discours est plutôt restreint. Nous définissons le domaine de discours comme l'étendue du vocabulaire, c'est-à-dire le nombre total de mots uniques employés pour l'ensemble des messages. Plus ce total est bas, moins il y a de chance qu'il y ait des mots propres à chaque catégorie et des termes discriminants pour bien classer les messages.

4.3.1 L'objet des messages

Pour vérifier nos hypothèses, nous avons identifié le ou les buts précis de chaque message. Cette identification ne laisse aucune place à la subjectivité. Il ne s'agit pas de regrouper les messages en classes générales mais seulement d'en connaître l'objet. Un message peut avoir plus d'un but et peut donc se retrouver dans plusieurs groupes en même temps. Par exemple, le message de la figure 4.5 a trois buts: rapport annuel, rapport trimestriel et liste de distribution. La table 4.10 montre que la majorité des messages ont un seul but.

```
J'aimerais recevoir un rapport annuel et trimestriel.
Et pouvez-vous m'ajouter à votre liste de distribution.  Merci
```

Fig. 4.5: Exemple de message avec plusieurs buts

objets	**messages**	**%**
1	973	83.1
2	147	12.6
3	36	3.1
4	11	0.9
5	4	0.3
total	1171	100.0

Tab. 4.10: Nombre d'objets par message pour BCE-3

Nous avons accumulé des statistiques sur deux versions de BCE-3. La première est le corpus intégral et contient donc tous les messages sans exception. La deuxième est une version stricte. Elle ne tient compte que des messages avec un seul but et qui ne sont pas des courriels additionnels d'un échange multiple. Les 25 objets les plus fréquents ainsi que le nombre de messages pour chaque groupe sont affichés dans la table 4.11. La table 4.12 présente les 42 objets les moins fréquents. La catégorie `incertain` contient les messages dont nous n'avons pas pu identifier clairement l'objet. Nous reviendrons sur la première colonne dans la section 5.2.

ensemble	**objet**	**intégral**	**strict**
A	share price	249	208
	annual report	175	100
	mailing list	100	55
	dividend	72	43
	stock split	70	41
B	* Nortel spin off	68	40
	dividend r. p.	68	39
	* Teleglobe takeover	42	37
	explanation	32	29
	investor kit	52	25
	preferred shares	26	25
	subsiduaries	32	24
C	BCE personnel	22	19
	negative comment	20	18
	certificate	29	17
	share purchase	32	15
	stock info	16	15
	earnings	31	14
	positive comment	42	13
	conference	23	12
	shares possessed	19	11
	share transfer	14	11
	annual meeting	15	10
	other financial info	15	10
	* incertain	10	10

Tab. 4.11: Nombre de messages pour les 25 objets les plus fréquents de BCE-3

Nous avons écarté certains groupes, identifiés par un astérisque, pour le reste de l'analyse. Le premier est bien sûr les messages appartenant au groupe `incertain`, parce qu'il est difficile de travailler avec des messages dont le but est inconnnu. Le

deuxième groupe retranché est `Teleclone`. Il comporte deux messages de clients ayant confondu les compagnies Teleclone et Teleglobe et n'ont donc aucun rapport avec BCE. Les autres groupes ignorés font référence à un événement ponctuel: `Aliant offer`, `B-split Corp`, `Bell Cablemedia acquisition`, `CTV takeover`, `Electrohome takeover`, `Nortel spin off`, `Teleglobe takeover` et `Unique Broadband Systems`.

4.3.2 La fréquence des mots

Nous avons calculé les mots les plus fréquents de tous les groupes pour nos deux versions de BCE-3. Avant nos calculs, nous avons éliminé les numéraux et les mots vides de sens. Aussi, nous avons tronqué ceux qui restaient à l'aide de l'algorithme de Porter [34]. Les statistiques sont loin d'être encourageantes. Pour les 22 plus grosses classes, plusieurs mots fréquents que nous pensions utiles pour bien séparer les catégories se retrouvent en grande quantité à l'intérieur de plusieurs classes. De plus, la majorité des mots sont présents dans au moins cinq classes, ce qui n'est guère mieux. Et les mots uniques à une ou deux catégories sont presques introuvables. Et pour les autres groupes de moindre importance, c'est aussi problématique. Les même mots se retrouvent autant dans ces classes que dans les plus importantes, et il y a très peu de mots propres aux petites catégories. Tout cela risque de nuire passablement à une classification qui contient plus d'une dizaine de catégories.

4.3.3 La version stricte

Nous sommes conscients que la version stricte de BCE-3 s'éloigne un peu du corpus original car plusieurs messages sont ignorés. Par contre, ces derniers sont loin d'être les plus représentatifs de BCE-3. Comme nous le verrons au prochain chapitre, nous avons débuté nos expériences avec la version originale. Mais en raison des résultats, nous avons poursuivi avec la version stricte. Cela nous a permis de vérifier si les messages additionnels et les messages à buts multiples causent réellement des problèmes supplémentaires.

objet	intégral	strict
business offer	10	9
quarterly report	42	8
∗ CTV takeover	6	5
BCE ownership	7	4
share sale	4	4
web site	8	3
name correction	3	3
resume	3	3
proxy	10	2
market share	5	2
number of employees	5	2
BCE services	3	2
analyst report	2	2
balance sheet	2	2
BCE history	2	2
BCED	2	2
beta coefficient	2	2
∗ Teleclone	2	2
P/E ratio	9	1
T5	3	1
BC-3658 form	2	1
management report	2	1
∗ Aliant offer	1	1
∗ B-split Corp.	1	1
BCE acronym	1	1
∗ Bell Cablemedia acquisition	1	1
blue letter of transmittal	1	1
capital gain calculation	1	1
∗ Electrohome takeover	1	1
financial information form	1	1
internship	1	1
promotion material	1	1
stock market report	1	1
T3	1	1
annual information form	4	0
environmental report	3	0
research report	3	0
20F form	2	0
8K report	2	0
eday	2	0
social report	2	0
∗ Unique Broadband Systems	1	0

Tab. 4.12: Nombre de messages pour les 42 objets les moins fréquents de BCE-3

Chapitre 5

Résultats

Dans ce chapitre, nous présentons les principaux résultats des expériences que nous avons effectuées. Pour prendre nos décisions, nous nous sommes basés principalement sur l'analyse de BCE-3, décrite au chapitre 4. Nous avons aussi pris en compte les aspects théoriques de la classification de textes et du courrier électronique, résumés au chapitre 2. De plus, nous n'avons pas négligé les considérations pratiques reliées à l'implantation de Merkure dans l'environnement de BCE que nous avons obervées au chapitre 3.

5.1 Les premières expérimentations

5.1.1 La nouvelle classification

Après l'analyse de BCE-3, nous avons établi une nouvelle classification du corpus. Nous avons demandé à deux personnes de lire tous les messages et de les séparer selon leur contenu. Seulement une des deux personnes possédait des connaissances en informatique et en traitement des langues naturelles, et aucune n'avait de connaissances dans le domaine des relations aux investisseurs. Comme point de référence, nous leur avons fourni la liste des catégories initiales telle qu'établie par les préposés de BCE. Les deux personnes n'avaient pas le droit de se consulter. Lorsqu'un message s'apparentait à plus d'une classe, les personnes avaient pour directive de l'inclure dans la catégorie prenant

le plus de place à l'intérieur du message. Et si la classe la plus importante d'un message ne pouvait pas être déterminée, le message devait être classé dans la plus petite catégorie.

Lorsque que les deux personnes ont complété leur classification respective, nous les avons jumelées. Nous avons rencontré très peu de problèmes pendant cette étape parce qu'elles étaient presque identiques. Les deux classifications contenaient les mêmes catégories, mais l'une d'elles possédait une classe mineure de plus. Nous avons décidé de l'inclure dans la catégorie `general` car elle ne contenait qu'une vingtaine de questions financières plus ou moins reliées entre elles. Un peu plus de 87% des messages étaient situés au même endroit. Nous avons apparié les cas litigieux avec la catégorie prédominante de leur contenu, ou avec la classe la plus petite en dernier recours. La table 5.1 illustre la nouvelle classification que nous avons utilisée pour nos premières expérimentations. Elle ressemble à une version épurée de celle que nous avons obtenue des préposés de BCE.

catégorie	**messages**	**description**
share price	226	tout ce qui concerne la valeur des actions de BCE
reports	179	demandes de rapports annuels, trimestriels, etc.
subsiduaries	107	questions sur les compagnies sous la tutelle de BCE
individuals	105	questions personnelles d'investisseurs
stock	92	messages concernant les actions, mais non la valeur
general	71	tout ce qui ne peut pas être placé ailleurs
dividend r. p.	68	questions sur le réinvestissement des dividendes
mailing list	66	demandes de modification aux listes de distribution
comments	62	commentaires positifs ou négatifs
dividend	61	tout ce qui concerne les dividendes
investor kit	48	demandes des documents pour les clients potentiels
earnings	30	messages concernant les gains
preferred shares	28	questions concernant les actions priviliégiées
conference	28	messages concernant les diverses conférences
total	1171	

Tab. 5.1: Nouvelle classification de BCE-3

5.1.2 Les premiers résultats

Pour nos expérimentations, nous avons utilisé trois classifieurs différents: un classifieur vectoriel (kppv, section 2.1.1), un classifieur probabiliste (Bayes, section 2.1.2), et un classifieur à base de règles (Ripper, section 2.1.3). Chaque message n'appartient qu'à une catégorie, et ne peut être classé que dans l'une d'entre elles. L'efficacité d'un classifieur est donc calculée par le ratio de bonnes prédictions sur le nombre total de cas. Nous avons observé les effets du prétraitement avec les trois classifieurs comme point de départ. Ces résultats nous ont aussi servi de référence et de comparaison pour la suite des expérimentations. Plusieurs sources de bruit possibles sont observées à l'aide des différents prétraitements:

simple Tous les mots du texte sont conservés.

numéraux Les numéraux sont enlevés.

mots vides Les mots vides de sens sont enlevés.

troncature Les mots sont tronqués selon l'algorithme de Porter.

mots rares Les mots dont la fréquence dans le corpus est inférieure à un seuil sont enlevés.

Nous avons conduit notre série d'expériences avec trois séparations différentes du corpus:

majoritaire Chaque classe est divisée en deux ensembles avec un ratio de quatre messages dans l'ensemble d'entraînement pour un dans l'ensemble de validation. L'ensemble d'entraînement, largement majoritaire, regroupe donc 873 messages tandis que l'ensemble de validation ne possède que 298 messages.

égale Chaque catégorie est divisée en deux ensembles de même taille. L'ensemble d'entraînement comprend un total de 582 messages et l'ensemble de validation en contient 589.

réduite Le nombre de messages de toutes les catégories est égalisé à la baisse, ce qui donne, pour chaque catégorie, 15 messages d'entraînement et 6 messages de validation. Cela représente donc 210 messages d'entraînement et 84 messages pour la validation.

Dans tous les cas, le choix d'un ensemble pour chaque message est entièrement aléatoire. Les séparations égale et majoritaire servent à observer le rendement des classifieurs avec les classes telles qu'elles sont. Normalement, pour un ensemble de classes homogènes, il ne devrait pas y avoir de différences majeures entre ces deux séparations et la séparation réduite. Par souci de simplicité, nous avons opté d'équilibrer les classes vers le bas. Les tables 5.3, 5.2 et 5.4 montrent l'efficacité, en pourcentage, des classifieurs pour les trois séparations selon le type de prétraitement effectué. Pour l'élimination des mots rares, le nombre entre parenthèses indique le seuil à atteindre pour être conservé. Nous avons essayé le classifieur kppv avec des valeurs de 10, 20, 30, 40 et 50 pour la constante k.

Les deux principales caractéristiques qui ressortent des résultats sont la faiblesse et l'uniformité des résultats. Le rendement est faible, mais ni le classifieur ni le prétraitement ne semble être en cause. Premièrement, peu importe la séparation, le prétraitement n'influence presque pas les résultats. Le fait de conserver ou non les numéraux altère à peine l'efficacité des classifieurs. Les mots vides de sens influencent un peu plus les résultats lorsqu'ils sont absents, mais généralement d'une façon négative. Pour la troncature, les variations sont aussi négligeables même si elles sont pour la plupart positives. Ce sont les mots rares qui apportent le plus d'impact. Alors que leur disparition entraîne une baisse du rendement presque partout, le 10ppv en profite car il ne se fie qu'à une poignée très restreinte de documents semblables pour faire sa prédiction. Ces mots peu fréquents augmentent de beaucoup la similarité entre deux documents qui en ont un en commun et propulsent de mauvais exemples en tête de liste. Si le classifieur ne se fie qu'à quelques documents pour prendre une décision, il utilise uniquement ces mauvais documents. S'il prend un ensemble de consulation plus grand, ces erreurs se perdent dans la masse de documents non affectés par les mots rares.

prétraitement	**Bayes**	**10ppv**	**20ppv**	**30ppv**	**40ppv**	**50ppv**	**Ripper**
simple	54.5	52.8	57.9	59.3	58.3	54.5	53.8
numéraux	53.4	53.4	58.3	59.7	59.7	57.2	51.4
mots vides	54.8	48.6	55.9	57.6	56.6	54.1	55.5
troncature	55.2	49.7	57.6	59.3	57.9	54.5	54.1
mots rares (5)	52.7	56.9	58.6	58.3	56.2	51.4	51.0
mots rares (10)	51.9	58.3	60.7	57.2	56.9	50.7	54.1

Tab. 5.2: Efficacité des classifieurs selon le prétraitement sur la séparation majoritaire

prétraitement	**Bayes**	**10ppv**	**20ppv**	**30ppv**	**40ppv**	**50ppv**	**Ripper**
simple	54.1	52.9	56.0	56.0	53.2	49.0	49.9
numéraux	54.4	51.8	56.0	57.2	53.9	49.0	46.6
mots vides	53.0	48.5	54.3	54.3	53.2	47.5	50.4
troncature	55.0	49.9	55.2	54.6	53.8	47.8	46.8
mots rares (5)	52.4	56.0	54.3	54.8	50.6	46.9	47.1
mots rares (10)	51.8	53.9	53.9	53.2	50.3	46.4	46.2

Tab. 5.3: Efficacité des classifieurs selon le prétraitement sur la séparation égale

prétraitement	**Bayes**	**10ppv**	**20ppv**	**30ppv**	**40ppv**	**50ppv**	**Ripper**
simple	42.9	46.4	50.0	46.4	45.2	44.2	44.0
numéraux	41.7	47.6	51.2	45.2	46.4	45.5	44.0
mots vides	39.3	47.6	45.2	46.4	41.7	42.9	45.2
troncature	42.9	50.0	48.8	54.8	48.8	47.6	44.0
mots rares (5)	39.2	47.6	46.4	44.0	45.2	42.9	39.3
mots rares (10)	37.5	46.4	36.9	41.7	40.5	39.3	29.8

Tab. 5.4: Efficacité des classifieurs selon le prétraitement sur la séparation réduite

Ensuite, aucun classifieur ne se distingue des autres. Bien que le prétraitement ne les affecte pas tous de la même manière, leur rendement reste à l'intérieur d'intervalles similaires. Ripper performe un peu moins bien sur la séparation égale, mais il s'agit de la seule différence notable. À part cela, les résultats de kppv sont légèrement mieux.

Finalement, c'est la séparation qui influence le plus le rendement. Les pourcentages augmentent proportionnellement au nombre de messages d'entraînement. Il y a un écart d'environ 5% entre les séparations réduite et égale, et autant entre les séparations égale et majoritaire. C'est tout fait normal puisque le nombre de messages dans le corpus n'est pas très grand. Toutefois, cela peut aussi signifier que les classes sont plus ou moins homogènes. Lorsqu'une catégorie ne regroupe que des messages vraiment semblables, la grosseur de l'échantillon d'entraînement ne change presque rien au-delà

d'un certain nombre. Par contre, une catégorie hétérogène regroupant plusieurs petites sous-classes similaires est habituellement avantagée par un ensemble d'entraînement plus grand. Une autre possibilité est la présence des courriels additionnels faisant partie d'un échange multiple et des messages avec plus d'un but. Finalement, il est aussi plausible que le vocabulaire employé pour chaque classe soit presque partout le même.

5.2 Les expérimentations sur la version stricte

5.2.1 Les classifications épurées

Pour la suite des expériences, nous n'avons utilisé que la version stricte de BCE-3, décrite dans la section 4.3. Nous avons donc laissé de côté les courriels additionnels d'un échange multiple et les messages ayant plus d'un but. Nous avons désigné trois ensembles de messages basés sur la grosseur des groupes:

A Les cinq plus gros groupes: `annual report`, `dividend`, `mailing list`, `share price` et `stock split`. Il s'agit des groupes avec 40 messages et plus.

B Les cinq plus gros groupes suivants ceux de A: `dividend reinvestment plan`, `explanation`, `investor kit`, `preferred shares` et `subsiduaries`. Tous ces groupes possèdent 20 messages et plus.

C Les douze plus gros groupes suivants ceux de B: `annual meeting`, `certificate`, `conference`, `earnings`, `financial info`, `negative comment`, `personnel`, `positive comment`, `share purchase`, `share transfer`, `shares possessed` et `stock info`. Ces groupes ont tous dix messages et plus.

Ces ensembles sont indiqués dans le tableau 4.11, et les groupes marqués d'un astérisque n'en font pas partie. À partir de ces ensembles, nous avons établi trois nouvelles classifications différant sur le nombre de catégories:

C5 Correspond à l'ensemble A. L'ensemble d'entraînement comprend 225 messages et l'ensemble de validation en a 222.

C10 Contient les ensembles A et B. L'ensemble d'entraînement possède 298 messages contre 291 pour l'ensemble de validation.

C22 Regroupe les ensembles A, B et C. L'ensemble d'entraînement contient 384 messages et l'ensemble de validation en compte 370.

Dans les trois cas, chaque catégorie est divisée également et aléatoirement entre l'ensemble d'entraînement et l'ensemble de validation. La classification C5 est idéale. Elle ne contient que cinq catégories, et le nombre de messages dans chacune est suffisant. La deuxième classification, C10, est moins intéressante. Elle ne regroupe dix classes, mais la moitié ont entre 20 et 40 messages, ce qui est peu en comparaison des cinq catégories de l'ensemble A. La troisième classification n'est pas adéquate du tout. Plus de la moitié des classes ne contiennent même pas 20 messages, ce qui est franchement insuffisant à des fins de classification. Nous l'avons tout de même utilisée pour voir si l'ajout des plusieurs petites catégories nuit aux prédictions faites sur les messages des classes les plus importantes.

Nous sommes conscients que nos expériences avec ces classifications ne doivent pas être considérées comme étant entièrement valides. Puisqu'elles ne contiennent que les messages avec les buts les plus fréquents, il est normal que le rendement des classifieurs soit meilleur que si nous utilisions tous les messages. Par contre, c'est justement l'idée derrière nos expérimentations. Nous cherchons à isoler certains facteurs pouvant nuire à la classification. Si ces résultats sont aussi désastreux que les précédents, il est fort probable que nous soyons aux prises avec un domaine de discours restreint. Sinon, dans le cas où un nombre plus petit de classes améliore le rendement, nous avons plutôt affaire à un surnombre de catégories. Il ne faut pas non plus écarter l'éventualité que ces deux possibilités soient en cause.

5.2.2 Les résultats biaisés

Nous avons testé les mêmes classifieurs que pour nos expériences précédentes. Mais au lieu de nous limiter aux prétraitements antérieurs, nous avons élargi nos expérimentations en jumelant les possibilités les plus prometteuses. Les tables 5.5, 5.6 et 5.7 affichent

les résultats des classifieurs selon le prétraitement et la séparation.

prétraitement	**Bayes**	**10ppv**	**20ppv**	**30ppv**	**40ppv**	**50ppv**	**Ripper**
simple	91.0	86.7	84.7	82.9	80.2	74.3	86.5
numéraux	90.5	87.8	84.7	82.9	80.6	74.3	88.3
mots vides	89.6	86.9	87.4	82.4	80.2	76.6	89.6
troncature	88.7	82.9	85.6	83.8	80.6	77.5	87.4
mots rares (5)	85.3	85.6	83.3	80.2	77.0	74.8	84.7
mots rares (10)	78.0	66.7	68.7	65.6	63.6	60.1	85.1
num. et m. v.	92.3	86.0	87.4	84.7	82.4	76.6	86.5
num. et tron.	89.6	85.1	86.9	83.8	81.5	76.6	91.4
m.v. et tron.	90.5	82.4	86.5	86.0	85.6	75.7	91.4
num., m.v. et tron.	91.0	83.8	88.3	87.4	86.5	77.0	90.5

Tab. 5.5: Efficacité des classifieurs selon le prétraitement sur C5

prétraitement	**Bayes**	**10ppv**	**20ppv**	**30ppv**	**40ppv**	**50ppv**	**Ripper**
simple	78.4	73.5	75.9	71.5	67.4	63.6	75.6
numéraux	77.7	73.9	74.9	72.5	68.4	65.3	74.2
mots vides	77.0	73.5	76.3	72.2	68.4	62.2	74.9
troncature	76.3	72.8	75.9	72.8	69.1	65.3	74.6
mots rares (5)	71.9	72.2	69.8	67.4	64.6	61.2	70.4
mots rares (10)	70.7	66.7	68.7	65.6	63.6	60.1	65.6
num. et m. v.	77.3	74.2	77.0	74.6	70.1	63.2	73.5
num. et tron.	77.0	71.1	77.3	75.3	70.8	67.7	73.7
m.v. et tron.	76.6	72.8	74.0	75.3	73.9	68.4	80.1
num., m.v. et tron.	77.7	74.2	77.0	78.0	73.5	70.8	80.8

Tab. 5.6: Efficacité des classifieurs selon le prétraitement sur C10

prétraitement	**Bayes**	**10ppv**	**20ppv**	**30ppv**	**40ppv**	**50ppv**	**Ripper**
simple	67.3	65.1	63.2	60.3	57.0	54.0	61.6
numéraux	67.8	65.4	64.0	61.6	57.6	54.9	59.7
mots vides	66.2	64.3	62.4	60.0	57.8	54.3	61.9
troncature	66.8	61.9	63.0	61.9	59.5	55.7	60.8
mots rares (5)	64.3	65.4	61.4	58.9	55.1	52.7	58.9
mots rares (10)	61.8	59.7	58.1	55.4	52.1	52.4	52.4
num. et m.v.	67.6	64.9	65.1	60.8	59.7	54.9	64.0
num. et tron.	68.1	62.2	65.9	61.9	59.5	56.2	60.8
m.v. et tron.	66.0	63.5	62.4	63.0	60.8	57.0	64.3
num., m.v. et tron.	67.3	62.7	64.3	65.4	61.9	58.9	64.6

Tab. 5.7: Efficacité des classifieurs selon le prétraitement sur C22

Cette seconde série d'expériences nous a donné des résultats nettement meilleurs. Ici encore, le prétraitement ne semble pas affecter outre-mesure les performances des classifieurs. La tendance générale est qu'une épuration complète des messages semble être la meilleure approche. Les trois classifieurs offrent un rendement supérieur lorsque les numéraux et les mots vides de sens sont enlevés, et que les mots restant sont tronqués. Par conséquent, nous avons poursuivi nos expérimentations uniquement avec ce prétraitement. Avec cette classification plus nette que la précédente, le classifieur kppv performe moins bien que les deux autres.

Une autre caractéristique notable est que les pourcentages diminuent en fonction du nombre de catégories utilisées, mais ils ne rejoignent pas ceux de notre première classification. Même C22 semble plus facile à gérer que notre première classification, avec un écart d'environ 10%. Elle contient pourtant huit classes de plus, mais seulement 754 messages comparativement aux 1171 messages initiaux. Cela nous porte à croire que les messages additionnels et à buts multiples sont effectivement deux sources de bruit non négligeables, et que la majorité des erreurs de classement se situent au niveau des messages qui n'appartiennent pas aux catégories les plus grosses.

5.2.3 La provenance des erreurs

Pour observer d'où viennent les erreurs, les tables 5.8, 5.9, 5.10 et 5.11 représentent les matrices de confusion pour C5. De même, les tables 5.12, 5.13 et 5.14 montrent les matrices de confusion pour C10. Une matrice de confusion sert à évaluer la qualité d'une classification. Pour chacune des catégories, les colonnes indiquent combien de messages ont été associés à chaque classe. Par exemple, pour Bayes avec C5, tous les messages appartenant à `share price` sont bien classés sauf deux: l'un a été étiqueté `dividend` et l'autre `stock split`. Puisque l'analyse de BCE-3 démontre que les mêmes mots sont présents dans plusieurs catégories, nous avons pensé qu'utiliser des expressions à la place de mots simples serait bénéfique. Nous en avons profité pour essayer un autre classifieur, celui-là à base de cooccurences de mots (section 2.1.4).

	prédiction					
catégorie	**a. r.**	**div.**	**m. l.**	**s. p.**	**s. s.**	
annual report	50	0	0	0	0	
dividend	0	15	0	5	1	
mailing list	1	0	23	2	1	
share price	0	1	0	102	1	
stock split	0	0	0	8	12	

Tab. 5.8: Matrice de confusion de Bayes pour C5

	prédiction					
catégorie	**a. r.**	**div.**	**m. l.**	**s. p.**	**s. s.**	
annual report	49	0	0	1	0	
dividend	1	14	0	6	0	
mailing list	7	0	19	1	0	
share price	1	0	0	101	2	
stock split	0	0	0	9	11	

Tab. 5.9: Matrice de confusion de 30ppv pour C5

	prédiction					
catégorie	**a. r.**	**div.**	**m. l.**	**s. p.**	**s. s.**	
annual report	47	0	2	1	0	
dividend	0	19	0	1	1	
mailing list	3	1	21	2	0	
share price	2	2	1	95	4	
stock split	0	0	0	1	19	

Tab. 5.10: Matrice de confusion de Ripper pour C5

	prédiction					
catégorie	**a. r.**	**div.**	**m. l.**	**s. p.**	**s. s.**	**?**
annual report	36	0	1	0	0	13
dividend	0	13	0	0	0	8
mailing list	1	0	24	0	0	2
share price	0	0	0	65	7	32
stock split	0	0	0	4	12	4

Tab. 5.11: Matrice de confusion du classifieur à base de cooccurences pour C5

En raison de la piètre performance de ce dernier, nous n'avons inclus que sa matrice de confusion pour C5. La dernière colonne de la table correspondante indique le nombre de messages qu'il n'a pas pu classer en raison d'un niveau de confiance trop bas. Comme les résultats obtenus avec cette classification sont supposés être biaisés vers le haut, nous avons décidé de ne pas nous attarder plus longtemps avec ce classifieur.

	prédiction									
catégorie	**a.r.**	**div.**	**d.r.p.**	**exp.**	**i.v.**	**m.l.**	**p.s.**	**s.p.**	**s.s.**	**sub.**
annual report	49	0	0	0	0	1	0	0	0	0
dividend	0	14	2	0	0	0	0	4	1	0
div. r. p.	1	1	14	0	0	0	0	3	0	0
explanation	0	1	0	4	0	1	0	5	3	0
investor kit	4	0	0	0	7	1	0	0	0	0
mailing list	1	0	2	0	1	21	0	2	0	0
pref. shares	1	0	0	2	0	0	5	5	0	0
share price	0	1	2	0	0	0	0	96	5	0
stock split	0	0	0	0	0	0	0	7	13	0
subsiduaries	2	0	0	1	0	0	0	4	2	3

Tab. 5.12: Matrice de confusion de Bayes pour C10

	prédiction									
catégorie	**a.r.**	**div.**	**d.r.p.**	**exp.**	**i.v.**	**m.l.**	**p.s.**	**s.p.**	**s.s.**	**sub.**
annual report	50	0	0	0	0	0	0	0	0	0
dividend	1	14	2	0	0	0	0	4	0	0
div. r. p.	0	1	16	0	0	0	0	2	0	0
explanation	0	0	0	0	1	0	0	13	0	0
investor kit	3	0	0	0	8	1	0	0	0	0
mailing list	3	0	0	0	0	23	0	1	0	0
pref. shares	2	0	0	0	0	0	4	6	0	0
share price	1	0	1	0	0	0	0	100	2	0
stock split	0	0	0	0	0	0	0	8	12	0
subsiduaries	3	0	0	0	0	0	0	9	0	0

Tab. 5.13: Matrice de confusion de 30ppv pour C10

Tout comme les pourcentages d'efficacité, les matrices de confusion des différents classifieurs pour C5 se ressemblent considérablement. Les divisions entre les classes sont assez nettes, à l'exception des catégories `share price` et `stock split` qui paraissent plus indivisibles. Cependant, les choses se compliquent un peu plus avec C10. Les messages appartenant aux cinq plus grosses classes restent presque aussi faciles à identifier,

	prédiction									
catégorie	**a.r.**	**div.**	**d.r.p.**	**exp.**	**i.v.**	**m.l.**	**p.s.**	**s.p.**	**s.s.**	**sub.**
annual report	45	0	1	0	0	1	0	3	0	0
dividend	0	18	2	0	0	0	0	0	1	0
div. r. p.	0	1	17	0	0	0	1	0	0	0
explanation	0	0	0	0	0	0	0	14	0	0
investor kit	0	0	0	0	7	0	0	5	0	0
mailing list	0	0	1	1	0	23	0	2	0	0
pref. shares	0	0	0	0	0	0	8	4	0	0
share price	2	1	3	0	0	1	0	94	4	0
stock split	0	0	0	0	0	0	0	1	19	0
subsiduaries	0	0	0	0	0	0	0	7	0	4

Tab. 5.14: Matrice de confusion de Ripper pour C10

mais les messages des nouvelles catégories posent d'énormes problèmes. En particulier, les messages appartenant au groupe `explanation` sont classés partout sauf à la bonne place. Le classement est aussi moins net que pour C5.

Afin de déterminer précisément d'où vient la baisse de rendement, nous avons observé l'efficacité des trois ensembles A, B et C à l'intérieur des classifications C5, C10 et C22. La table 5.15 illustre le rendement des classifieurs sur les trois ensembles selon la classification utilisée. Par exemple, la première ligne indique le pourcentage des messages appartenant à A qui sont bien classés en utilisant la classification C5. Malgré le fait que les matrices de confusion se ressemblent énormément, nous avons décidé d'essayer une combinaison des trois classifieurs. La prédiction de cette combinaison est simplement la catégorie majoritaire parmi les prédictions des trois classifieurs. Lorsque les trois classifieurs donnent des réponses différentes, celle ayant le plus haut niveau de confiance est gardée.

ensemble	**classification**	**Bayes**	**30ppv**	**Ripper**	**combinaison**
A	C5	91.0	87.4	90.5	91.0
A	C10	86.9	89.6	89.6	92.8
A	C22	85.1	88.7	87.8	89.6
B	C10	47.8	40.6	52.2	50.7
B	C22	40.6	49.3	42.0	43.8
C	C22	40.5	13.9	19.0	20.4

Tab. 5.15: Efficacité des ensembles A, B et C pour C5, C10 et C22

Avec cette table, nous voyons que l'écart entre les classifieurs et la combinaison des classifieurs est minime pour A. Pour B et C, les écarts plus grands sont dus au fait que les classifieurs ne commettent pas leurs erreurs sur les mêmes messages. De plus, nous constatons que les messages de l'ensemble A sont toujours faciles à classer.

5.3 La classification proposée

Nos expériences précédentes sont intéressantes en théorie, mais le sont beaucoup moins en pratique. Nous avons donc entrepris quelques essais supplémentaires en vue d'en faire une implantation concrète. À l'aide des matrices de confusion et de quelques tests additionnels, nous avons convenu de la classification illustrée à la table 5.16. Cette classification est composée de 818 messages, soit seulement ceux appartenant à la version stricte de BCE-3. De plus, ils ne doivent pas faire partie des catégories laissées de côté comme les classes temporelles.

catégorie	**messages**	**description**
dividend r. p.	39	questions sur le réinvestissement des dividendes
stock split	41	questions sur le fractionnement des actions
dividend	43	tout ce qui concerne les dividendes
mailing list	55	demandes de modification aux listes de distribution
report	137	demandes de rapports annuels, trimestriels, etc.
share price	234	tout ce qui concerne la valeur des actions de BCE
general	269	tout ce qui ne peut pas être placé ailleurs
total	818	

Tab. 5.16: Classification proposée

Nous avons vu avec les matrices de confusion qu'une telle classification, sans la catégorie générale, donne de bons résultats. La dernière difficulté est de bien séparer les messages de ces classes des messages généraux. Pour trouver le meilleur rendement, nous avons aussi essayé une classification binaire. La classe `general` et une deuxième catégorie comprenant tous les messages des 6 autres classes mentionnées ci-dessus. Pour ces derniers essais, nous nous sommes limités au classifieur Ripper avec un prétraitement complet (élimination des numéraux et des mots vides de sens, et utilisation de la troncature). La table 5.17 affiche les résultats pour la classification binaire et la classification

de la table 5.16. Le rendement de cette dernière classification est aussi vérifiée sur un sous-ensemble de BCE-4. Il regroupe les 144 messages en anglais reçus au cours du mois de mars 2002. Pour ces trois dernières classifications, nous avons utilisé une validation croisée pour mesurer l'efficacité.

classification	**corpus**	**efficacité**
2 classes	BCE-3	82.5
7 classes	BCE-3	80.1
7 classes	BCE-4	79.6

Tab. 5.17: Efficacité avec la classification proposée

L'option de n'utiliser que deux catégories est meilleure pour différencier les messages généraux des autres, mais l'écart avec l'alternative n'est pas assez grand. Le rendement du classifieur sur les six autres classes n'est pas assez élevé pour nous permettre une classification en deux étapes: la séparation des messages généraux des autres messages, et la classification de ces derniers en six catégories. De plus, cette classification semble être adéquate pour traiter les messages actuels (BCE-4) avec autant d'efficacité que ceux de BCE-3. Cela veut dire que le corpus BCE-3 représente bien les courriels envoyés quotidiennement aux adresses de BCE, malgré ce que nous pensions initialement à cause de la distribution très inégale des courriels.

Ces résultats atteignent l'efficacité que nous escomptions et la tâche effectuée avec cette classification est adéquate dans le cadre de Merkure. Elle ne comporte que six catégories (plus une classe générale) mais elle ne peut pas contenir de classes plus appropriées. En effet, les messages des classes `dividend reinvestment plan` et `mailing list` ne sont pas traités par les préposés des relations aux investisseurs mais sont plutôt redirigés à leur agent de transfert. De plus, les demandes de rapport ne nécessitent aucun traitement supplémentaire à part l'envoi du document. Aussi, les messages appartenant à la classe `share price` sont tous similaires et requièrent une réponse très factuelle. Le départage de ces messages, qui n'ont pas besoin d'être analysés plus en profondeur, permet d'éviter une quantité appréciable de traitement qui aurait dû être effectué par les autres modules de Merkure. Pour ces raisons, nous avons proposé la classification de la table 5.16 et l'avons implantée dans l'environnement de BCE.

Chapitre 6

Conclusion

Nous avons décrit Merkure, dont l'objectif est d'élaborer un système capable de répondre automatiquement aux courriels destinés au département des relations aux investisseurs de BCE. Merkure est décomposé en plusieurs modules complémentaires. Le module dont il est question dans ce mémoire s'occupe de la classification du courrier électronique.

Nous avons analysé en profondeur le corpus BCE-3, constitué de courriels envoyés à BCE et des réponses à ces courriels. L'un des points importants de cette analyse est qu'il y a peu de mots typiques à une seule classe. Peu importe l'objet du message, ce sont invariablement les mêmes mots qui en décrivent le contenu. C'est aussi avec ce corpus que nous avons essayé un peu plus de 500 combinaisons d'approche de classification différentes.

Nous avons expérimenté avec quelques classifieurs différents: vectoriel, probabiliste, à base de règles, à base de cooccurences de mots et même une combinaison des trois premiers. Nous avons aussi vérifié l'apport de plusieurs formes de prétraitement comme l'élimination des numéraux, des mots vides de sens et des mots rares, la troncature des mots et plusieurs combinaisons de ces différents aspects. Malheureusement, aucune approche ne s'est révélée vraiment supérieure. Nous avons obtenu nos meilleurs résultats en restreignant l'apprentissage sur des messages n'ayant qu'un seul but, en enlevant les numéraux et les mots vides de sens, et en tronquant les mots restants. Nous avons dû nous limiter à six catégories distinctes, plus une catégorie générale, pour obtenir 80%

d'efficacité.

L'efficacité de nos expérimentations a plafonné aux environs de ce résultat, et les deux principales causes semblent être la taille réduite de l'échantillon et un domaine de discours trop restreint. Dans le cas de BCE, les messages envoyés aux relations aux investisseurs sont tous trop semblables pour pouvoir bien définir des classes distinctes. Peu importe le but du message, les mêmes mots reviennent presque immanquablement pour en décrire le contenu. Cela est très gênant, surtout à cause de la superficialité des classifieurs, qui ne font pas d'analyse sémantique profonde. En raison de la quantité et de la variété des tests effectués, nous ne pensons pas que les classifieurs actuels soient appropriés pour un domaine aussi restreint. Bien qu'ils soient capables de classer adéquatement trois messages sur quatre, il y a encore une grande place pour l'amélioration. En ce sens, nous croyons que le meilleur chemin à suivre pour améliorer les résultats est d'utiliser des caractéristiques additionnelles en plus des mots. De plus, nous sommes convaincus que le classifieur devrait établir plus que des liens de cooccurrence avec les mots, puisque ceux-ci sont souvent les mêmes d'un message à l'autre.

Cependant, nous ne croyons pas qu'utiliser des techniques plus approfondies soit une solution intéressante dans le cadre de Merkure. En effet, il ne serait pas justifié d'utiliser des méthodes aussi complexes que celles employées par les modules de question-réponse et de raisonnement à base de cas, qui compléteront la classification. C'est d'ailleurs en partie pour cette raison que nous considérons satisfaisants ces résultats. Le classifieur est capable de reconnaître la majorité des messages qui n'ont pas besoin de traitement complexe à cause du suivi à utiliser, et laisse le traitement des autres messages aux autres modules ayant des techniques plus approfondies. Par contre, dans l'intérêt général de la classification de textes, il serait intéressant de voir ce que pourrait donner une analyse sémantique poussée pour classer un ensemble de documents assez similaires, comme c'est le cas pour les relations aux investisseurs de BCE.

Bibliographie

[1] I. Androutsopoulos, J. Koutsias, K. V. Chandrinos et C. D. Spyropoulos. An experimental comparison of naive bayesian and keyword-based anti-spam filtering with personal e-mail messages. *Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval*, Athène, Grèce, 2000, p. 160 - 167.

[2] R. Baeza-Yates et B. Ribeiro-Neto. *Modern information retrieval*, Addison-Wesley, 1999, 513 p.

[3] L. Breiman, J. H. Friedman, R. A. Olshen et C. J. Stone. *Classification and Regression Trees*, Wadsworth Int. Group, 1984.

[4] X. Carreras et L. Màrquez. *Boosting trees for anti-spam email filtering [extended].* Rapport de recheche LSI-01-44-R, Universitat Politècnica de Catalunya, 2001. www.lsi.upc.es/dept/techreps/html/R01-44.html

[5] J. Cheng et R. Greiner. Comparing bayesian network classifiers. *Proceedings of the 15th conference on uncertainty in artificial intelligence*, Stockholm, Suède, 1999, p. 101 - 108.

[6] W. W. Cohen. Fast effective rule induction. *Proceedings of the 12th International Conference on Machine Learning*, 1995, p. 115 - 123.

[7] M. Collot et N. Belmore. Electronic language: a new variety of English. *Computer-mediated Communication: Linguistic, Social and Cross-Cultural Perspectives*, John Benjamins, 1996, pp. 13-28.

[8] G. T. Dietterich. Machine learning research: four current directions. *AI Magazine*, vol. 18, no. 4, p. 97 - 136.

[9] R. O. Duda et P. E. Hart. *Pattern classification and scene analysis*, John Wiley & Sons, 1973, 680 p.

[10] J. Dysart. Email marketing grows up: a primer for the new millenium. *NetWorker*, vol. 3, no. 4, 1999, p. 40 - 41.

[11] K. S. Eklundh et C. Macdonald. The use of quoting to preserve context in electronic mail dialogues. *IEEE Transactions on Professional Communication*, vol. 37, no. 4, 1994, p. 197-202.

[12] N. Freed et N. Borenstein. MIME part one: format of internet message bodies. *RFC 2045*, novembre 1996. www.imc.org/rfc2045

[13] N. Freed et N. Borenstein. MIME part two: media types. *RFC 2046*, novembre 1996. www.imc.org/rfc2046

[14] N. Freed et N. Borenstein. MIME part five: conformance criteria and examples. *RFC 2049*, novembre 1996. www.imc.org/rfc2049

[15] S. Geva. Boosting the Performance of Nearest Neighbour Methods with Feature Selection. *Proceedings of the 5^{th} Pacific-Asia Conference on Knowledge Discovery and Data Mining*, Hong Kong, Chine, 2001, p. 210 - 221.

[16] E. S. Han, G. Karypis et V. Kumar. Text categorization using weight adjusted k-nearest neighbor classification. *Proceedings of the 5^{th} Pacific-Asia Conference on Knowledge Discovery and Data Mining*, Hong Kong, Chine, 2001, p. 53-65.

[17] J. B. Harris et C. Figg. Participating from the sidelines, online: facilitating telementoring projects. *ACM Journal of Computer Documentation*, vol. 24, no. 4, novembre 2000, p. 227 - 236.

[18] P. J. Hayes, P. M. Andersen, I. B. Nirenburg et L. M. Schmandt. TCS: a shell for content-based text categorization. *Proceedings of the 6th IEEE Conference on Artificial Intelligence Applications*, Santa Barbara, États-Unis, 1990, p. 320-326.

[19] L. S. Jensen et T. R. Martinez. Improving text classification by Using conceptual and contextual Features. *Proceedings of the sixth ACM SIGKDD international conference on Knowledge discovery and data mining*, Boston, États-Unis, 2000, p. 101 - 102.

[20] T. Joachims. A probabilistic analysis of the Rocchio algorithm with TFIDF for text categorization, *Proceedings of the 14th International Conference on Machine Learning*, Nashville, États-Unis, 1997, p. 143 - 151.

[21] L. Kosseim. *Système de réponse automatique : État de l'art.* Document Interne, Université de Montréal, 2000. www-rali.iro.umontreal.ca/Publications/etat2.pdf

[22] A. Krogh et J. Vedelsby. Neural network ensembles, cross validation, and active learning. *Advances in Neural Information Processing Systems 7*, MIT Press, 1995, p. 231 - 238.

[23] L. S. Larkey et W. B. Croft. Combining classifiers in text categorization. *Proceedings of the 19th annual international ACM SIGIR conference on Research and development in information retrieval*, Zurich, Suisse, 1996, p. 289 - 297.

[24] D. D. Lewis. An evaluation of phrasal and clustered representations on a text categorization task. *Proceedings of the 15th annual international ACM SIGIR conference on Research and development in information retrieval*, Copenhagen, Danemark, 1992, p. 37 - 50.

[25] G. F. Luger et W. A. Stubblefield. *Artificial intelligence: structures and strategies for complex problem solving*, Addison-Wesley, 1998, 824 p.

[26] M. L. Markus. Finding a happy medium: explaining the negative effects of electronic communication on social life at work. *ACM Transactions on Information Systems*, vol. 12, no. 2, avril 1994, p. 119 - 149.

[27] A. McCallum et K. Nigam. A comparison of event models for naive bayes text classification. *AAAI/ICML-98 Workshop on Learning for Text Categorization*, Madison, États-Unis, 1998, p. 41 - 48.

[28] C. J. Merz. Using correspondence analysis to combine classifiers. *Machine Learning*, vol. 36, no. 1-2, 1999, p. 33 - 58.

[29] K. Moore. MIME part three: message header extensions for non-ASCII text. *RFC 2047*, novembre 1996. www.imc.org/rfc2047

[30] D. Opitz et R. Maclin. Popular ensemble methods: an empirical study. *Journal of AI Research*, vol. 11, 1999, p. 169 - 198.

[31] R. M. Palloff et K. Pratt. Building learning communities in cyberspace: effective strategies for the online classroom. Jossey-Bass Publishers, 1999, 240 p.

[32] J. Palme. Common internet message headers. *RFC 2076*, février 1997. www.imc.org/rfc2076

[33] D. Popolov, M. Callaghan et P. Luker. Conversation space: visualising multi-threaded conversation. *Proceedings of the Working Conference on Advanced Visual Interfaces*, Palermo, Italie, 2000, p. 246 - 249.

[34] M. F. Porter. An algorithm of suffix stripping. *Program*, vol. 14, no. 3, juillet 1980, p. 130 - 137.

[35] J. R. Quinlan. Induction of decision trees. *Machine learning*, vol. 1, no. 1, p. 81 - 106.

[36] J. R. Quinlan. *C4.5: programming for machine learning* Morgan Kaufmann, 1993, 302 p.

[37] J. D. Rennie. *Improving Multi-class Text Classification with Naive Bayes* Master's thesis, Massachusetts Institute of Technology, 2001.

[38] P. Resnick. Internet message format. *RFC 2822*, avril 2001. www.imc.org/rfc2822

[39] C.J. van Rijsbergen. A theoretical basis for the use of co-occurrence data in information retrieval. *Journal of Documentation*, vol. 33, no. 2, 1977, p. 106 - 119.

[40] E. Riloff. Little Words Can Make a Big Difference for Text Classification, *Proceedings of the 18th annual international ACM SIGIR conference on Research and development in information retrieval*, Seattle, États-Uni, 1995, p .130 - 136.

[41] G. Salton et C. Buckley. Term-weighting approaches in automatic text retrieval. *Information Processing & Management*, vol. 24, no. 5, 1988, p. 513 - 523.

[42] R. E. Schapire. The strength of weak learnability. *Machine Learning*, vol. 5, no. 2, 1990, p. 197 - 227.

[43] L. Sproull et S. Kiesler. Reducing social context cues: electronic mail in organizational communication. *Management Science*, vol. 32, no. 11, novembre 1986, p. 1492 - 1512.

[44] V. Vapnik. *Estimation od dependences bsaed on empirical data*, Springer-Verlag, 1982.

[45] V. Vapnik. *The nature of statistical learning theory*, Springer, 1995, 304 p.

[46] S. S. Venkatesh, R. R. Snapp et D. Psaltis. Bellman strikes again!: the growth rate of sample complexity with dimension for the nearest neighbour classifier. *Proceedings of the 5th annual workshop on Computational learning theory*, Pittsburgh, États-Unis, 1992, p. 93 - 102.

[47] C. Welty. Backtracking: the demise of "!" *Intelligence*, vol. 12, no. 3, automne 2001, p. 56.

[48] S. Whittaker et C. Sidner. Email overload: exploring personal information management of email. *Conference proceedings on Human factors in computing systems*, Vancouver, Canada, 1996, p. 276 - 283.

[49] D. Wolpert. Stacked generalization. *Neural Networks*, vol. 5, p. 241 - 249.