ACTES DE L'ATELIER

DATA MINING, APPLICATIONS, CAS D'ÉTUDES ET SUCCESS STORIES (ACE 2011)

en conjonction avec

11ème Conférence Internationale Francophone
sur l'Extraction et la Gestion des Connaissances
(EGC 2011)

25 janvier 2011, Brest, FRANCE

Organisé par :

Dominique Gay (Orange Labs Networks and Carriers, Lannion, France)
Hakim Hacid (Alcatel-Lucent Bell Labs, Paris, France)

Préface

Le data mining et l'apprentissage, au niveau national, sont principalement focalisés sur le point de vue théorique. Les différents travaux et initiatives le démontrent. Il est certain que le data mining a pris beaucoup de sens dans les pays anglo-saxons où il a trouvé de l'intérêt dans plusieurs domaines comme les communications, la bourse, le commerce, la médecine, etc. Ces dernières années, l'industrie nationale commence à s'intéresser de prêt à ce domaine et voit en lui une opportunité unique pour se distinguer sur le marché en investissant dans l'innovation.

Afin de permettre à l'industrie nationale de mieux intégrer ce domaine dans leur chaine d'innovation et leurs lignes de produits, il est nécessaire à la communauté académique de travailler main dans la main avec les industriels pour résoudre leurs problèmes concrets et de prouver que ce domaine pourrait régler certaines problématiques qui les intéresse directement pour enrichir leurs portefeuilles. Il est à noter que la thématique de cet atelier n'est pas extraordinairement nouvelle montrant déjà la conscience du monde académique et industriel sur la nécessité de travailler ensemble. A titre d'exemple, la conférence EGC intègre déjà un aspect applicatif dans les thématiques de la conférence.

Cette édition est la première d'une série que nous espérons longue et bénéfique pour la communauté. L'objectif est de réunir des travaux pouvant être : *(i)* des applications industrielles ou concrètes et récentes du data mining, *(ii)* des nouvelles techniques pouvant être directement à des problèmes industriels, *(iii)* des cas d'études et success stories. La liste des domaines d'application du data mining couvre (mais n'est pas restreinte à) :

- Processus industriels du data mining et du machine learning ;
- Planification des ressources industrielles ou d'entreprises ;
- Surveillance/Contrôle/Monitoring des réseaux ;
- Customers Relationship Management (CRM) ;
- Systèmes de recommandation ;
- Marketing et publicité (en ligne ou sur nouveaux supports) ;
- E-commerce, Vente à distance, Vente par correspondance ;
- Analyse d'audience ;
- Analyse de traces de site/logiciel ;
- Analyse de données bancaires, financières ;
- Recherche d'informations ;
- Moteurs de recherche.

Les organisateurs de l'atelier ACE - EGC 2011
Dominique Gay, Orange Labs, France
Hakim Hacid, Alcatel-Lucent Bell Labs, France

Comités

Comité d'organisation

- Dominique Gay (Orange Labs Networks and Carriers, France)
- Hakim Hacid (Alcatel-Lucent Bell Labs, France)

Comité de programme

- Nicolas Becourt (Alcatel-Lucent, France)
- Younès Bennani (LIPN, Université Paris 13, France)
- Jérôme Besombes (ONERA, France)
- Marc Boullé (Orange Labs, France)
- Fabrice Clérot (Orange Labs, France)
- Sylvie Galichet (LISTIC, Université de Savoie, France)
- Gregory Grefenstette (Exalead, France)
- Brigitte Hoeltzener (E3I2, ENSIETA, France)
- Stéphane Lorin (CeNTAI, Thales Communications, France)
- François-Xavier Jollois (LIPADE, Université Paris Descartes, France)
- Vincent Lemaire (Orange Labs, France)
- Nicolas Méger (LISTIC, Université de Savoie, France)
- Gilbert Saporta (CEDRIC, CNAM, France)
- Maguelonne Teisseire (CEMAGREF, France)
- AbdelMalek Toumi (E3I2, ENSIETA, France)
- Tanguy Urvoy (Orange Labs, France)

Remerciements

Les organisateurs de l'atelier souhaitent remercier vivement toutes les personnes ayant contribué à la tenue de cet atelier. En particulier :

- les membres du comité de programme dont le travail d'évaluation était crucial pour assurer la qualité de l'atelier ;
- les auteurs pour la qualité de leurs contributions constituant la base essentielle de discussions fructueuses ;
- les deux conférenciers invités, Pr. Philippe Lenca et Pr. Stéphane Lallich pour avoir accepté de partager leur expérience dans le cadre de cet atelier ;
- les organisateurs d'EGC qui ont mis en place l'environnement et les moyens pour la réussite des ateliers.

Table des matières

Première partie

Conférence invitée

Chapitre 1

Le choix d'une bonne mesure de qualité, condition du succès d'un processus de fouille de données

Le choix d'une bonne mesure de qualité, condition du succès de la fouille de données

Philippe Lenca[*,***], Stéphane Lallich[**]

[*]Institut Télécom, Télécom Bretagne
UMR CNRS 3192 Lab-STICC
philippe.lenca@telecom-bretagne.eu
[**]Université Lyon, Lyon 2
Laboratoire ERIC
stephane.lallich@univ-lyon2.fr
[***]Université européenne de Bretagne, France

Notre réflexion se situe dans le domaine de l'apprentissage supervisé ou non supervisé par induction de règles. La fouille de données est couronnée de succès lorsque l'on parvient à extraire des données des connaissances nouvelles, valides, exploitables, etc. (Fayyad et al. (1996) Kodratoff et al. (2001)). L'une des clefs du succès est, bien sûr, le choix d'un algorithme qui soit bien adapté aux caractéristiques des données et au type de connaissances souhaitées : par exemple les règles d'association en non supervisé ; les arbres de décision, les règles d'association de classe et le bayésien naïf, en supervisé. Cependant, le succès dépend d'autres facteurs, notamment la préparation des données (représentation des données, *outliers*, variables redondantes) et le choix d'une bonne mesure d'évaluation de la qualité des connaissances extraites, tant dans le déroulement de l'algorithme que dans l'évaluation finale des résultats obtenus. C'est de ce dernier facteur que nous allons parler.

En introduction, nous évoquerons rapidement le problème de la représentation des données. Puis, après avoir rappelé le principe de la recherche des règles d'association (Agrawal et Srikant (1994)) ou des règles d'association de classe intéressantes (Liu et al. (1998)), nous montrerons, à partir de quelques exemples, la diversité des résultats obtenus suivant la mesure d'intérêt choisie, que ce soit en comparant les pré-ordres obtenus ou en calculant les meilleures règles (Vaillant et al., 2004). Ces exemples illustrent le fait qu'il n'y a pas de mesure qui soit intrinsèquement bonne, mais différentes mesures qui, suivant leurs propriétés, sont plus ou moins bien adaptées au but poursuivi par l'utilisateur. Une mesure favorise tel ou tel type de connaissance, ce qui constitue un biais d'apprentissage que nous illustrerons par la mesure de Jaccard (Plasse et al. (2007)) .

Nous proposerons ensuite une synthèse des travaux concernant les mesures de qualité des règles d'association en présentant les principaux critères d'évaluation des mesures et en montrant concrètement le rôle de chacun de ces critères dans le comportement des mesures (e.g. Lenca et al. (2003), Tan et al. (2004), Geng et Hamilton (2006), Lenca et al. (2008), Suzuki (2008), Guillaume et al. (2010), Lerman et Guillaume (2010), Gras et Couturier (2010) ; nous renvoyons également le lecteur aux ouvrages édités par Guillet et Hamilton (2007) et Zhao et al. (2009)). Nous illustrerons le lien qui existe entre les propriétés des mesures sur les critères retenus et leur comportement sur un certain nombre de bases de règles (Vaillant et al., 2004).

A côté de ces critères qui permettent d'étalonner les propriétés des mesures, nous présenterons d'autres critères de choix très importants. En premier lieu, nous nous intéresserons aux propriétés algorithmiques des mesures afin de pouvoir extraire les motifs intéressants en travaillant directement sur la mesure considérée, sans fixer de seuil de support, ce qui permet d'accéder aux pépites de connaissances (Wang et al. (2001), Xiong et al. (2003), Li (2006), Le Bras et al. (2009), Le Bras et al. (2009), Le Bras et al. (2010)). Nous exhiberons des conditions algébriques sur la formule d'une mesure qui assurent de pouvoir associer un critère d'élagage à la mesure considérée. Nous nous poserons ensuite le problème de l'évaluation de la robustesse des règles suivant la mesure utilisée (Azé et Kodratoff (2002), Cadot (2005), Gras et al. (2007), Le Bras et al. (2010)).

Enfin, nous traiterons le cas des données déséquilibrées (Weiss et Provost (2003)) en apprentissage par arbres (Chawla (2003)) et nous montrerons comment le choix d'une mesure appropriée permet d'apporter une solution algorithmique à ce problème qui améliore de façon significative à la fois le taux d'erreur global, la précision et le rappel (Zighed et al. (2007), Lenca et al. (2008)). Si l'on veut privilégier la classe minoritaire, cette solution peut être encore améliorée en introduisant, dans la procédure d'affectation des étiquettes opérant sur chaque feuille de l'arbre, une mesure d'intérêt adéquate qui se substitue à la règle majoritaire (Ritschard et al. (2007), Pham et al. (2008)). Une discussion sur les mesures de qualité de bases de règles est présentée dans (Holena, 2009).

En définitive, comment aider l'utilisateur à choisir la mesure la plus appropriée à son projet ? Nous proposerons une procédure d'assistance au choix de l'utilisateur qui permet de retourner à celui-ci les mesures les plus appropriées, une fois qu'il a défini les propriétés qu'il attend d'une mesure (Lenca et al. (2008)).

Références

Agrawal, R. et R. Srikant (1994). Fast algorithms for mining association rules. In *VLDB*, pp. 487–499.

Azé, J. et Y. Kodratoff (2002). Evaluation de la résistance au bruit de quelques mesures d'extraction de règles d'association. In *EGC*, pp. 143–154.

Cadot, M. (2005). A simulation technique for extracting robust association rules. In *CSDA*, pp. 143–154.

Chawla, N. (2003). C4.5 and imbalanced datasets : Investigating the effect of sampling method, probalistic estimate, and decision tree structure. In *ICM Workshop on Learning from Imbalanced Data Sets*.

Fayyad, U., G. Piatetsky-Shapiro, P. Smyth, et R. Uthurusamy (Eds.) (1996). *Advances in Knowledge Discovery and Data Mining*. AAAI/MIT Press.

Geng, L. et H. J. Hamilton (2006). Interestingness measures for data mining: A survey. *ACM* (3, Article 9).

Gras, R. et R. Couturier (2010). Spécificité de l'A.S.I. par rapport à d'autres mesures de qualité de règles d'association. In *Analyse Statistique implicative*, pp. 175–198.

Gras, R., J. David, F. Guillet, et H. Briand (2007). Stabilité en A.S.I. de l'intensité d'implication et comparaisons avec d'autres indices de qualité de règles d'association. In

Qualité des Données et des Connaissances, pp. 35–43.

Guillaume, S., D. Grissa, et E. M. Nguifo (2010). Propriété des mesures d'intérêt pour l'extraction des règles. In *Qualité des Données et des Connaissances*, pp. 15–28.

Guillet, F. et H. J. Hamilton (Eds.) (2007). *Quality Measures in Data Mining*. Springer.

Holena, M. (2009). Measures of ruleset quality for general rules extraction methods. *Int. J. Approx. Reasoning* (6), 867–879.

Kodratoff, Y., A. Napoli, et D. Zighed (2001). Bulletin de l'Association Française d'Intelligence Artificielle, Extraction de connaissances dans des bases de données.

Le Bras, Y., P. Lenca, et S. Lallich (2009). On optimal rules mining: a framework and a necessary and sufficient condition for optimality. In *PAKDD*, Volume 5476 of *Lecture Notes in Computer Science*, pp. 705–712. Springer-Verlag Berlin Heidelberg.

Le Bras, Y., P. Lenca, et S. Lallich (2010). Mining interesting rules without support requirement: A general universal existential upward closure property. *Annals of Information Systems 8*, 75–98.

Le Bras, Y., P. Lenca, S. Moga, et S. Lallich (2009). On the generalization of the all-confidence property. In *ICMLA*, pp. 759–764. IEEE Press.

Le Bras, Y., P. Meyer, P. Lenca, et S. Lallich (2010). A robustness measure of association rules. In *ECML/PKDD*, Volume 6322 of *Lecture Notes in Computer Science*, pp. 227–242. Springer-Verlag Berlin Heidelberg.

Lenca, P., S. Lallich, T.-N. Do, et N.-K. Pham (2008). A comparison of different off-centered entropies to deal with class imbalance for decision trees. In *PAKDD*, Volume 5012 of *Lecture Notes in Computer Science*, pp. 634–643.

Lenca, P., P. Meyer, P. Picouet, B. Vaillant, et S. Lallich (2003). Critères d'évaluation des mesures de qualité en ECD. *Revue des Nouvelles Technologies de l'Information* (RNTI-1), 123–134.

Lenca, P., P. Meyer, B. Vaillant, et S. Lallich (2008). On selecting interestingness measures for association rules: user oriented description and multiple criteria decision aid. *EJOR* (2), 610–626.

Lerman, I.-C. et S. Guillaume (2010). Analyse comparative d'indices d'implication discriminants fondés sur une échelle de probabilité. Technical Report IRISA 1942/INRIA 7187.

Li, J. (2006). On optimal rule discovery. *TKDE 18*(4), 460–471.

Liu, B., W. Hsu, et Y. Ma (1998). Integrating classification and association rule mining. In *Knowledge Discovery and Data Mining*, pp. 80–86.

Pham, N.-K., T.-N. Do, P. Lenca, et S. Lallich (2008). Using local node information in decision trees: Coupling a local decision rule with an off-centered entropy. In *DMIN*, Volume 1, pp. 117–123.

Plasse, M., N. Niang, G. Saporta, A. Villeminot, et L. Leblond (2007). Combined use of association rules mining and clustering methods to find relevant links between binary rare attributes in a large data set. *Computational Statistics & Data Analysis 52*(1), 596–613.

Ritschard, G., D. A. Zighed, et S. Marcellin (2007). Données déséquilibrées, entropie décentrée et indice d'implication. In *Analyse Statistique implicative*, pp. 315–327.

Suzuki, E. (2008). Pitfalls for categorizations of objective interestingness measures for rule discovery. In *Statistical Implicative Analysis, Theory and Applications*, pp. 383–395.

Tan, P.-N., V. Kumar, et J. Srivastava (2004). Selecting the right objective measure for association analysis. *IS* (29), 293–313.

Vaillant, B., P. Lenca, et S. Lallich (2004). A clustering of interestingness measures. In *DS*, pp. 290–297.

Wang, K., Y. He, et D. W. Cheung (2001). Mining confident rules without support requirement. In *CIKM*, pp. 89–96. ACM.

Weiss, G. M. et F. Provost (2003). Learning when training data are costly: The effect of class distribution on tree induction. *J. of Art. Int. Research 19*, 315–354.

Xiong, H., P.-N. Tan, et V. Kumar (2003). Mining strong affinity association patterns in data sets with skewed support distribution. In *ICDM*, pp. 387–394.

Zhao, Y., C. Zhang, et L. Cao (Eds.) (2009). *Post-Mining of Association Rules: Techniques for Effective Knowledge Extraction*. IGI Global.

Zighed, D. A., S. Marcellin, et G. Ritschard (2007). Mesure d'entropie asymétrique et consistante. In *EGC*, pp. 81–86.

Deuxième partie

Articles ACE

Chapitre 2

Etude des résultats des systèmes de RI à grande échelle

Etude des résultats des systèmes de RI à grande échelle
Sélection des mesures de performance pour l'évaluation

Sébastien Déjean*, Josiane Mothe**, Julia Poirier***,
Benoît Sansas***, Joelson Randriamparany **,

*IMT, UMR, Université de Toulouse
**IRIT, URM5505, Université de Toulouse
***INSA, Université de Toulouse

Résumé. Cet article présente un usage des méthodes d'analyse de données dans le domaine de la recherche d'information. Plus spécifiquement, nous analysons l'ensemble des données issues de la campagne d'évaluation TREC et montrons que les mesures utilisées pour évaluer les systèmes peuvent se réduire à 7 au lieu des 128 qui sont intégrées dans le programme d'évaluation trec_eval.

1 Introduction

Les moteurs de recherche d'information (RI) visent à retrouver l'information pertinente par rapport à un besoin formulé sous forme de requête d'un utilisateur. Pour permettre cette RI, les systèmes se basent sur plusieurs principes fondamentaux dont l'indexation, qui vise à représenter chaque document à l'aide de termes d'indexation et l'appariement de la requête avec chacun des documents. Ainsi, les différents moteurs de recherche diffèrent par les mécanismes mettant en œuvre l'indexation d'une part et l'appariement d'autre part.

Les performances d'une chaine de traitement sont évaluées lors de campagnes internationales comme Text Retrieval Conference (TREC). Pour une tâche donnée, TREC propose l'ensemble des ressources nécessaires aux participants pour tester leurs systèmes. Ainsi, par exemple, la tâche TREC *adhoc* vise à évaluer les moteurs qui restituent une liste de documents à partir d'une requête. TREC fourni pour cela un ensemble de documents et un ensemble de besoins d'information. Les participants fournissent en retour la liste des documents que leurs systèmes retrouvent. L'évaluation est ensuite réalisée en considérant les réponses attendues (données par des acteurs humains) et des mesures de performance.

Actuellement, plusieurs centaines de requêtes de test sont disponibles et plusieurs dizaines de résultats provenant de variantes de chaines de traitement. Cette masse d'information n'est que très peu exploitée dans le domaine de la recherche d'information. Cet article vise à proposer quelques pistes d'analyse qui pourraient ouvrir de nouvelles voies dans le domaine de la recherche d'information. Les outils statistiques nous ont paru pertinents pour une analyse poussée de ces masses de données.

La suite de l'article est organisée comme suit : dans la section 2, nous présentons quelques travaux reliés. Dans la section 3, nous présentons les données que nous avons analysées. Dans la section 4 nous montrons qu'il est possible, grâce à l'analyse de données, de regrouper les mesures d'évaluation selon des classes homogènes. Dans la section 5, nous montrons que le classement d'un ensemble de systèmes est stable lorsque nous considérons 6 mesures au lieu des 128 initiales, que l'on se base sur le score ou sur le rang moyen des systèmes. Nous concluons ces travaux et présentons plusieurs pistes pour les prolonger.

2 Travaux reliés

Peu de travaux se sont intéressés à l'analyse des résultats issus des campagnes d'évaluation.

Banks *et al.* (1999) décrivent différentes analyses sur différentes données de TREC. En particulier, ils considèrent une matrice dans laquelle les lignes et les colonnes représentent les systèmes et les besoins d'information ; les cellules correspondent à la *précision moyenne* (moyenne des précisions à chaque fois qu'un document pertinent est retrouvé). Cette matrice est alors utilisée pour étudier les groupes qui pourraient en être extraits. Pour cela, les auteurs utilisent une classification hiérarchique par simple liage. Cette méthode combine les groupes qui minimisent la distance entre les éléments les plus proches. Les figures présentées dans l'article cachent la distance entre les groupes détectés ; de plus l'article ne discute pas l'arbre obtenu ni les groupes qui auraient été obtenus en coupant l'arbre à différents niveaux. Les auteurs concluent que l'analyse n'apporte pas d'information vraiment utilisable.

Une analyse différente est présentée par Mizzaro et Robertson (2007). L'objectif de leur étude est d'identifier un petit ensemble de requêtes qui pourrait être utilisé pour distinguer les systèmes efficaces des systèmes qui ne le sont pas. Les auteurs utilisent les mesures de performance sur les paires systèmes/requêtes. Les auteurs concluent que certains systèmes sont meilleurs pour distinguer les requêtes faciles de celles qui sont difficiles. Ils concluent également que les systèmes peu performants le sont quelque soit la requête alors que de bons systèmes peuvent échouer sur certaines requêtes. Enfin, les auteurs montrent que les requêtes les plus faciles sont les plus aptes à distinguer les systèmes par rapport à leurs performances.

3 Données et objectifs d'analyse

3.1 Matrice de données à analyser

Les données que nous utilisons sont obtenues à partir des résultats des participants à une tâche d'évaluation. Plus précisément, les données initiales sont les listes de documents retrouvées par chaque système participant à une tâche. Nous nous sommes basés sur la tâche *adhoc* de TREC. Cette tâche se compose chaque année de 50 requêtes à traiter ; le nombre de systèmes et les systèmes eux-mêmes varient d'une année sur l'autre. Par exemple, en 1999, 129 systèmes ont traité les requêtes. A partir de ces éléments, pour cette seule année, nous pouvons considérer un ensemble de 6450 individus à analyser (un individu pour chaque système et pour chaque requête). Pour chaque individu, nous calculons un ensemble de valeurs correspondant à des mesures de performance, ce seront les variables. Elles sont calculées par l'outil trec_eval (Buckley, 1991) et sont au nombre de 128. La majorité de ces mesures ont des valeurs ou scores compris entre 0 et 1.

Un extrait de la matrice qui nous sert de base dans l'analyse est présenté à la figure 1.

3.2 Objectif d'analyse

Nous nous sommes fixés deux objectifs pour l'analyse :

- Compte tenu du nombre important de mesures permettant d'évaluer la performance des systèmes, nous souhaitons vérifier le niveau de redondance et la complémentarité

de celles-ci. En effet, lorsqu'un nouveau système de RI est mis au point, la mesure de ses performances est une étape cruciale. Comparer le nouveau système aux autres systèmes existants sur 128 mesures s'avère fastidieux. A l'opposer, se limiter à la mesure *Mean Average Precision* (moyenne sur un ensemble de besoins d'information des *précisions moyennes*) risque de cacher une partie des performances. L'analyse des données pourrait permettre de déterminer un nombre optimal de mesures à considérer lors d'une évaluation, voire à les identifier (Baccini *et al.*, 2010).

- La mise au point de nouveaux modèles de RI ou l'adaptation de modèles existants visent à obtenir un « meilleur » système ; il s'agit donc de pouvoir comparer des systèmes entre eux. Nous étudions l'impact de l'utilisation d'un nombre réduit de mesures de performance sur les rangs obtenus par les systèmes.

campagnes	taches	systemes	requetes	0.20R-prec	0.40R-prec	0.60R-prec
TREC1999	adhoc	1	401	0.016700	0.008300	0.011100
TREC1999	adhoc	1	402	0.000000	0.000000	0.000000
TREC1999	adhoc	1	403	0.000000	0.000000	0.000000
TREC1999	adhoc	1	404	0.000000	0.000000	0.000000
[...]						
TREC1999	adhoc	weaver2	447	0.750000	0.571400	0.600000
TREC1999	adhoc	weaver2	448	0.000000	0.000000	0.035700
TREC1999	adhoc	weaver2	449	0.000000	0.000000	0.000000
TREC1999	adhoc	weaver2	450	0.762700	0.686400	0.596600

Figure 1 : Extrait de la matrice analysée

4 Redondance et complémentarité des mesures

4.1 Données et méthode d'analyse

Pour étudier la redondance dans les mesures de performance des systèmes de RI, nous nous sommes appuyés sur l'ensemble des données collectées pour la tâche adhoc de TREC (TREC-2 à TREC-8). Nous avons éliminé la première année qui a servi de mise au point de la définition de la tâche (TREC-1). Au total, nous obtenons une matrice composée de 23 518 individus. Les variables sont celles présentées dans la section 3 (mesures de performances obtenues par trec_eval).

D'un point de vue outils mathématiques, pour analyser ces données, nous nous sommes appuyés sur une **classification ascendante hiérarchique**. En effet, notre objectif est de regrouper les mesures de performance en groupes le plus homogènes possibles. N'ayant pas d'idée *a priori* sur le nombre de classes à obtenir, une telle classification, nous permet de choisir à postériori un ensemble de classes. Seber (1984) préconise, lorsque cela est possible de stabiliser ce type de classification en la faisant suivre d'une classification supervisée du type k-means. Lors de la classification hiérarchique, nous avons utilisé la mesure de Ward qui est la plus utilisée. Elle consiste à fusionner à chaque étape les groupes qui minimisent l'augmentation du total de la somme des carrés des distances dans les groupes.

4.2 Résultats d'analyse

4.2.1 Classes de mesures de performances des systèmes de RI

La figure 2 présente le dendrogramme résultant de la classification hiérarchique des mesures de performance des systèmes de RI. Le graphique en haut à droite de la figure représente la distance entre les nœuds. Ce graphique suggère une coupe pertinente en considérant 3 groupes ; une autre en considérant 5 groupes et une dernière en considérant 7 groupes (*cf.* la pente sur le graphique cité plus haut).

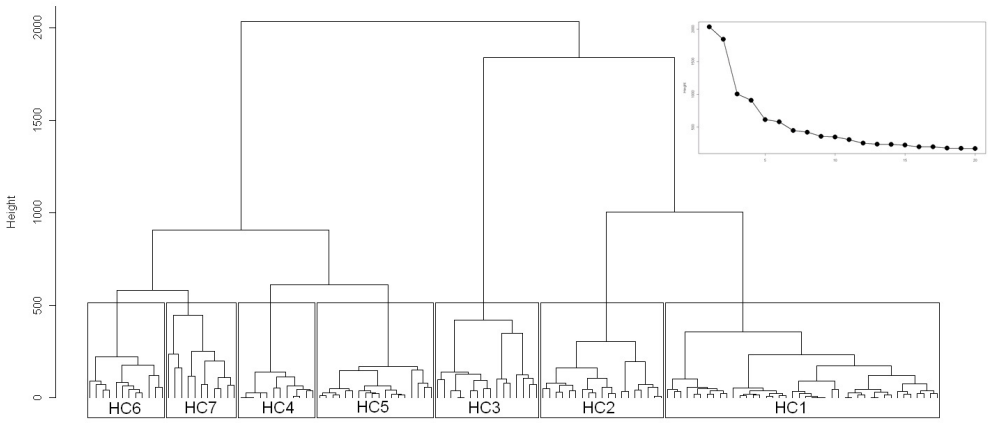


Figure 2. Dendrogramme représentant la classification hiérarchique des mesures de performance

L'application d'une classification k-means à partir des groupes ainsi obtenus à montré la bonne stabilité des groupes puisque seulement 13 mesures ont changé de groupe après son application. La classification supervisée a isolé un groupe de 3 mesures ne correspondant pas à des scores (*nombre de documents retrouvés*, *rang du premier document pertinent retrouvé* et *nombre de documents jugés non pertinents retrouvés).*

Au final, le nombre de groupes (7) correspond donc au nombre minimal de mesures permettant de couvrir les différents aspects mesurés par l'ensemble complet des mesures.

4.2.2 Homogénéité des groupes et représentant de groupe

Une fois les groupes déterminés, pour permettre leur utilisation, nous avons sélectionné un représentant de groupe : pour un groupe donné, il s'agit de la mesure qui devrait être étudiée pour évaluer le système de RI. Selon les résultats de l'analyse, utiliser une autre mesure du même groupe conduirait aux mêmes conclusions lors de la comparaison du système évalué avec d'autres. En effet, les mesures issues d'un même groupe sont trouvées comme redondantes.

Avant de choisir le représentant de chaque groupe, nous nous sommes assurés de leur homogénéité. Compte tenu de la forte homogénéité des groupes obtenus, n'importe quelle mesure pourrait servir de représentant de groupe. Cependant, nous avons arbitré en choisissant, soit la mesure du groupe la plus communément utilisée dans la littérature, soit la mesure la plus proche du centroïde. Le détail des groupes obtenus n'est pas présenté dans cet article.

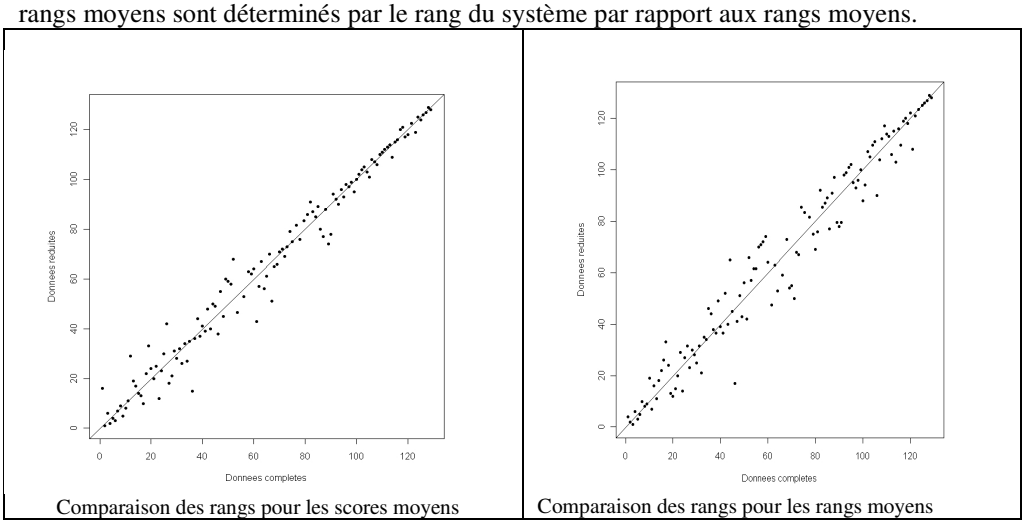
Les mesures représentatives des groupes ainsi que les caractéristiques des groupes sont les suivantes :

- MAP (Mean Average Precision) est un bon représentant du premier groupe. Ce groupe contient également les mesures permettant de réaliser les courbes de rappel/précision. La mesure MAP est connue pour permettre une représentation globale des performances.
- P10 (Précision lorsque 10 documents sont retrouvés) représente bien le groupe qui associe dans le résultat de l'analyse les mesures de hautes précisions.
- P100 représente bien le groupe qui associe les mesures de précision lorsque de grands ensembles de documents retrouvés sont considérés.
- Exact recall (Rappel exact, non interpolé) peut représenter le groupe correspondant aux mesures orientées rappel et qui ont été regroupées par l'analyse.
- rank first rel est le représentant des 3 mesures regroupées qui ne sont pas des scores.
- Recall 30 (rappel lorsque 30 documents sont retrouvés) et bpref topnonrel sont deux mesures que nous avons choisies pour représenter les deux groupes restants.

5 Classement des systèmes : effet de score et de rang

Généralement, la mise au point d'un système de RI implique la comparaison avec d'autres et donc un classement du système. Le principal objectif de l'analyse présentée dans cette section est de comparer les méthodes de classement des systèmes lorsque l'ensemble des mesures sont utilisées et lorsque l'ensemble réduit de mesures est choisi (Poirier et Sansas, 2009).

Nous définissons le **score moyen d'un système** par la moyenne des valeurs sur l'ensemble des mesures de performance pour chaque système. Le rang d'un système est déterminé par rapport aux scores moyens après avoir ordonné les systèmes par ordre croissant des scores moyens. Le **rang moyen d'un système** sur l'ensemble des mesures correspond à la moyenne des rangs obtenu pour chaque mesure. Les systèmes qui ont les meilleurs rangs moyens sont déterminés par le rang du système par rapport aux rangs moyens.



Comparaison des rangs pour les scores moyens | Comparaison des rangs pour les rangs moyens

Figure 3 – Comparaison des rangs obtenus pour les données complètes et réduites (req. 425)

La figure 3 montre les différences dans le classement obtenu lorsque l'ensemble des mesures de performance sont considérées (axe des X) et lorsque l'ensemble réduit de mesure est utilisé (axe des Y) pour la requête 425. La partie de droite considère les scores moyens alors que celle de gauche considère les rangs moyens. Quelle que soit la méthode, on constate que les résultats sont stables. Cependant, le classement est plus homogène en utilisant la méthode des scores moyens. Ceci est dû à un nombre d'ex-aequo beaucoup plus important en utilisant la méthode du rang des rangs moyens qu'en utilisant la méthode du rang des scores moyens. Le test de corrélation de Kendall indique une corrélation de 0,916 pour les rangs des scores moyens et de 0,878 pour les rangs des rangs moyens.

6 Conclusions et perspectives

Dans cet article, nous nous sommes intéressés à l'analyse de données issues du domaine de la RI. Grâce aux méthodes d'analyse de données, nous avons pu réduire le nombre de mesures de performance à utiliser pour comparer deux systèmes. Nous avons montré que 7 mesures de score étaient suffisantes pour représenter une plus large gamme de mesures.

Le prolongement de ces travaux vise à utiliser les méthodes d'analyse de données pour l'étude fine des impacts des différents modules de RI et des caractéristiques des requêtes. En effet, dans cette présente étude, nous avons considéré les systèmes de RI comme des boites noires et n'avons pas considéré leurs caractéristiques. De plus, nous avons considéré de la même façon tous les besoins d'information. Nous souhaitons dans le futur réaliser une étude plus fine qui prendrait en compte ces aspects. Ces travaux s'inscrirons dans le cadre du projet ANR CAAS (Analyse Contextuelle et Recherche d'information Adaptative) dans lequel deux partenaires industriels du domaine de la RI collaborent.

Références

Baccini A., Déjean S., Mothe J., (2010). Analyse des critères d'évaluation des systèmes de recherche d'information. Technique et Science Informatiques, 29(3) :289-308.

Banks D., Over P., Zhang N.-F., (1999). Blind Men and Elephants: Six Approaches to TREC data, Information Retrieval, 1(1-2), 7-34.

Buckley, C. (1991). Trec eval,available at http://www-nlpir.nist.gov/projects/trecvid/trecvid.tools/trec_eval_video/README

Poirier J. and Sansas B. (2009). Comparaison des classements de systèmes de recherche d'information en fonction des mesures de performances utilisées. Rapport Interne IRIT/RR–2009-31–FR, IRIT.

Seber, G.A.F. (1984). Multivariate Observations, Wiley.

Summary

This paper presents the use of data analysis methods for information retrieval. More specifically, we analyze the set of the data resulting from TREC survey and show that performance measures used to evaluate systems can be reduced to 7 instead of the 128 which are integrated in the trec_eval evaluation program.

Chapitre 3

Introduction et exploitation de la sémantique dans les systèmes pair à pair hybrides

Introduction et exploitation de la sémantique dans les systèmes pair à pair hybrides

Zekri Lougmiri*, Kenniche Ahlem**, Beldjilali Bouziane***

Département Informatique, Faculté des Sciences

Université d'Oran Es-Sénia, BP 1524 El- M'naouer, Maraval, Oran 31000, ALGERIE

* zekri.lougmiri@univ-oran.dz,

**ahlemkenniche@yahoo.fr,

*** bouziane.beldjilali@univ-oran.dz

Résumé. Les systèmes pair à pair non structurées, comme Gnutella, sont intéressants pour la recherche d'information à large échelle. Le problème significatif dans ce type d'architectures est que le routage des requêtes est purement automatique et n'inclut aucune sémantique. Ce routage dépend du nombre de sauts(Time To live: TTL) qu'effectue la requête et de la dynamique du système pour localiser les réponses.

Nous nous concentrons sur l'approche hybride *(Gnutella.V0.6)* et nous proposons une amélioration, faisant usage de la théorie de la percolation, qui permet le calcul d'un seuil critique que nous appelons seuil de satisfaction. Celui-ci permettra aux ultrapairs une prise de décision intelligente dans le processus de routage, au lieu de l'utilisation automatique du TTL. Cette décision est basée sur l'analyse de la qualité des réponses retournées par les pairs. Aussi, nous proposons une réorganisation dynamique du réseau, afin de rapprocher les profils sémantiques des nœuds.

1 Introduction

Les systèmes pair à pair sont récemment devenus un média populaire pour le partage d'énormes quantités de données. Leur première architecture était centralisée où le serveur indexait les données et les pairs. Mais lorsque le nombre de requêtes augmente, les frais engagés sur l'index centralisé deviennent très élevés; ce qui constitue un goulot d'étranglement. Leur seconde architecture est devenue décentralisée où les fonctionnalités des pairs sont similaires. Gnutella est le modèle représentatif de ces systèmes. Dans ce système, la localisation de données se fait par diffusion. Cette diffusion est pénalisante car elle noie rapidement le système; plus encore, la localisation de fichiers n'est pas exhaustive car elle est sujette à la dynamique du réseau et au TTL qui contrôle la propagation de requêtes. Ainsi, il devient clair que Gnutella n'est pas un protocole intelligent.

Nous nous intéressons aux systèmes hybrides et plus précisément à *(Gnutella.V0.6)*. Cette version introduit le schéma des *ultrapairs* et des *nœuds feuilles* pour créer une structure hybride de Gnutella. L'avantage de ce réseau est la combinaison de l'efficacité des systèmes centralisés et décentralisés.

Plusieurs travaux ont implémenté la sémantique dans les systèmes pairs à pairs. Crespo et Garcia (2004) ont proposé SON où les pairs inscrivent explicitement leurs domaines d'intérêts. Sauf que le maintien des index est coûteux. Alexander et al (2005) ont proposé INGA qui est un système décentralisé composé d'un ensembles de niveaux sémantiques que les pairs atteignent par des index et des raccourcis. La même fonction de calcul d'INGA fut implémentée dans Bibster par Hasse et al (2004). Zaharia et al. (2007) proposent des systèmes hybrides basés sur des algorithmes probabilistes de routage. Nos principales contributions dans ce papier sont :

- Alléger la fonctionnalité des ultrapairs en faisant participer les nœuds feuilles à la recherche. Donc notre recherche se fait en deux étapes une recherche locale puis une recherche globale.
- Nous calculons un seuil, issu de la théorie de la percolation, qui permettra une prise de décision sur la propagation de la requête vers d'autres nœuds. Ce seuil mesure la qualité des réponses fournies, au lieu de l'utilisation du TTL qui est une méthode purement automatique. A notre connaissance, c'est la première fois que ce seuil, riche en sémantique, est défini.
- Nous proposons une méthode de tri (Ranking) des fichiers selon les mots de la requête et selon la popularité de ces mots.
- Nous réorganisons le réseau, pour lui permettre d'être plus dynamique et ouvert, nous proposons que les pairs qui ont répondu souvent et efficacement aux requêtes émises par un ultrapair, autre que le leur, seront recâblés avec celui-ci.

2 Théorie de la percolation:

La **percolation,** Broadbent et al. (1957) et Wikipedia, est un *processus physique critique* qui décrit pour un système, une transition d'un état vers un autre. C'est un phénomène de seuil associé à la transmission d'une « information » par le biais d'un réseau de sites et de liens qui peuvent, selon leur état, relayer ou non l'information aux sites voisins. Pour ***z*** voisins dans un site, le seuil de percolation selon Stauffer et al. (1992) se calcule par: $p_c = 1 / (z - 1)$ **(1)**

3 Modèle proposé :

Notre méthode passe par deux étapes (figure2) présentées ci dessous :

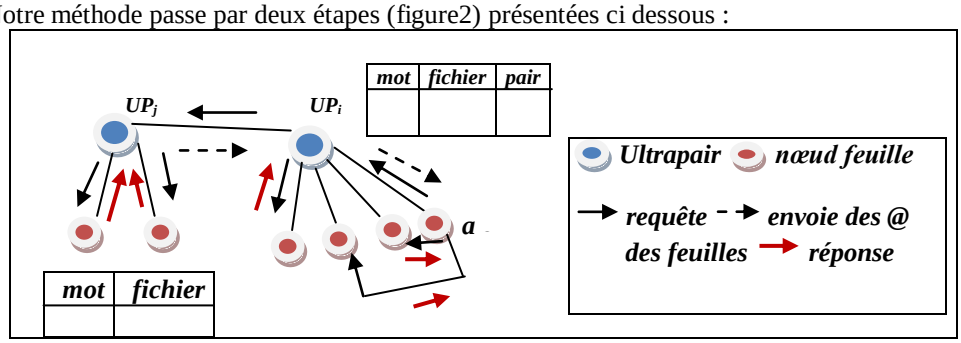


FIG. 2- *La recherche locale et globale dans la méthode proposée*

3.1 Les outils d'implémentation :

Nous dotons les pairs des structures de données suivantes :

- Un index local mot-fichier : implémenté au niveau de chaque feuille. Il capture pour chaque mot la liste des fichiers qui le contiennent.
- Un index global mot-fichier-pair : implémenté au niveau de chaque ultrapair. Il capture les pairs, leurs fichiers et les mots contenus dans ces fichiers.
- La matrice d'incidence mot-mot : C'est une matrice où M[mi,mi] est le nombre d'apparitions du mot mi dans les requêtes antérieures et M[mi,mj] est le nombre de fois où les deux mots apparaissent ensembles. Cette matrice étant symétrique, nous sauvegardons uniquement sa partie supérieure non creuse.
- Index résultat : Un index au niveau de chaque ultrapair. Il est mis à jour à chaque fois qu'un pair feuille répond positivement à une requête d'un autre ultrapair.

3.2 La recherche locale :

La feuille ***a*** (figure2), désirant récupérer une information, réalise une recherche locale par un broadcast vers tous ses voisins directs avec un TTL=3 Yang et al(2003). Ce qui signifie qu'au niveau du cluster, la recherche deviendra purement décentralisée. Lorsque la recherche se termine, ***a*** calcule le seuil de satisfaction pour décider s'il route la requête vers UP_i ou s'il est satisfait des réponses fournies par ses voisins. Au dessous de ce seuil, la requête est limitée au cluster où elle a été initiée et nous avançons que le nœud ***a*** sera satisfait des réponses de ses voisins. Au dessus du seuil, elle sera acheminée vers son ultrapair $\boldsymbol{UP_i}$, jusqu'à la satisfaction du seuil.

La situation ainsi décrite est une transition de phase type, basée sur un riche seuil. Ce seuil est inspiré du seuil de percolation, puisque notre réseau satisfait les hypothèses fondamentales de la percolation:

- Nous avons un réseau avec un espace contenant un nombre de nœuds illimité.
- La relation entre les nœuds repose sur un aspect local puisque chaque nœud achemine sa requête vers son voisin direct du même cluster.
- Les liens entre ces nœuds ont un caractère aléatoire.

Un document est dit de qualité s'il contient le maximum de mots clés de la requête. A partir de là, le nœud ***a*** procèdera au calcul du seuil comme suit :

Soit une requête de *K* mots, et soit *NL* le nombre de documents retournés contenant exactement *K* mots de la requête, alors :

Si NL= 0 alors la requête sera transmise vers l'ultrapair $\boldsymbol{UP_i}$.

Sinon si *NL≠0* alors calculer le seuil, suivant la formule : $p= \frac{NL}{résultats} < \frac{1}{Z-1}$ (2)

Où ***z*** et le nombre de voisins du pair émetteur et **résultats** est le nombre de résultats trouvés.

Un fichier contenant un grand nombre de mots de la requête aura plus de probabilité de satisfaire la demande de ***a***. Le nombre de résultats fournis avec des documents contenant le

maximum de mots clés de la requête, nous est décisif pour pouvoir décider de la satisfaction du pair demandeur. En effet, ce test est un gain pour la réduction du trafic dans le système.

3.3 La recherche globale

La recherche globale se fait au niveau des ultrapairs. Dans le cas où le seuil n'est pas respecté au niveau de ***a***, UP_i consulte son index et réunit un nombre de résultats; puis il recalcule le seuil de satisfaction selon la formule (3). Si celui-ci est respecté, la recherche se termine et UP_i enverra les réponses au nœud ***a***, sinon la requête sera propagée en dehors du cluster vers les UP voisins. Dans notre exemple, elle est acheminée vers UP_j. Chaque UP procèdera de la même manière pour la recherche et calculera son seuil de satisfaction.

$P = \frac{NL}{résultat+} < \frac{1}{Z'-1}$ (3) où z' est le nombre de feuilles de l'UP et Résultat+ est la somme des résultats obtenus.

3.4 Le tri des résultats

Après avoir terminé la recherche, le nœud source passe à la recherche de documents les plus pertinents pour sa requête. Les réponses retournées sont sous la forme de triplets (Nom du Fichier, Mots, Score) où Mots contient l'ensemble de mots qui figurent dans la requête ayant été trouvés dans le document. Score est la popularité de ce dernier. Il est calculé à partir de la matrice des mots Zekri(2010). C'est à ce niveau que l'index résultat est mis à jours (figure3).

3.5 Réorganisation du réseau

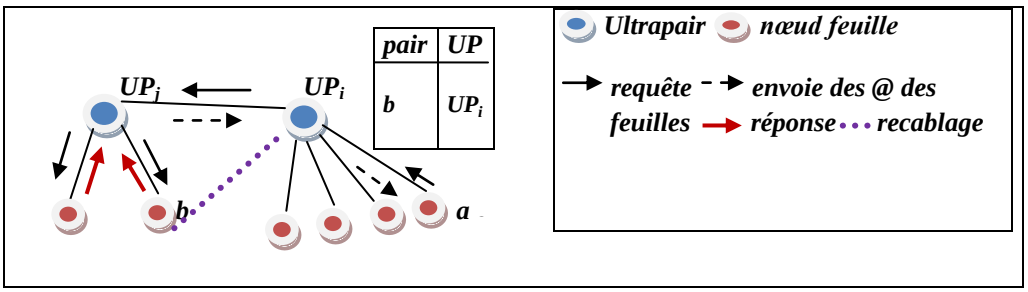


FIG. 3- *recâblage des nœuds et réorganisation du réseau*

Nous rassemblons les pairs qui ont échangé des ressources. Si un pair satisfait plus les membres d'un autre, selon Sripanidkulchai et al. (2003) qui stipulent que si un pair a un morceau particulier de contenu qui intéresse un autre pair alors il est très probable qu'il ait des ressources qui l'intéressent aussi. Cela voudrait dire que les pairs de ce cluster partagent un intérêt commun et détiennent des ressources qui intéressent le nœud en question. Le cas idéal serait qu'ils communiquent avec ces pairs en formant un seul cluster. Suivant la théorie de la percolation, une connexion est ouverte avec une probabilité p $(0 \leq p \leq 1)$ et fermée avec une probabilité $(1-P)$. Dans notre étude, nous créons un lien entre le nœud feuille qui répond positivement aux requêtes d'un autre UP. Dans la figure 3, le nœud feuille ***b*** répond souvent positivement aux requêtes de UP_i. La probabilité p pour que le nœud ***b*** soit lié avec UP_i est

mesurée par la formule (4): $p=\frac{réponse}{Résultat} < \frac{1}{Z''-1}$ (4) Où **Réponse** : est le nombre de réponses de ce nœud et **Résultat** : la somme des résultats positifs donnés par l'ultrapair au nœud source et z'' étant les voisins de $\boldsymbol{UP_j}$.

4 Expérimentation:

Nous avons comparé la performance des deux systèmes *Gnutelle0.6* et notre approche en se basant sur deux métriques : le coût de l'inondation (nombre de messages au niveau des ultrapairs) et la qualité des résultats obtenus pour chaque méthode. Nous avons développé notre propre plate forme de simulation en utilisant le langage java (plate forme *Eclipse*). Nos expérimentations ont été exécutées sur un PC portable de type « *Sony* » avec 2 GB de mémoire vive et un processeur « *Intel(R) Pentium(R) Dual CPU T2330 2.00GHz (2 CPU)* ». Notre configuration est définie les paramètres donnés en début de simulation suivants:

TTL : profondeur maximale de recherche, initialisé à 3.
Taille du réseau : nombre de pairs dans le réseau, initialisé à 100.
Nombre de requêtes : nombre de requêtes émises par les différents pairs du système.

FIG. 4- *Nombre de messages*

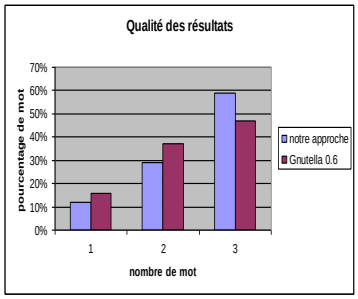


FIG. 5- *Qualité des résultats*

On remarque sur la figure 4 que le nombre de messages dans notre approche a énormément diminué par rapport à Gnutella 0.6, ce qui évite le goulot d'étranglement des ultrapairs. Ceci s'explique par le fait que le routage vers les ultrapairs est guidé par le calcul d'un seuil qui permet une prise de décision qualitative suivant la satisfaction du nœud source. La figure 5 montre les résultats obtenus en mesurant la qualité des documents obtenus par les deux méthodes. La qualité des résultats obtenus par notre approche pour une requête de 3 mots est supérieure à celle de Gnutella 0.6, sachant que la recherche avec cette méthode avait touché à un plus grand nombre de nœuds et a généré un plus grand nombre de messages. Nous pouvons dire que le calcul du seuil a permis de donner des réponses de qualité avec un gain en nombre de messages. Pour les documents avec 1 mot et 2 mots la recherche dans Gnutella 0.6 a retourné un plus grand nombre de documents, cela s'explique par le grand nombre de nœuds touchés par l'inondation faite au niveau des ultrapairs.

5 Conclusion

Nous avons proposé, une méthode de routage des requêtes dans les systèmes p2p hybrides basée sur la théorie de la percolation. Cette méthode permet aux feuilles et aux ultrapairs une

prise de décision intelligente et sémantique, puisque notre méthode prend en compte le calcul du seuil de satisfaction et la qualité des résultats retournés. Des tests, réalisés en utilisant notre plateforme de simulation, ont montré que notre algorithme de routage est plus performant qu'un algorithme de routage classique, en terme de réduction en nombre de messages au niveau des UP et en qualité des résultats.

Références

Alexender Loser, steffen Staab, Christoph Tempich : *« Semantic Social Overlay Networks »*, IEEE Journal On Selected Areas In Communications, December 2005.

P. Haase, J. Broekstra, M.Ehrig, M. Menken, P.Mika, M. Plechawski,P. Pyszlak, B. Schnizler, R. Siebes, S. Staab, and C. Tempich, "Bibster- a semantics-based bibliographic peer-to-peer system," in 3rd. Int. Semantic Web Conference (ISWC), 2004.

Broadbent, S-R. et J-M. Hammersley (1957). Percolation Processes I. Crystals and Mazes. Proceedings of the Cambridge Philosophical Society. Vol. 53, n° 3, pp. 629-641.

Crespo.A, Garcia-Molina.H. Semantic Overlay Networks. Comp. Science Dep., Stanford University, Tech. Rep., 2002

Gnutella website. http://www.gnutella.com.

Sripanidkulchai, K, B. Maggs, H. Zhang(2003). Efficient content location using interest-basedlocality in peer-to-peer systems, in: Proceedings of the 22nd IEEE INFOCOM, vol. 3.

Stauffer et A. Aharony(1992). Introduction to Percolation Theory, Second Edition, **pp**. 68

Yang, B. et H. Garcia-Molina(2003). Designing a Super-Peer Network . Data Engineering, International Conference on, pp. 49. Proceedings of the 1st NSDI

Zaharia,M et S. Keshav(2007). Gossip-based Search Selection in Hybrid Peer-to-Peer Networks. Proc. Ann. Int'l Workshop Peer-to-Peer Systems (IPTPS).

Zekri.L et Beldjillali.B(2010). Semantic Search method based on keywords in p2p systems. Soumis à publication.

Summary

Unstructured Peer to peer systems, like Gnutella, are interesting for retrieving information on a large scale. The significant problem is that the routing of queries is purely automatic including no semantics, it depends on the value of the TTL and on the dynamics of the system to find the answers.
We focus on the hybrid approach (GnutellaV0.6) and we propose an improvement, making use of percolation theory, which allows the calculation of a critical threshold, that we call satisfaction threshold and enable peers to take intelligent decision in the routing process, instead of the automatic use of TTL. Thus, we propose a dynamic reorganization of the hybrid network, in order to approximate the semantic profiles of nodes.

Chapitre 4

FLM-rule-based prognosis

FLM-rule-based prognosis

Florent Martin*,**, Nicolas Méger*
Sylvie Galichet*, Nicolas Bécourt**

*Université de Savoie, Polytech Annecy-Chambéry, LISTIC laboratory,
B.P. 80439, F-74944 Annecy-le-Vieux Cedex, France
{firstname.lastname}@univ-savoie.fr,
http://www.polytech.univ-savoie.fr
** Adixen, 98 Avenue de Brogny, F-74009 Annecy, France
{firstname.lastname}@adixen.fr,
http://www.adixen.fr

Abstract. This paper presents a local pattern-based method that addresses system prognosis. It also details a successful application to complex vacuum pumping systems. The results that we got for production data are very encouraging. We are now deploying our patent pending solution for a customer of the semiconductor market.

1 Introduction

In the current economic environment, industries have to minimize production costs and optimize the profitability of equipments. Fault prognosis is a promising way to meet these objectives. Most of the prognosis applications come from aerospace (e.g., Letourneau et al. (1999); Schwabacher and Goebel (2007)) or medicine (e..g, Magoulas and Prentza (2001)). In Schwabacher and Goebel (2007), prognosis is defined as detecting the precursor signs of a system disfunction and predicting how much time is left before a major failure. In industrial applications, most of data-driven approaches are neural network based (e.g., Schwabacher and Goebel (2007)). Unfortunately, neural networks are limited by their inability to explain their conclusions (e.g, Magoulas and Prentza (2001)). Thus we propose to extract local patterns from historical data, namely *First Local Maximum-rules (FLM-rules)* (Méger and Rigotti (2004)), to perform prognosis using real-time production data. A successful application to the prognosis of vacuum pumping system failures is detailed in this paper. When dealing with such pumping systems, common sensing technology (power levels, temperature, pressure and flow rates) is not sufficient to deal efficiently with failure prognosis. We thus use vibratory data, more informative about system status, but also more difficult to handle. Indeed, because of their complex kinematic, vacuum pumping systems may generate high vibration levels even if they are in good running conditions. This paper is structured as follows: Section 2 briefly reviews the related work and gives an overview on FLM-rules. Section 3 introduces our industrial application and describes the way data are selected and preprocessed. Section 4 presents the process used to select the most predictive FLM-rules. In Section 5, real time prognosis

is detailed. Finally, experimental results are presented in Section 6 while Section 7 ends this paper.

2 Related work

Although maintenance is a manufacturing area that could benefit a lot from data mining solutions, few applications have been identified so far. Most of them are diagnosis applications (i.e. identifying problems without predicting them) (Schwabacher and Goebel (2007)). Prognosis applications often meet difficulties in predicting how much time is left before a failure. End-users indeed have to set the width of the temporal window that is used to learn a model and predict failures (e.g, Letourneau et al. (1999); Cho et al. (2007)). As further expounded in Section 3, the dataset we deal with is a large sequence of events, i.e. a long sequence of time-stamped symbols. Such a context has been identified in Hatonen et al. (1996). More precisely, in Hatonen et al. (1996), a data mining application, known as the *TASA project*, is presented. It aims at extracting episode rules that describe a network alarm flow. These rules syntactically take the temporal aspect into account. They are known as *episode rules*. They are selected according to a frequency (or support) measure and a confidence measure (for more formal definitions, the reader is referred to Mannila et al. (1997)). In practice, users have to set the maximum temporal window width of the episode rule occurrences so as to make extractions tractable. Though this approach has not been designed to perform prognosis, it inspired some works that actually aim at predicting failures. For example, in Cho et al. (2007), the authors propose a method that searches for previously extracted episode rules in datastreams. It has only be tested on synthetic data. As soon as an episode rule is recognized in a datastream, it is used for predicting future events by adding the maximum time width to the occurrence date of the first symbol of the episode rule. This method is interesting but still, a crucial temporal information, i.e. the maximum window width, has to be set by users and is considered to be the same for all possible rules. However, it is quite difficult to justify the use of a same maximum window width for each rule. The same kind of problem can be found when extracting global models as well (Letourneau et al. (1999)). It has thus been proposed in Méger and Rigotti (2004) to extract *FLM-rules*. These episode rules indeed come along with their respective *optimal temporal window widths*. For example, rule $A \rightarrow B \Rightarrow F$, whose optimal window width is w time units, can be interpreted as: "if A occurs at t_A followed by B at t_B ($t_A < t_B < t_A + w$), then F should occur within $]t_B, tA_+w]$ with a confidence c_w higher than γ, a minimum confidence threshold". Confidence value c_w is the first local maximum of confidence such that there exists a larger window width for which the confidence of the rule is at least $decreaseRate$ % lower than c_w. Threshold $decreaseRate$ allows selecting more or less pronounced maxima of confidence. Finally, this rule has also been observed at least σ times for window widths less or equal to w. Threshold σ is termed as the minimum support. All thresholds are user-defined. Due to space limitations, the reader is referred to Méger and Rigotti (2004) for more detailed and formal definitions. In Le Normand et al. (2008), we proposed a first approach for using the optimal window widths so as to establish prediction dates. As explained in Section 5, it was not consistent with respect to the definition of FLM-rules.

3 Application and data preprocessing

We consider complex (i.e. with a complex kinetic) vacuum pumping systems running under really severe and unpredictable conditions. One major default mode that can affect these systems is the seizing of the pump axis. Seizings can be provoked by many causes such as heat expansion or gas condensation. Preventive maintenance plannings are not efficient. Therefore, Alcatel Vacuum Technology initiated a predictive maintenance project. With this aim in view, the quadratic mean (Root Mean Square, denoted RMS) of the vibration speed over 20 frequency bands has been collected over time at a frequency of $1/80$ Hz. RMS is the standard deviation of the vibration speed, i.e. the power content of the signal. Available data cover more than 2 years for 64 identical pumping systems. In order to build the learning dataset, we selected data collected before a serious and fairly common type of failures we want to anticipate, i.e. first seizings. So we build learning sequences that start at the system startup date and end at the first failure occurrence. We got 13 doubtless sequences that end with the first occurrence of a seizing. In order to detect evolutions of systems, for each system, the experts first determine the signal power content P_0 corresponding to good running conditions. The measured signal power content P is then compared to this reference by computing the ratio P/P_0 which is in turn quantized using 3 levels. This defines the default severity and it is done for each frequency band. Then, we define a dictionary of 240 symbols, each symbol being associated with three pieces of information: the frequency band, the default severity and the duration at that severity level. We also introduce a specific symbol to represent seizing occurrences. Finally, we got 13 sequences containing 2000 symbols on average along with their occurrence dates. They were concatenated into a single *large sequence* (Hatonen et al. (1996); Mannila et al. (1997); Méger and Rigotti (2004)). In order to avoid the extraction of FLM-rule occurrences spreading over various subsequences, a large time gap between each initial sequence is imposed and the *WinMiner* algorithm (Méger and Rigotti (2004)) is used to extract FLM-rule occurrences. WinMiner can indeed handle a maximum time gap constraint (denoted as $maxgap$) between occurrences of symbols.

4 Rule selection

As we aim at predicting seizings, we only retain FLM-rules concluding on the symbol "seizing". Furthermore, we select the most predictive ones. Our selection process is inspired from the well known *leave-one-out cross validation* technique. It involves considering alternatively initial sequence ending with a failure, as a validation set while other subsets form the learning set. Each FLM-rule triggering a false alarm is rejected. More details can be found in Martin et al. (2010). The set of selected FLM-rules is termed as the *FLM-base*. Back to our application, as seizings can originate from very different causes, and as we only have 13 subsets relating to a seizing, the minimum support threshold σ is set to 2. In order to extract the most confident rules, the minimum confidence threshold γ is set to 100%. Parameter $decreaseRate$ is set to 30% to select pronounced/singular optimal window widths and the *maxgap* constraint is set to 1 week to consider very large optimal window widths. Finally, we asked for FLM-rules containing 4 event types as a maximum so as to consider generic rules and to make extractions tractable. At the end of the rule selection process, we got a FLM-base containing 29 FLM-rules with their respective optimal window widths. Running such a process remains tractable: execu-

tion times does not exceed 3 hours on a standard PC (proc. Intel Xeon CPU 5160 @ 3.00GHz, 3.9 Go ram, linux kernel 2.26.22.5).

5 Real time prognosis

To match the premisses of the FLM-rules belonging to the FLM-base, we maintain a queue of event occurrences whose time span is lower than W, the largest optimal window width of the FLM-base. We thus make sure that enough data are kept for being able to identify the premisses of all the FLM-rules that form the FLM-base. We recall that FLM-rules occurrences are minimal ones, i.e. they can not overlapp. For a more formal definition, the reader is referred to Méger and Rigotti (2004). Each time a new event occurs, it is added to the queue. If that one corresponds to the suffix (last event) of the premiss of a FLM-rule r belonging the the FLM-base, a premiss matching process is launched. We scan the queue through an observation window $W_r^o =]t_r, t_0[$ with t_0 the date at which rule matching process is launched and $t_r = t_0 - w_r$, with w_r the optimal window width of rule r. Indeed, in the worst case scenario, the conclusion of rule r is about to occur at $t_0 + 1$ and the earliest date of occurrence of the first symbol of its premiss is $t_0 + 1 - w_r = t_r + 1$. As proposed in Cho et al. (2007), in this observation window, we search for the latest minimal occurrence of the premiss of r which amounts in finding the occurrence date of its first event, denoted ts_r and defined as follows: let Ts_r be the set of the occurrence dates of the first event of the premiss occurrences of rule r that occurs in $]t_r, t_0]$. The date ts_r is the single element in Ts_r such that $\nexists\, t \in Ts_r$ with $ts_r \neq t$ and $t > ts_r$ and such that the conclusion of rule r does not appear in $]ts_r, t_0]$. Then, by definition of FLM-rules, we forecast the conclusion of rule r in $w_r^f =]t_0, tc_r]$ with $tc_r = ts_r + w_r$. By construction $t_0 < tc_r < t_0 + maxgap$. In Le Normand et al. (2008), for each matched premiss of rule r, its conclusion is forecasted at tc_r though it may appear in $]t_0, tc_r]$ by definition of FLM-rules. The prognosis approach proposed in Le Normand et al. (2008) is thus not consistent with respect to the definition of FLM-rules.

Let Tc be the set of all prediction dates tc_r that are active, i.e. that are greater than t_0 (those prediction dates can be computed before and at t_0). The associated failure prediction time interval $]t_s^f, t_e^f]$, also termed as the forecast window W^F, is such that $t_s^f = t_0 \wedge t_e^f = min(Tc)$. Figure 1 provides a forecast window established using rules α, β, δ that have been recognized at t_0^{n-1} and t_0^n.

6 Experimental evaluation of prognosis

In order to evaluate the accuracy of our forecast, we consider two cases:

- each time our forecast method foresees a seizing, we check if seizing really occurs in the given forecast window $W^F =]t_s^f, t_e^f]$.
- each time t_0 a new event arises, and if no warning is triggered, we check if a seizing occurs in $]t_0, t_0 + maxgap]$. We extracted rules under $maxgap$, a maximum time gap between events. We thus can not predict any occurrence of conclusions of FLM-rules after $t_0 + maxgap$.

Two datasets were considered: the 13 sequences used to build our FLM-base (dataset 1) and 21 new sequences of production data (dataset 2). Using these datasets, we simulated 2 datas-

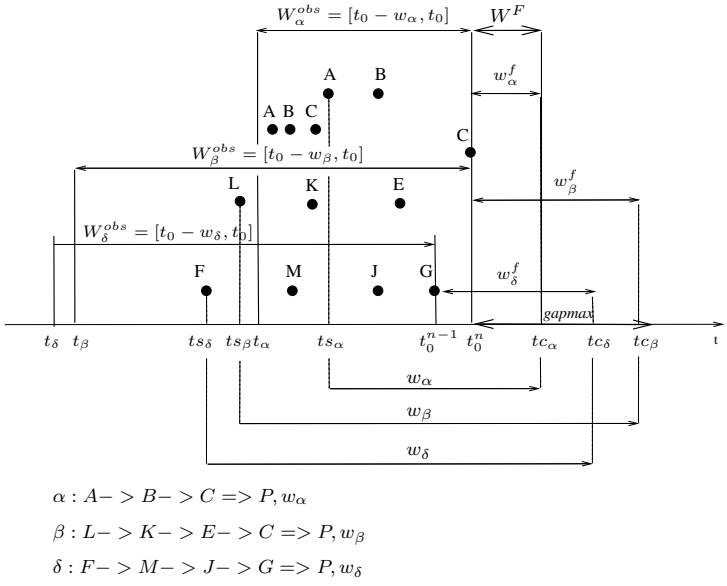


FIG. 1 – *Merging prediction information of FLM-rules.*

treams and made respectively 24125 and 32525 forecasts. Results of evaluations are given in Table 1 and in Table 2, using confusion matrices. Though we could access few data relating to failures so far, results are really encouraging as we predicted 10 seizings out of 13 with $99,97\%$ of accuracy on dataset 1 and as we foresaw 2 upcoming seizings with $98,75\%$ accuracy on dataset 2. The 262 misforecasts on dataset 1 all relate to the 3 failures that have not been detected. Furthermore, the 20 and 404 false alarms (tables 1 and 2) stating that pump will seize, were generated on pump that really seized few days later. Earliest failure predictions provided by our software prototype arise at the latest 3 hours before the seizing really occurs and, most of the time, more than 2 days before. This is enough to plan an intervention. Those results are good and the solution is beeing deployed for a client of the semi-conductor market.

	$\widehat{failure}$	$\widehat{healthy}$
failure	492	262
healthy	20	23351

TAB. 1 – *Results for dataset 1.*

	$\widehat{failure}$	$\widehat{healthy}$
failure	300	0
healthy	404	31821

TAB. 2 – *Results for dataset 2.*

7 Conclusion and perspectives

In this paper, we present a local pattern-based approach for prognosing failures using FLM-rules. We applied this approach in an industrial context in which vacuum pumping systems are

running under severe and unpredictable conditions. Results are encouraging as we forecast failures with a good accuracy, i.e. more than 98% on both learning data and new data. Moreover, using our predictions, enough time is left to technical teams for planning an intervention. The presented prognosis method is patent pending. Others applications to telecommunication networks, constant frequency rotating machines or supply chain management can be considered. Future work directions include introducing fuzzy logic to merge prediction dates so as to provide end-users with gradual warnings.

References

Cho, C., Y. Zheng, and A. Chen (2007). Continuously matching episode rules for predicting future events over event streams. In *Advances in Data and Web Management*, Volume 4505, pp. 884–891.

Hatonen, K., M. Klemettinen, H. Mannila, P. Ronkainen, and H. Toivonen (1996). Tasa: Telecomunications alarm sequence analyzer or: How to enjoy faults in your network. In *1996 IEEE Network Operations and Management Symposium (NOMS'96)*, pp. 520–529.

Le Normand, N., J. Boissiere, N. Méger, and L. Valet (2008). Supply chain management by means of flm-rules. In *12th European Conf. on Principles and Practice of Knowledge Discovery in Databases (PKDD'08), Induction of Process Models Workshop*, pp. 29–36.

Letourneau, S., F. Famili, and M. S. (1999). Data mining for prediction of aircraft component replacement. *Jr. of Intelligent Systems and their Applications, IEEE 14*(6), 59–66.

Magoulas, G. and A. Prentza (2001). Machine learning in medical applications. *Jr. of Machine Learning and Its Applications 2049*, 300–307.

Mannila, H., H. Toivonen, and A. Verkamo (1997). Discovery of frequent episodes in event sequences. *Data Mining and Knowledge Discovery 1*(3), 259–298.

Martin, F., N. Méger, S. Galichet, and N. Bécourt (2010). Episode rule-based prognosis applied to complex vacuum pumping systems using vibratory data. In *Advances in Data Mining: Applications and Theoretical Aspects, 10th Industrial Conference on Data Mining ICDM 2010*, pp. 376–389.

Méger, N. and C. Rigotti (2004). Constraint-based mining of episode rules and optimal window sizes. *In Proc. of 8th European Conference on Principles and Practice of Knowledge Discovery in Databases (PKDD'04) 3202*, 313–324.

Schwabacher, M. and K. Goebel (2007). A survey of artificial intelligence for prognostics. In *Working Notes of 2007 American Institute in Aeronautics and Astronautics Fall Symposium: AI for Prognostics*.

Résumé

Cet article présente une méthode à base de motifs locaux permettant de faire du pronostic. Une application à des systèmes de pompage complexes est détaillée. Les résultats obtenus pour des données de production sont très encourageants. Notre solution, dont le brevet est en cours d'homologation, est en train d'être déployée chez un client du marché du semi-conducteur.

Chapitre 5

Gestion de la QoS des services ADSL à l'aided'un processus de data mining

Gestion de la QoS des services ADSL à l'aide d'un processus de data mining

Vincent Lemaire*, Françoise Fessant *

* Orange Labs, 2 avenue Pierre Marzin, 22300 Lannion

Résumé. Dans cet article l'intérêt d'une approche "fouille de données" est explorée dans le cadre du contrôle de la qualité de service (QoS) des lignes ADSL du réseau de France Télécom. Cet article présente la plateforme et les mécanismes qui ont été mis en place. Ces derniers permettent la détection et la classification des lignes bruitées au sein du réseau. L'utilisation de la classification permet une amélioration de la QoS. L'interprétation de la classification permet la découverte de connaissances actionnables.

1 Introduction

Internet est devenu la plateforme pour les services de voix et de vidéo. Toutefois la qualité des services multimédia offerts sur Internet dépend des congestions, indisponibilités, et autres anomalies survenant dans le réseau. La mesure de la qualité de service (QoS) permet de détecter quand un processus ou un élément du réseau opère en dehors de sa plage de fonctionnement ou ne fonctionne pas correctement. La collecte de mesures de QoS de bout en bout permet, en général, de déterminer la source du problème : problème lié à un élément du réseau (DSLAM (Digital Subscriber Line Access Multiplexor)...), problème lié à la ligne physique, problème lié à la Box (LiveBox^{TM}dans cet article). La connaissance de la source du problème doit permettre de réagir promptement et de manière adéquate.

Orange a mis en place un suivi de la QoS pour ses services ADSL de manière à améliorer la satisfaction de ses clients. Cet article n'a pas l'ambition de décrire l'ensemble de ce processus. Il se concentre sur la détection d'un type particulier de problème : les lignes ADSL "bruitées" et montre en quoi cette classification permet une amélioration de la QoS et l'extraction de nouvelles connaissances. Dans l'application qui est décrite, l'élément de QoS considéré est lié à la disponibilité du service de téléphonie sur IP (VOIP) par le biais des LiveBox (LB). L'ensemble du processus de data mining (Fayyad et al., 1996) qui a été mis en place est présenté Figure 1.

2 Collecte et préparation des données

Provenance des données : La sonde Audiphone est un agent logiciel embarqué dans les Live Box qui, dès que la LB est allumée, surveille la disponibilité du service de téléphonie sur IP et peut également effectuer des mesures de la qualité vocale durant les appels. L'avantage essentiel de l'agent Audiphone est qu'il est positionné au plus près de l'usager. Il est capable de détecter l'indisponibilité du service et de déterminer la cause de cette indisponibilité telle

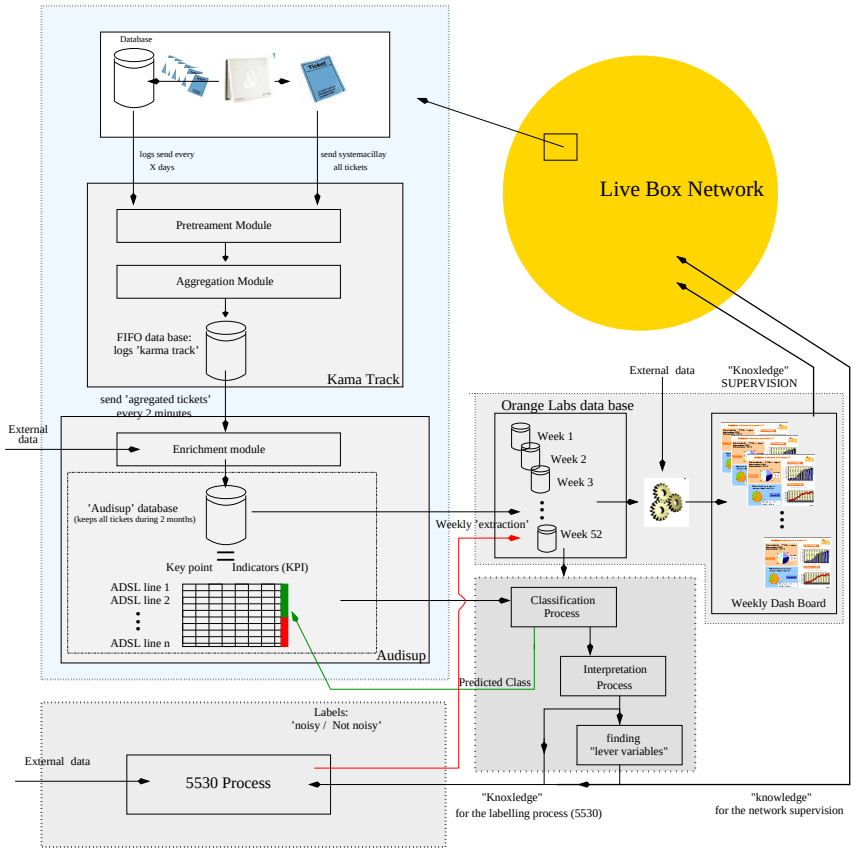


FIG. 1 – *Le processus complet de data mining : des tickets Audiphone aux indicateurs de QoS*

que vue par la livebox en extrémité de la chaîne de transmission. Il permet alors une vision du service tel qu'il est perçu par l'usager. Dans cet article nous nous intéressons plus spécifiquement à la disponibilité du service. A chaque évènement d'indisponibilité, l'agent logiciel génère un ticket qui est envoyé ensuite vers une chaîne de collecte pour stockage et traitement. Il est important de préciser que le contenu des tickets ne viole en rien la vie privée de l'utilisateur de la LB. Le contenu d'un ticket ne comporte que des informations anonymes sur la qualité et la conformité de la LB vis-à-vis des services qu'elle doit rendre.

Mise en base des données : Les tickets sont remontés vers une plateforme d'analyse et de traitement (KarmaTrack-Audisup) qui réalise l'enrichissement des tickets à l'aide d'informations réseau et le calcul de Key Point Indicators (KPI), sur différents axes tels que les types de problèmes et la durée d'indisponibilité. Les tickets sont ensuite mis en base au sein d'une application nommée 'Audisup'. Audisup traite les logs de tickets pour produire une base de données "à plat" (un tableau de N instances représentées par J variables). Cette base de données contient l'ensemble des tickets émis par les LB au cours des 35 derniers jours. Un ticket

contient différents champs. Certains de ces champs sont de type 'variable continue' d'autres sont des champs à valeurs catégorielles.

Etiquetage des lignes ADSL : L'application 5530 (Le Meur et Santos-Ruiz, 2010), développée par Alcatel, permet d'observer la qualité des lignes ADSL, par la collecte de paramètres représentatifs de leur état, au travers de l'interrogation de DSLAM télé surveillés. Le 5530 est un moyen de service après vente (SAV) supplémentaire et de gestion de QoS permettant de fiabiliser à distance le diagnostic de signalisations complexes sur l'ADSL et le multiservice. La notion de stabilité (notion extrêmement corrélée à la notion d'indisponibilité) au sein du 5530 est évaluée à l'aide de deux compteurs : le MTBR (Mean Time Between Retrains) et le MTBE (Mean Time Between Errors) basés essentiellement sur les nombres et durées de resynchronisations de la ligne. En fonction du MTBR et des MTBE, le 5530 classe les lignes en 3 catégories : Stable ; Risquée ; Instable (Le Meur et Santos-Ruiz, 2010). Il est à noter que pour connaître la stabilité de la ligne à l'aide du 5530, il faut que l'inspection ait une durée de plus de six heures de synchronisation entre le modem client et le modem DSLAM.

3 Modélisation : Classification des lignes ADSL

Notations utilisées par la suite : (*) une table de modélisation T contenant K instances et J variables explicatives ; (*) un problème de classification à C classes ; (*) un classifieur probabiliste f entraîné sur la table de modélisation ; (*) une instance x_k représentée sous la forme d'un vecteur à J dimensions.

3.1 Données et protocole expérimental

On dispose pour l'analyse de la classification des lignes ADSL d'une extraction des tickets remontés par la plateforme Audisup sur une période de 5 jours. Tous les tickets associés à une même date journalière sont stockés dans le même fichier. Chaque ligne ADSL est caractérisée par différents types d'indicateurs : les paramètres d'identification de la LB et du réseau (fixes quel que soit le jour considéré) et les paramètres qui concernent l'évènement d'indisponibilité (variables selon le jour de l'évènement). Une ligne ADSL est décrite par 123 indicateurs (variables explicatives). A ces indicateurs on rajoute l'information sur la caractéristique de la ligne "bruitée/non bruitée" provenant du service 5530. La base de modélisation contient 71164 lignes ADSL. Les priors sur les classe 'Stable', 'Risquée et 'Instable' sont respectivement de 0.881, 0.054, 0.064. Les expérimentations de classification ont été réalisées à l'aide du logiciel Khiops (développé par Orange Labs, `www.khiops.com`). Le classifieur utilisé est un classifieur naif de Bayes moyenné (Boullé, 2007).

3.2 Résultats

On a utilisé une procédure de validation croisée par k-folds (avec k=10) pour produire les résultats. Les performances en taux de bonne classification (ACC) et d'AUC sont respectivement de 0.8924 +/- 0.0017 et de 0.8185 +/- 0.0078 ; ce qui en font de très bons résultats. Les erreurs de classification proviennent majoritairement de l'affectation à tort de l'étiquette stable à une ligne jugée instable par l'application 5530. Ce résultat est cohérent avec le service 5530 qui a identifié un problème d'étiquetage pour certaines lignes étiquetées instables qui le sont à

tort. En effet certaines cartes implémentées dans les DSLAM ne permettent pas de distinguer les extinctions de modem des resynchronisations liées à la transmission ADSL. Ceci a pour conséquence d'augmenter le compteur de resynchronisations utilisé par le 5530 pour déterminer l'état d'une ligne et donc d'aboutir à l'évaluation d'une ligne instable à tort. La courbe de lift obtenue est présentée Figure 2.

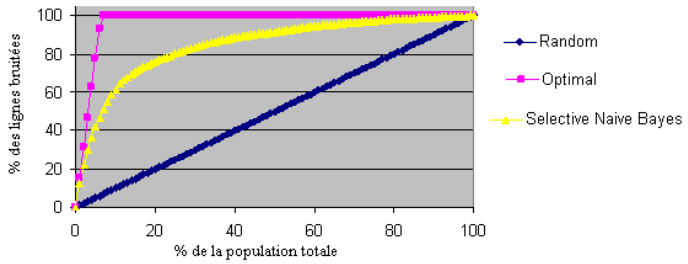


FIG. 2 – *Courbe de lift*

4 Evaluation et interprétation des résultats

4.1 Importances individuelles des variables explicatives

On propose de calculer l'importance individuelle des variables explicatives pour chaque ligne ADSL (et non en "moyenne"). On utilise dans cette section une méthode de calcul, G, permettant, connaissant f (le classifieur utilisé section 3), de calculer l'importance d'une variable en entrée du classifieur. Etant donné l'état de l'art et le type de classifieur utilisé (un classifieur naif de Bayes) on choisit d'utiliser comme mesure d'importance le "Weight of Evidence" (WoE) décrit dans (Robnik-Sikonja et Kononenko, 2008). L'indicateur WoE mesure le log d'un odds ratio. Il est calculé pour l'ensemble des variables explicatives présentes en entrée du classifieur et pour une classe d'intérêt. La classe d'intérêt (q) est en général la classe d'appartenance prédite de l'instance x_k. Une variable qui a une importance (WoE) positive contribue positivement à définir la classe prédite, à l'inverse une variable qui a une importance (WoE) négative contribue négativement à définir la classe prédite (donc contribue positivement à définir une autre classe du problème de classification).

La figure 3 illustre ce point avec 3 lignes ADSL : l'une classée par le classifieur comme 'STABLE', la seconde classée 'RISKY' et la troisième comme 'UNSTABLE'. On remarque que : (i) la ligne classée 'Stable' est caractérisée par une multitude de petites contributions positives (importances positives) et quelques contributions négatives ; (ii) la ligne classée 'Risky' est caractérisée par une multitude de petites contributions positives (importances positives) et presque aucunes contributions négatives ; (iii) la ligne classée 'Unstable' par quelques fortes contributions positives. On a donc une interprétation complètement individuelle pour chaque ligne ADSL, permettant un diagnostic précis. Pour chaque ligne on a : (i) la classe prédite par le classifieur ; (ii) un score de confiance sur cette prédiction ; un ordonnancement des variables explicatives en fonction de leur importance ; (iii) la valeur de l'importance pour chaque variable explicative.

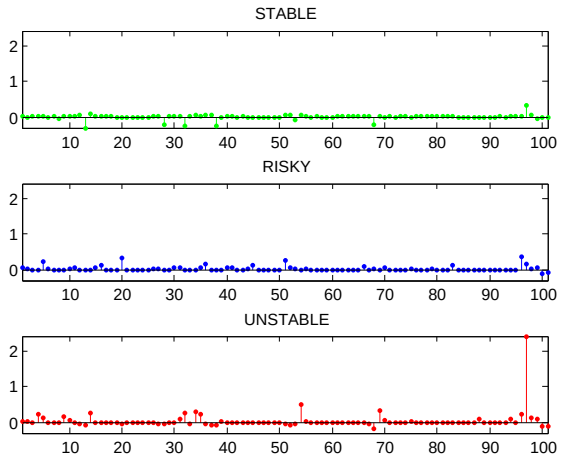


FIG. 3 – *Exemple d'importance des variables explicatives pour 3 lignes ADSL (de haut en bas) pour une ligne classée stable, une ligne classée risky et une ligne classée unstable.*

4.2 Indices d'amélioration de la stabilité des lignes

On utilise ici la méthodologie décrite dans (Lemaire et Hue, 2010). Soit C_z la classe cible d'intérêt parmi les C classes cible, par exemple ici, la classe "ligne bruitée". Soit f_z la fonction qui modélise la probabilité d'occurrence de cette classe cible $f_z(X = x) = P(C_z|X = x)$ étant donné l'égalité du vecteur X des J variables explicatives à un vecteur donné x de J valeurs. La méthode proposée ici recherche à augmenter la valeur de $P(C_z|X = x_k)$ pour chacun des K exemples de la base de données. Parmi les variables disponibles en entrée du classifieur, on exclura de l'exploration celles pour lesquelles les valeurs ne peuvent pas être modifiées. On conserve les variables dites levier, c'est à dire celles sur lesquelles on pense pouvoir agir.

On utilise, à titre d'exemple pour l'article, la variable 'LB_firmware' (LBF) comme variable levier. D'après la table de modélisation (la table ayant servi à créer le classifieur) cette variable peut prendre 4 modalités différentes que l'ont nommera A, B, C, D pour des raisons de confidentialité. L'étape de prétraitement des variables catégorielles (groupage de modalités (Boullé, 2005)) qui est la première étape lors la construction du classifieur naif de Bayes a déterminé que ces 4 groupes étaient optimaux (pas de création d'un groupe contenant plusieurs de ces modalités).

On s'intéresse ensuite plus précisément aux lignes ADSL 'Unstable' et effectivement prédites comme 'Unstable' par le classifieur (soit 1611 lignes ADSL). 1231 d'entre elles peuvent voir leur probabilité d'être stable augmenter. Pour celà la variable LBF doit prendre comme valeur 'D'. La probabilité de stabilité des 380 autres lignes ADSL ne peut pas être améliorée (la variable LBF est déjà à 'D'). On présente Figure 4 l'amélioration de la probabilité. Dans cette figure l'axe des abscisses coincide avec une modalité de LBF, puis au sein d'une modalité de LBF les valeurs de $PCa(x_k) - PCi(x_k)$ ont été ordonnées de manière croissante. On a (i) pour $x_k \in [1 : 407]$ LBF='A' ; (ii) pour $x_k \in [408 : 778]$ LBF='B' ; (iii) pour $x_k \in [779 : 1231]$ LBF='C'. Les points bleus montrent une amélioration sans changement de classe. Les points

rouges montrent une amélioration avec changement de classe. On en conclut que la variable 'L_firmware' est effectivement une variable levier. Elle permet lorsque l'on force sa valeur à 'D' d'obtenir des lignes plus stables (1231 cas sur 1611) voir d'obtenir des lignes stables (les 51 carrés rouges dans la figure 4).

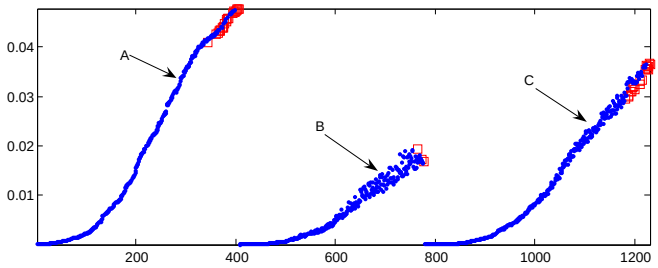


FIG. 4 – *Amélioration possible* $(PCa(x_k) - PCi(x_k))$ *de la stabilité pour les 1280 lignes ADSL* (x_k).

On a donc une exploration des corrélations ligne ADSL par ligne ADSL permettant de rechercher des moyens d'améliorer leur stabilité. Pour chaque ligne classée instable on a : (i) la probabilité initiale d'instabilité $(PCi(.))$; (ii) la probabilité améliorée (diminuée) d'instabilité $(PCa(.))$; (iii) la variable explicative qui permet le gain ; (iv) la valeur que doit prendre cette variable explicative pour obtenir le gain.

5 Discussion

De bonnes performances de classification ont été obtenues, validant le processus d'extraction et de création des données. L'analyse des variables les plus importantes montre que l'information de ligne bruitée est très corrélée à l'information de désynchronisation de la ligne ADSL (les deux variables les plus informatives pour la cible sont les tickets de perte de service de type 306_1 et de retour de service 308_1 qui comptabilisent le nombre de désynchronisations de la ligne). Ce qui est complètement cohérent avec la manière dont l'application 5530 étiquette l'état d'une ligne.

Deux manières d'utiliser le modèle de classification se dégagent de l'étude.

Une première piste pour l'utilisation du modèle en mode opérationnel est le filtrage des lignes instables pour ne garder que les lignes stables. De cette manière on filtre un grand nombre de tickets qui sont produits parce que la ligne est instable.

Une autre utilisation possible serait de renforcer la connaissance du 5530 pour améliorer l'étiquetage des lignes. En effet, un certain nombre de lignes étiquetées instables le sont à tort. Certains types de DSLAM remontent mal les compteurs des nombres de resynchronisations journaliers sur lesquels se base l'application 5530 pour en déduire l'état de stabilité d'une ligne. Ces DSLAM ne font pas la différence entre les resynchronisations et les extinctions/allumages électriques. Par exemple, il suffit que le client laisse son modem allumé moins de 4 heures dans une journée pour être déclaré instable à coup sur. Toutes les lignes pour lesquelles le client éteint son modem la nuit et pendant la journée de travail sont déclarées instables à tort. La sonde Audiphone remonte les désynchronisations de la ligne ADSL de manière indépendante des

DSLAM et le compte des tickets de ce type de remonté pourrait aider à préciser les compteurs alimentant la décision de stabilité.

Dans le cadre de l'étude qui a été menée la 'classification' des lignes ADSL les trois quarts du processus a été entièrement industrialisé : de la collecte des données à la prédiction des lignes ADSL bruitées. L'information de prédiction est elle en cours d'industrialisation : analyse précise des résultats et utilisation dans la compréhension du phénomène, exploitation des résultats dans l'amélioration du service 5530 et enfin étiquetage automatique des tickets à des fins de filtrage.

La phase d'interprétation des résultats de classification permet quant à elle d'obtenir une interprétation complètement individuelle d'une ligne ADSL autorisant un diagnostic précis. La phase d'exploration des corrélations existantes au sens du classifieur permet de rechercher des moyens d'améliorer la stabilité des lignes et ainsi d'avoir des propositions de plan d'intervention.

Références

Boullé, M. (2005). A bayes optimal approach for partitioning the values of categorical attributes. *Journal of Machine Learning Research 6*, 1431–1452.

Boullé, M. (2007). Compression-based averaging of selective naive Bayes classifiers. *Journal of Machine Learning Research 8*, 1659–1685.

Fayyad, U. M., G. Piatetsky-Shapiro, et P. Smyth (1996). *Advances in Knowledge Discovery and Data Mining*, Chapter From data mining to knowledge discovery : An overview., pp. 1–34. AAAI/MIT Press.

Le Meur, R. et L. Santos-Ruiz (2010). Evaluation de la stabilité d'une ligne ADSL par le 5530. Technical report, France Telecom Research and Development.

Lemaire, V. et C. Hue (2010). *Correlation Analysis in Classifiers*, Chapter From data mining to knowledge discovery : An overview., pp. 1–34.

Robnik-Sikonja, M. et I. Kononenko (2008). Explaining classifications for individual instances. *IEEE TKDE 20*(5), 589–600.

Summary

In this paper we explore the interest of computational intelligence tools in the management of the Quality of Service (QoS) for ADSL lines. The paper presents the platform and the mechanism used to monitoring the quality of service of the Orange ADSL network in France. This platform allows the detection and the classification of the noisy lines in the network. The interpretation of results given by the classification process allows the discovery of a knowledge used to improve the process which labels the lines (noisy / not noisy) and to prevent inefficient supervision of the network.

Chapitre 6

Un Automate Cellulaire pour la Détection de Spam

Un Automate Cellulaire pour la Détection de Spam

Fatiha. BARIGOU *, Naouel. BARIGOU**
Laboratoire d'Informatique d'Oran
Université d'Oran, BP 1524, El M'Naouer, 31 000 Oran, Algérie
*fatbarigou@gmail.com, **barigounaouel@gmail.com

Résumé. Dans le contexte du filtrage de courriels indésirables (appelé aussi spam), nous proposons l'utilisation d'une classification supervisée booléenne à base d'automate cellulaire. Nous examinons, par des expériences sur le corpus LingSpam, la performance de cette approche en variant les méthodes de prétraitement du corpus : utilisation d'une stop-liste, racinisation, et sélection des termes.

1 Introduction

Aujourd'hui, le courriel est devenu un moyen rapide et économique pour échanger des informations. Cependant, on se retrouve assez vite submergés de quantités de messages indésirables appelé aussi spam. Pour faire face à cette charge croissante des spam, de nombreuses solutions ont vu le jour (Sanz et al, 2008). Certaines solutions sont basées sur l'en-tête du courrier électronique tel que les listes noires, les listes blanches et grises. D'autres solutions sont basées sur le contenu textuel du message tel que le filtrage à base d'apprentissage (Guzella et Caminhas, 2009). Dans ce papier, nous proposons pour la première fois une approche de détection de spam basée sur l'induction symbolique par automate cellulaire (Atmani et Beldjilali, 2007). Le principe de cette approche consiste à construire un modèle booléen à partir d'un ensemble de courriels d'apprentissage pour la classification des emails entrant en spam ou légitime. La suite de cet article est organisée de la manière suivante : la section 2 est consacrée à l'étude de l'approche proposée. La section 3 présente l'étude expérimentale pour l'évaluation de cette nouvelle solution. La section 4 présente nos conclusions et quelques orientations pour les travaux futurs.

2 Approche Cellulaire de Classification

Cette section est consacrée à l'étude de la classification supervisée à base d'automate cellulaire adoptée pour la détection de spam. Le principe de cet automate est tout d'abord décrit.

2.1 Principe de l'automate cellulaire CASI

CASI (Cellular Automata for System Induction) issue des travaux de (Atmani et Beldjilali, 2007) est une méthode cellulaire de génération, de représentation et d'optimisation des graphes d'induction (Zighed, 2000) générés à partir d'un ensemble d'exemples d'apprentissage. Ce système cellulo-symbolique est organisé en cellules où chacune d'elles, est reliée seulement avec son voisinage. Toutes les cellules obéissent en parallèle à la même règle appelée fonction de transition locale, qui a comme conséquence une transformation globale du système. Deux composants coopèrent entre eux pour la construction du modèle booléen : le COG (Cellular

Optimization and Generation) qui s'occupe de la génération du graphe d'induction cellulaire et de son optimisation et le CIE (Cellular Inference Engine), un moteur d'inférence cellulaire, qui génère un ensemble de règles cellulaires utilisées pendant la phase de filtrage. Pour se faire, ils utilisent une base de connaissances sous forme de deux couches finies d'automates finis. La première couche, CelFact[1], pour la base des faits et, la deuxième couche, CelRule[2], pour la base de règles. Le voisinage des cellules est défini par deux matrices d'incidence d'entrée R_E et de sortie R_S. La dynamique de l'automate cellulaire, utilise deux fonctions de transitions $\delta fact$ qui simule les phases de sélection et de filtrage dans un système expert et $\delta rule$ qui correspond à la phase d'exécution :

$$(EF, IF, SF, ER, IR, SR) \xrightarrow{\delta_{fact}} (EF, IF, \boldsymbol{EF}, \boldsymbol{ER} + (\boldsymbol{R_E^T} \times \boldsymbol{EF}), IR, SR)$$

$$(EF, IF, SF, ER, IR, SR) \xrightarrow{\delta_{rule}} (\boldsymbol{EF} + (\boldsymbol{R_S} \times \boldsymbol{ER}), IF, SF, ER, IR, \overline{\boldsymbol{ER}})$$

2.2 Architecture

Nous présentons dans la figure1 l'architecture de notre système à base d'automate cellulaire que nous avons baptisé CASD («Cellular Automaton for Spam Detection).

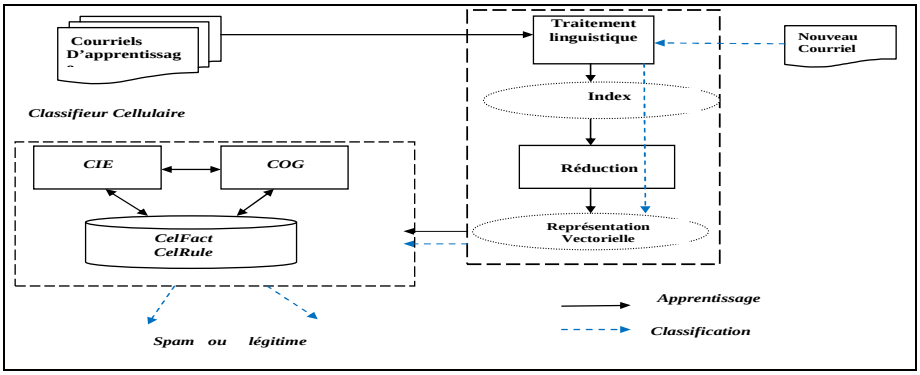


FIG.1– *Architecture de CASD*

2.3 Prétraitement Linguistique et Réduction

L'ensemble des courriels d'apprentissage doit être prétraité pour extraire les termes qui le représentent. A l'aide du module d'indexation que nous avons implémenté, nous pouvons établir une première liste de termes en procédant au découpage du texte en mots, à l'élimination des mots vides, comme nous pouvons aussi utiliser une variante de l'algorithme de Porter[3] pour la racinisation des différents mots retenus dans cette première phase. Cet ensemble de termes est par la suite réduit, par le choix de l'une des trois mesures implémentées dans CASD : l'information mutuelle (MI), le gain d'information (GI), et la statistique de Chi-2(χ^2) (Sebastiani, 2002). La sélection des termes est effectuée dans le but de choisir les éléments les plus

[1] Toute cellule de CelFact est considérée comme fait établi si sa valeur est 1, sinon, elle est considérée comme fait à établir. Elle se présente sous trois états : état d'entrée (EF), état interne (IF) et état de sortie (SF)

[2] Toute cellule de CelRule est considérée comme une règle candidate si sa valeur est 1, sinon, elle est considérée comme une règle qui ne doit pas participer à l'inférence. Elle se présente sous trois états : état d'entrée (ER), état interne (IR) et état de sortie (SR

[3] http://tartarus.org/~martin/PorterStemmer/

pertinents de tous les courriels. Le système calcule la mesure élue pour tous les termes, puis prend les K (seuil) premiers termes correspondant aux plus grandes valeurs calculées.

Comme dans la majorité des algorithmes de classification de textes, nous utilisons une représentation vectorielle (Salton et al., 1975) des courriels : le sac de mots. Ainsi chaque courriel est représenté par un vecteur d = {w_1, w_2, . . . ,$w_{|n|}$ }de $\Re^n$ où chaque coordonnée représente la présence ou l'absence (=0) du mot dans le courriel et n désigne le nombre de termes de l'index.

2.4 Apprentissage

Le processus d'apprentissage effectué par notre système est résumé dans six étapes :

1. transformation de la représentation vectorielle des courriels dans le format « arff » adopté par l'automate cellulaire,
2. production du graphe d'induction avec la méthode Sipina,
3. représentation cellulaire du graphe d'induction,
4. inférence en chainage avant : passer de la configuration G(t) vers la configuration G(t+1) en utilisant les deux fonctions de transition δfact, δrule,
5. répéter (4) jusqu'à stabilisation (i.e. G(t+1) = G(t))
6. sauvegarde du modèle booléen généré.

2.4.1 Un exemple Illustratif

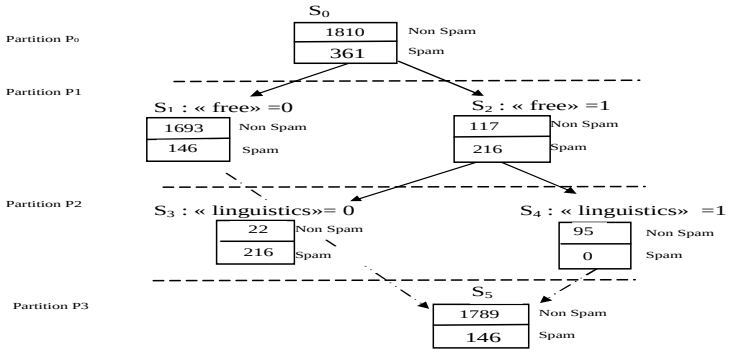


FIG. 2 – *Extrait d'un graphe d'induction : seulement les premières partitions sont présentées*

(a) Génération des règles cellulaires : un ensemble de règles est généré à partir du graphe d'induction comme illustré dans le tableau 1

Règle	Si	Prémisse	Alors	Conclusion
Règle 1	si	{S_0 }	alors	{«free » = 0, S_1}
Règle 2	si	{S_0}	alors	{«free» = 1, S_2}
Règle 3	si	{S_2}	alors	{«linguistics »=0, S_3}
Règle 4	si	{S_2}	alors	{«linguistics »=1, S_4}
Règle 5	si	{S_1, S_4}	alors	{S_5}

TAB. 1 –*Règles générées à partir du graphe de la figure 2*

(b) Représentation booléenne des règles générées : cet ensemble de règles est représenté par CelFact, CelRule, R_E et R_S dans l'automate cellulaire.

Fait n° i	
1	S_0
2	free=0
3	S_1
4	free =1
5	S_2
6	linguistics=0
7	S_3
8	linguistics=1
9	S_4
10	S_5

EF	IF	SF
1	0	0
0	1	0
0	0	0
0	1	0
0	0	0
0	1	0
0	0	0
0	1	0
0	0	0
0	0	0
CelFact		

Règle n°j	
1	R1
2	R2
3	R3
4	R4
5	R5

ER	IR	SR
0	1	0
0	1	0
0	1	0
0	1	0
0	1	0
CelRule		

TAB 2. – *Représentation cellulaire des partitions* $P_0 = \{s_0\}$, $P_1 = \{s_1, s_2\}$, $P_2 = \{s_3, s_4\}$, *et* $P_3 = \{s_5\}$.

$\mathbf{R_E}$	R_1	R_2	R_3	R_4	R_5
S_0	1	1	0	0	0
«free »= 0	0	0	0	0	0
S_1	0	0	0	0	1
« free »=1	0	0	0	0	0
S_2	0	0	1	1	0
« linguistics»=0	0	0	0	0	0
S_3	0	0	0	0	0
« linguistics»=1	0	0	0	0	0
S_4	0	0	0	0	1
S_5	0	0	0	0	0

$\mathbf{R_S}$	R_1	R_2	R_3	R_4	R_5
S_0	0	0	0	0	0
« free »= 0	1	0	0	0	0
S_1	1	0	0	0	0
« free»=1	0	1	0	0	0
S_2	0	1	0	0	0
« linguistics»=0	0	0	1	0	0
S_3	0	0	1	0	0
« linguistics»=1	0	0	0	1	0
S_4	0	0	0	1	0
S_5	0	0	0	0	1

TAB. 3 – *Les matrices d'incidence d'entrées/sorties pour la figure 2*

(c) Inférence : les Tableaux 2 et 3 représentent la configuration initiale G(0). Le chainage avant va permettre au modèle de passer de cette configuration aux configurations G(1), G(2)…G(i). L'inférence s'arrête après stabilisation avec une configuration finale. Le tableau 4 présente le modèle cellulaire final correspondant à la figure 2.

Fait		EF	IF	SF
1	free=0	0	1	0
2	free=1	0	1	0
3	linguistics=0	0	1	0
4	linguitics=1	0	1	0
5	S3 : class=spam	0	1	0
6	S5 : class=legitimate	0	1	0
CelFact				

Règle		ER	IR	SR
1	R1	0	1	0
2	R2	0	1	0
3	R3	0	1	0
CelRule				

$\mathbf{R_E}$	R_1	R_2	R_3
«free » = 0	1	0	0
« free »=1	0	1	1
« linguistics»=0	0	1	0
« linguistics»=1	0	0	1
S_3 : class=spam	0	0	0
S_5 :class=legitimate	0	0	0

$\mathbf{R_S}$	R_1	R_2	R_3
«free »= 0	0	0	0
« free »=1	0	0	0
« linguistics»=0	0	0	0
« linguistics»=1	0	0	0
S_3 :class=spam	0	1	0
S_5 :class=legitimate	1	0	1

TAB.4 – *Un extrait de la Configuration finale de l'automate*

Du modèle booléen représenté en tableau 4, des règles de classification sont déduites, par exemple la règle R1 se lit : *Si « free =0 » Alors légitime (classe majoritaire de S5).*

2.5 Classification

Cette étape utilise comme entrée le modèle élaboré depuis la phase d'apprentissage. Nous résumons les principales étapes comme suit :

1. Charger le modèle booléen : CelFact[4], CelRule, R_E, et R_S
2. Prétraiter le nouveau courriel et calculer sa représentation vectorielle : soit V.
3. Initialiser la base de faits CelFact :
   Pour chaque terme j dans CelFact faire
   Si terme $_j$ présent dans V **Alors** EF(terme$_j$ =1) ← 1
   Sinon EF(terme$_j$ =0)← 1 **Fin Si** **Fin Pour**
4. Appliquer la fonction de transition globale ∇=δfact°δrule
5. **Si** (EF (class=spam) ==1) **Alors** le courriel est classifié spam
   Sinon (EF (class= legitimate) =1) le courriel est classifié légitime **FinSi.**

3 Etude Expérimentale et Résultats

Afin d'évaluer cette approche de classification cellulaire pour le filtrage de spam, nous avons entamé plusieurs expériences sur le corpus Ling-Spam[5]. Et en nous appuyant sur la validation croisée, et en suivant les travaux effectués dans ce domaine (Androutsopoulos et al, 2000), nous mesurons le rappel de la classe spam (r), la précision de la classe spam (p), la F-mesure de la classe spam (f1) et enfin l'exactitude (e).

Soient N(LL) : le nombre de courriels légitimes classifiés légitimes; N(SS) : le nombre de courriels spam classifiés spam; N(LS) : le nombre de courriels légitimes classifiés spam et N(SL) le nombre de courriels spam classifiés légitimes, nous avons alors :

$$p = \frac{N(SS)}{N(SS)+N(LS)} \qquad r = \frac{N(SS)}{N(SS)+N(SL)} \qquad f1 = \frac{2*pr}{p+r} \qquad e = \frac{N(SS)+N(LL)}{N(SS)+N(LL)+N(SL)+N(LS)}$$

.

A travers ces expériences, nous avons constaté que la qualité de prédiction devient de plus en plus meilleure en termes de précision, rappel et exactitude à partir de 300 termes lorsqu'il y a racinisation des mots et élimination des mots vides avec les trois mesures de sélection. Nous avons constaté aussi que la mesure de sélection GI amène à une meilleure qualité de prédiction que les deux autres mesures. Enfin, nous avons constaté que l'approche proposée se stabilise à partir de 500 termes sélectionnés avec la fonction GI et amène à une qualité de prédiction intéressante : une précision = 98,1%, un rappel =84,2%, et une exactitude = 97.1%. Afin de comparer ces résultats avec les autres techniques, nous incluons les résultats des expériences réalisées sur le corpus LingSpam avec deux autres classifiers proposés dans la littérature :

NB : nous incluons les meilleurs résultats déclarés par (Androutsopoulos et al, 2000) pour l'approche bayésienne naïve. En utilisant une version lemmatisée du corpus LingSpam et l'information mutuelle (MI) comme métrique pour la sélection des termes, Ils trouvent que le classifier NB fonctionne de manière optimale avec un ensemble de termes égale à 100.

K-NN : à partir du même papier, nous incluons les meilleurs résultats déclarés pour une variante de l'algorithme du plus proche voisin. Comme dans le cas de NB, ils effectuent la sélection des termes en se basant sur la métrique MI, et obtiennent des résultats optimaux avec un plus petit nombre de termes (égale à 50) pour k=1 et K=2.

[4] Le EF de chaque cellule est initialisé à 0: aucun fait n'est établit
[5] Ling-Spam Corpus, http ://www.aueb.gr/users/ion/data/lingspam*public:tar:gz*

Classifier	Mesure de Sélection	Nbr Termes	Spam Précision	Spam Rappel	Spam F-mesure	Exacti-tude
NB	MI	100	99,02	82,35	89,92	96,926
TiMBL(1)	MI	50	95,92	85,27	90,28	96,890
TiMBN(2)	MI	50	97,10	83,19	89,61	96,753
CASD	**GI**	**500**	**98,10**	**84,20**	**90,62**	**97,100**

TAB. 5– *Résultats avec les meilleures configurations sur le corpus LingSpam*

Le tableau 5 présente les meilleurs résultats obtenus en utilisant notre classifieur CASD à côtés de ceux publiés précédemment et cité ci-dessus. Les résultats indiquent une amélioration des performances lorsque le classificateur CASD est utilisé. Il est clair qu'il surpasse NB et Knn en exactitude et en F-mesure.

4 Conclusion

Dans ce papier, nous avons proposé l'utilisation d'une nouvelle approche basée sur un automate cellulaire pour la détection de spam. Nos premières évaluations indiquent que l'approche proposée est intéressante. Bien que les résultats obtenus soient encourageants, beaucoup de points sont susceptibles d'être étudiés dans le cadre de travaux futurs. Nous devons approfondir nos expériences pour bien discerner les avantages et faiblesses de cette approche, cela nous permettra de mieux comprendre les situations où l'approche deviendra plus intéressante. Nous devons aussi mener une comparaison plus poussée de cette approche avec d'autres algorithmes d'apprentissage utilisés dans le filtrage de spam et en considérant d'autres corpus comme Spam Assassin et critères d'évaluation.

Références

Androutsopoulos, I., Koutsias, J (2000), "An Evaluation of Naive Bayesian Networks.", In: Machine Learning in the New Information Age. Barcelona Spain 9-17

Atmani B. et Beldjilali B. (2007). Knowledge Discovery in Database: Induction Graph and Cellular Automaton, Computing and Informatics Journal, 26, 171-197.

Guzella T. S., Caminhas W. M. (2009), "Review: A review of machine learning approaches to spam filtering", Expert Systems with Applications, 36(7), 10206-10222.

Salton G., Wong A., Yang C. S., (1975) "A vector space model for automatic indexing", Communications of the ACM, 18(11), 613-620

Sanz . E.P, Hidalgo J M G, Perez J C C, (2008), "Email spam filtering", in Zelkowitz M. (Ed.), Advances in computers, vol.74, 45-114.

Sebastiani F. (2002) Machine Learning in Automated Text Categorization, ACM Computing.

Zighed. (2000). Graphe d'induction: Apprentissage et data mining. HERMES.

Summary

Spam, also known as junk mail quickly became a major problem on the Internet. To address this growing burden of this type of spam, we propose the use of a supervised classification based on Boolean cellular automata to automatically classify incoming emails as spam or legitimate. To evaluate the performance of this new approach, we conduct a series of experiments on the corpus LingSpam.

Index des auteurs