Université de Paris Est

THÈSE

pour obtenir le grade de
Docteur de l'Université de Marne-La-Vallée

Spécialité : Informatique

présentée et soutenue publiquement le 15 décembre 2009 par

Chahnez ZAKARIA

Contributions à la détection de conflits relationnels dans les
échanges d'e-mails entre personnes en situation de travail
coopératif.

Une approche fondée sur les modèles statistiques et les ontologies

Directeurs de thèse
Olivier Curé et Gabriella Salzano

Jury

Président : Kamel Smaïli, LORIA, Nancy

Rapporteurs : Jean Claude Martin, LIMSI, Université Paris-Sud 11
Catherine Pelachaud, LTCI, TELECOM ParisTech

Examineurs : Olivier Curé (directeur), Université Paris-Est, Marne la Vallée
Gabrielle Salzano (co-directeur), Université Paris-Est, Marne la Vallée

Résumé

Dans ce mémoire de doctorat nous identifions un nouveau projet de recherche qui aborde le problème concret de la détection et de l'interprétation de conflits dans des documents échangés entre plusieurs personnes ou groupes de personnes. Nous considérons que la proposition d'une solution à cette problématique peut intéresser de nombreuses compagnies et organismes travaillant avec des outils de Travail Collaboratif Assisté par l'Ordinateur (TCAO). Par exemple, en disposant d'un tel outil, un chef de projet pourra détecter automatiquement les conflits émergents entre membres de son équipe et intervenir rapidement pour rétablir une situation propice à un travail de qualité. Ainsi, un effet direct de l'exploitation d'un tel outil sera d'améliorer la communication au sein du personnel, d'enrichir la gestion des ressources humaines et d'augmenter la productivité de l'entreprise.

L'approche adoptée dans nos travaux est pluridisciplinaire et implique les domaines suivants : psychologie et sociologie pour la gestion des émotions négatives, sémantique pour supporter l'interprétation du contenu des documents, linguistique pour la manipulation des corpus et statistique pour permettre la classification des documents en adéquation avec les conflits rencontrés.

Notre première étape de travail a consisté à définir un état de l'art de la recherche et des applications relatives à la détection d'émotions dans des documents textuels. L'absence de projet validé nous a conforté à définir une première approche dans cette direction de recherche. Une problématique importante implique l'interprétation, au travers d'un accès à une forme de sémantique, du contenu des documents échangés. Nous sommes convaincus que les technologies relatives aux ontologies correspondent à une réponse adaptée à cette situation. Notre seconde démarche a porté sur la recherche de l'existence d'une ontologie dans ce domaine. Nos recherches ne nous ont pas permis de trouver une ontologie répondant à notre attente et nous nous sommes donc inspirés de travaux en psychologie et sociologie pour concevoir une première version de notre taxinomie en élargissant la notion de conflit aux émotions négatives. Nous avons donc étudié différentes taxinomies d'émotions pour construire une première hiérarchie de notre ontologie de conflits relationnels. Ensuite nous avons procédé à un enrichissement de celle-ci en extrayant, depuis différents corpus, des termes discriminants et relatifs aux conflits. Cette étape de découverte a nécessité des traitements linguistiques des corpus ainsi que l'exploitation du modèle des triggers. L'objectif de notre ontologie consiste à caractériser un e-mail comme comportant (ou non) un (ou des) conflit(s). Pour cela notre approche détecte le type de conflit exprimé dans le document, en le classifiant selon les concepts de l'ontologie. Pour nos expérimentations, cette association entre concept de l'ontologie et document est réalisé à l'aide d'un ensemble de classifieurs, à savoir : les SVM (Support Vector Machine), le classifieur TFIDF (Term Frequency - Inverse Document Frequency), les arbres de décision, le Naïve Bayes, les K plus proches voisins et les réseaux de neurones. Dans cette thèse, nous nous concentrons sur des documents textuels et en particulier nous nous intéressons au contexte des emails, la méthode la plus populaire d'échange de documents et largement intégrée dans les outils de TCAO.

Mots-clés: Travail coopératif assisté par ordinateur, e-mails, émotions négatives, conflits relationnels, ontologies, classification textuelle.

Abstract

In this PhD dissertation we identify a novel research project that tackles the concrete problem of detecting and interpreting conflicts in documents exchanged between several persons or groups of persons. We consider that proposing a solution to this problem may interest many companies and organisations that make an intensive use of Computer Supported Cooperative Work (CSCW) tools. For instance, with such a system, project leaders could automatically detect conflicts emerging between team members and would enable them to rapidly react according to the gravity of the conflict, for instance by restoring a non-conflicting environment. Hence a direct effect of using such a tool would be to improve the communication between human agents and to enhance the productivity of the company.

The approach we have adopted in this work is pluridisciplinary and involves the following domains : psychology and sociology to manage the role of negative emotions, semantic related technologies to handle the interpretation of the content of the documents, linguistic to manipulate large corpora and statistical aspects to classify the documents according to the precise conflicts encountered.

The first major step of our work involves the definition of a state of the art related to emotion detection in textual documents. The absence of validated project motivated us to define a first approach. An important issue of this project implies the interpretation of the content of exchanged documents via an access to a form of semantics. We are convinced that ontologies are the ideal representation paradigm to represent a particular aspect, e.g. conflicts, of knowledge conveyed by these documents. Another step of our work was to search if ontologies exist in the domain of conflicts. According to our needs, we did not find any satisfactory ontologies for this domain and thus we studied works in psychology and sociology to define a first version taxonomy of conflicts at large, that is negative emotions. This taxonomy was not containing sufficient knowledge, in order to classify the documents with precision, so it was necessary to enrich it with new properties and concepts. We achieved this goal using an approach that retrieves discriminating, conflict related terms encountered in several text corpora. The discovery of these terms requires linguistic treatments of the documents as well as detecting triggers of terms.

Finally, the goal of our enriched ontology consists in characterising a document comprising a set of conflicts. This operation is performed by associating a set of ontology concepts with a given document. This association is performed by using the following classifiers : SVM (Support Vector Machine), TF-IDF (Term Frequency - Inverse Document Frequency), decision trees, naïve bayes solutions and neural networks. In this work, we concentrate on textual documents and in particular we are interested in the context of the most popular method of exchanging messages : emails.

Keywords: Computer supported cooperative work, emails, negative emotions, relational conflicts, ontologies, text classification.

Table des matières

Table des figures	vii
Liste des tableaux	ix
Introduction Générale	1
Chapitre 1 Contexte psychologique, sociologique et technologique du projet de recherche	5
1.1 La perception de l'homme dans l'organisation	6
1.1.1 L'école classique	6
1.1.2 L'école des relations humaines et les expériences de Hawthorne	8
1.2 Conséquences des conflits relationnels au sein de l'entreprise	9
1.3 Le Travail Coopératif Assisté par Ordinateur	11
1.3.1 L'aspect pluridisciplinaire du domaine de TCAO	11
1.3.2 Les taxinomies d'outils de TCAO	12
1.4 Les phénomènes linguistiques de l'e-mail : les nouvelles formes de communication écrite	13
1.4.1 Les erreurs d'orthographe	14
1.4.2 Néologisme	14
1.4.3 Néographie	14
1.4.4 Didascalies électroniques	15
1.5 Synthèse	15
Chapitre 2 Les ontologies comme support du traitement sémantique	17
2.1 Introduction à la notion d'ontologie	18
2.1.1 Présentation générale	18
2.1.2 Conception manuelle ou classique d'ontologie	19
2.1.3 Conception d'ontologie à partir du texte	20
2.1.4 Techniques de mise à jour d'ontologies	21
2.2 Un inventaire des vocabulaires et taxinomies d'émotions	23

2.3	Les ontologies d'émotions pour la classification des données	25
2.3.1	Des exemples d'ontologies d'émotions pour la classification de la parole et des images	25
2.3.2	Des exemples d'ontologies d'émotions pour la classification textuelle	28
2.4	Synthèse	29
Chapitre 3 La classification textuelle		30
3.1	Définition de la classification textuelle	30
3.2	Corpus et évaluation des modèles de classification	31
3.2.1	Corpus d'apprentissage	31
3.2.2	Corpus de développement	31
3.2.3	Corpus de test	31
3.2.4	Les mesures d'évaluation des modèles de classification	32
3.3	La représentation textuelle	33
3.3.1	La représentation en " <i>sac de mots</i> "	33
3.3.2	La représentation en signal numérique	35
3.4	La construction du vocabulaire et la réduction de dimension	35
3.4.1	L'analyse linguistique	36
3.4.2	L'analyse statistique	39
3.5	La classification textuelle	40
3.5.1	Le classifieur TFIDF	40
3.5.2	Les réseaux de neurones	41
3.5.3	Les arbres de décision	42
3.5.4	Les k plus proches voisins	43
3.5.5	Le classifieur Naïve Bayes	44
3.5.6	Les SVM	45
3.6	Synthèse	47
Chapitre 4 La conception d'une ontologie de conflits relationnels		48
4.1	Schéma et description générale de notre approche de détection des conflits relationnels	49
4.2	Présentation des corpus utilisés	52
4.2.1	Le corpus d'e-mails	52
4.2.2	Le corpus du journal "Le Monde"	53
4.2.3	Le corpus de forums	54
4.3	Prétraitements des corpus textuels et construction du vocabulaire	55
4.3.1	Les entités nommées	55

4.3.2	Les marques d'emails	55
4.3.3	Les erreurs d'orthographe	57
4.3.4	La néographie	57
4.3.5	Les binettes	57
4.3.6	La lemmatisation	58
4.3.7	Les mots outils	58
4.3.8	Translittération et les phénomènes liés à l'origine et la culture	59
4.3.9	Construction du vocabulaire	60
4.4	Construction de l'ontologie de conflits relationnels	63
4.4.1	Détermination de l'objectif opérationnel de l'ontologie	64
4.4.2	Construction de la hiérarchie de l'ontologie de conflits relationnels .	65
4.4.3	Collecte et traitement des corpus recouvrant le domaine de conflits relationnels	69
4.4.4	Extraction automatique de Triggers des corpus et leur intégration à l'ontologie	69
4.4.5	Mise à jour de l'ontologie	72
4.5	Synthèse	72
Chapitre 5 Détection des conflits relationnels : classification des e-mails		73
5.1	La résolution de la détection des conflits relationnels dans les e-mails	73
5.2	Le scénario des expérimentations	74
5.3	Le classifieur TFIDF	76
5.4	Les SVM et M-SVM	79
5.5	Les arbres de décision	80
5.6	Le classifieur Naïve Bayes	82
5.7	Les K plus proches voisins	83
5.8	Les réseaux de neurones	85
5.9	Synthèse	86
Conclusion et Perspectives		89
Bibliographie		93
Annexes		105
Annexe A Exemples de concepts émotions avant et après de l'ajout des triggers		105
A.1	Concept émotion "Colère"	106
A.2	Concept émotion "Haine"	107

A.3	Concept émotion "Trahison"	107
A.4	Concept émotion "Jalousie"	108
A.5	Concept émotion "Peur"	109
A.6	Concept émotion "Dévalorisation"	109
A.7	Concept émotion "Honte"	110
Annexe B Matrices de confusion de la classification avant et après de l'ajout des triggers		111
B.1	Arbre de décision	112
B.2	Naïve Bayes	113
B.3	Les K plus proches voisins	114
B.4	Réseau de neurones	116
Annexe C Article CoopIS'09		118

Table des figures

1.1	Répartition des participants selon les facteurs liés au travail [168]	9
1.2	Issue du processus de réinsertion professionnelle selon les facteurs liés au travail [168]	10
2.1	Rapport entre l'émergence des émotions avec les systèmes de traitement de l'information humaine [106]	26
2.2	Le modèle activation-évaluation d'émotion de Whissel [62]	27
3.1	La représentation vectorielle d'un document	36
3.2	Un exemple de texte lemmatisé avec TreeTagger	37
3.3	L'architecture d'un réseau perceptron multicouche [147]	42
3.4	Hyperplan avec faible marge vs. Meilleur hyperplan séparateur [70]	45
4.1	Notre approche de détection d'e-mails de conflits relationnels	51
4.2	Taux d'erreur en apprentissage et test, selon la taille du corpus [173]	54
4.3	Quelques exemples de micro publicité	56
4.4	Un exemple d'entête d'un e-mail	57
4.5	émotions négatives importantes vs. émotions négatives semi-importantes	65
4.6	Un extrait de la taxinomie de Damasio	66
4.7	Un extrait de la taxinomie de Lariévy	66
4.8	Un extrait de la hiérarchie de l'ontologie de conflits relationnels	67
4.9	Un extrait du concept " Triste "	71
4.10	L'enrichissement du concept " Triste " par les termes déclenchés	71
5.1	Taille du corpus des " forums " par concept émotion	76
5.2	Nombre de paragraphes détectés dans le corpus " Le Monde " par concept émotion en utilisant l'ontologie sans les triggers	78
5.3	Nombre de paragraphes détectés dans le corpus " Le Monde " par concept émotion en utilisant l'ontologie avec les triggers	78
A.1	Extrait du concept " Colère "	106
A.2	Extrait du concept " Colère " après l'ajout des termes déclenchés	106
A.3	Extrait du concept " Haine " avant l'ajout des termes déclenchés	107
A.4	Extrait du concept " Haine " après l'ajout des termes déclenchés	107
A.5	Extrait du concept " Trahison " avant l'ajout des termes déclenchés	107
A.6	Extrait du concept " Trahison " après l'ajout des termes déclenchés	108
A.7	Extrait du concept " Jalousie " avant l'ajout des termes déclenchés	108
A.8	Extrait du concept " Jalousie " après l'ajout des termes déclenchés	108

A.9	Extrait du concept " Peur " avant l'ajout des termes déclanchés	109
A.10	Extrait du concept " Peur " après l'ajout des termes déclanchés	109
A.11	Extrait du concept " Dévalorisation " avant l'ajout des termes déclanchés .	109
A.12	Extrait du concept " Dévalorisation " après l'ajout des termes déclanchés .	110
A.13	Extrait du concept " Honte " avant l'ajout des termes déclanchés	110
A.14	Extrait du concept " Honte " après l'ajout des termes déclanchés	110

Liste des tableaux

2.1	Les classes d'émotions [114]	28
3.1	L'indexation des documents	33
4.1	Le corpus d'e-mails	53
4.2	Le corpus "Le Monde"	53
4.3	Le corpus de forums	54
4.4	Un extrait de la liste des mots outils	59
4.5	Le corpus d'e-mails	60
4.6	Le corpus d'e-mails avant suppression des noms propres	60
4.7	Le corpus d'e-mails après suppression des noms propres	60
4.8	Le corpus "Le Monde"	61
4.9	Le corpus "Le Monde" avant suppression des noms propres	61
4.10	Le corpus "Le Monde" après suppression des noms propres	61
4.11	Le corpus de forums	61
4.12	Le corpus de forums avant suppression des noms propres	61
4.13	Le corpus de forums après suppression des noms propres	61
4.14	Exemples de mots de faible occurrence et retenus comme pouvant indiquer la présence de conflits relationnels	62
5.1	Les résultats du classifieur TFIDF en utilisant l'ontologie avec les triggers pour le corpus "Le Monde"	77
5.2	Les résultats du classifieur TFIDF en utilisant l'ontologie sans et avec les triggers pour le corpus "Forums"	79
5.3	Les résultats du classifieur M-SVM en utilisant l'ontologie sans et avec les triggers pour le corpus "Forums"	80
5.4	Les résultats du classifieur M-SVM en utilisant l'ontologie avec les triggers pour le corpus "Le Monde"	80
5.5	Les résultats du classifieur arbre de décision en utilisant l'ontologie sans et avec les triggers pour le corpus "Forums"	81
5.6	Les résultats du classifieur arbre de décision en utilisant l'ontologie sans les triggers pour le corpus "Le Monde"	82
5.7	Les résultats du classifieur arbre de décision en utilisant l'ontologie avec les triggers pour le corpus "Le Monde"	82
5.8	Les résultats du classifieur naïve bayes en utilisant l'ontologie sans les triggers pour le corpus "Le Monde"	82
5.9	Les résultats du classifieur naïve bayes en utilisant l'ontologie avec les triggers pour le corpus "Le Monde"	83

5.10	Les résultats du classifieur naïve bayes en utilisant l'ontologie sans et avec les triggers pour le corpus "Forums"	83
5.11	Les résultats du classifieur KNN en utilisant l'ontologie sans les triggers pour le corpus "Le Monde"	84
5.12	Les résultats du classifieur KNN en utilisant l'ontologie avec les triggers pour le corpus "Le Monde"	84
5.13	Les résultats du classifieur KNN en utilisant l'ontologie sans et avec les triggers pour le corpus "Forums"	84
5.14	Les résultats du classifieur réseau de neurones en utilisant l'ontologie sans les triggers pour le corpus "Le Monde"	85
5.15	Les résultats du classifieur réseau de neurones en utilisant l'ontologie avec les triggers pour le corpus "Le Monde"	85
5.16	Les résultats du classifieur réseau de neurones en utilisant l'ontologie sans et avec les triggers pour le corpus "Forums"	86
5.17	Matrice de confusion du réseau de neurones avec les triggers pour le corpus "Forums"	86
5.18	Synthèse de classification des messages du corpus "Forum"	87
B.1	Matrice de confusion du classifieur arbre de décision sans les triggers pour le corpus " Forums "	112
B.2	Matrice de confusion du classifieur arbre de décision avec les triggers pour le corpus " Forums "	112
B.3	Matrice de confusion du classifieur arbre de décision sans les triggers pour le corpus " Le Monde "	112
B.4	Matrice de confusion du classifieur arbre de décision avec les triggers pour le corpus " Le Monde "	113
B.5	Matrice de confusion du classifieur Naïve Bayes sans les triggers pour le corpus " Forums "	113
B.6	Matrice de confusion du classifieur Naïve Bayes avec les triggers pour le corpus " Forums "	113
B.7	Matrice de confusion du classifieur Naïve Bayes sans les triggers pour le corpus " Le Monde "	114
B.8	Matrice de confusion du classifieur Naïve Bayes avec les triggers pour le corpus " Le Monde "	114
B.9	Matrice de confusion du classifieur KNN sans les triggers pour le corpus " Forums "	114
B.10	Matrice de confusion du classifieur KNN avec les triggers pour le corpus " Forums "	115
B.11	Matrice de confusion du classifieur KNN sans les triggers pour le corpus " Le Monde "	115
B.12	Matrice de confusion du classifieur KNN avec les triggers pour le corpus " Le Monde "	115
B.13	Matrice de confusion du classifieur réseau de neurones sans les triggers pour le corpus "Forums"	116
B.14	Matrice de confusion du classifieur réseau de neurones avec les triggers pour le corpus "Forums"	116
B.15	Matrice de confusion du classifieur réseau de neurones sans les triggers pour le corpus "Le Monde"	116

B.16 Matrice de confusion du classifieur réseau de neurones avec les triggers pour le corpus "Le Monde"	117
---	-----