Université Lumière Lyon 2

THÈSE pour obtenir le grade de DOCTEUR en INFORMATIQUE

présentée et soutenue publiquement par

Jérémie Clech

Le 2 mars 2004

Contribution Méthodologique à la Fouille de Données Complexes

préparée au sein du Laboratoire ERIC

Faculté des Sciences Economiques et de Gestion

Humanités

Informatique

sous la direction de Djamel A. ZIGHED

Copyright CLECH Jérémie et Université Lumière - Lyon 2 - 2004. Ce document est protégé en vertu de la loi du droit d'auteur.

COMPOSITION DU JURY M. Henri BRIAND, Examinateur, Professeur, Université de Nantes M. Mohand-Saïd HACID, Examinateur, Professeur, Université Lyon 1 M. Yves KODRATOFF, Examinateur, Professeur, Université Paris-Sud M. Ludovic LEBART, Rapporteur, Professeur,, ENST de Paris M. Pierre-François MARTEAU, Rapporteur, Professeur, Université de Bretagne Sud M. Djamel Abdelkader ZIGHED, Directeur de thèse, Professeur, Université Lyon 2

Table des matières

Remerciements

Clech_j_remerciements

Copyright CLECH Jérémie et Université Lumière - Lyon 2 - 2004.Ce document est protégé

Copyright CLECH Jérémie et Université Lumière - Lyon 2 - 2004.Ce document est protégé

Notations

Clech_j_notations

en vertu de la loi du droit d'auteur.

en vertu de la loi du droit d'auteur.

INTRODUCTION générale

Clech_j_intro

Copyright CLECH Jérémie et Université Lumière - Lyon 2 - 2004.Ce document est protégé

Copyright CLECH Jérémie et Université Lumière - Lyon 2 - 2004.Ce document est protégé

Corps de thèse

Chapitre 1. Représentation des données complexes pour la fouille

Clech_j_chapitre01

Chapitre 2. Apprentissage à base d'instances

Clech_j_chapitre02

Chapitre 3. Visualisation des données complexes

Clech_j_chapitre03

en vertu de la loi du droit d'auteur.

Chapitre 4. Recherche d'information au sein de données complexes

Clech_j_chapitre04

Chapitre 5. Applications sur des données complexes

Clech_j_chapitre05

en vertu de la loi du droit d'auteur.

Conclusion générale

Clech_j_conclu

Copyright CLECH Jérémie et Université Lumière - Lyon 2 - 2004.Ce document est protégé

Copyright CLECH Jérémie et Université Lumière - Lyon 2 - 2004.Ce document est protégé

Annexe A

Les 14 caractéristiques de texture définies par HARALICK *et al.* (1973) sont résumées dans le tableau ci-dessous :

en vertu de la loi du droit d'auteur.

Caractéristiques de texture	Formule
Moment angulaire second	$f_1 = \sum_{i=1}^{N_g}\sum_{j=1}^{N_g} \tilde{P}(i,j)^2$
Contraste	$f_2 = \sum_{n=0}^{N_g-1} \left(\sum_{\substack{i=1 \\ \lvert i-j \rvert = n}}^{N_g}\sum_{j=1}^{N_g} \tilde{P}(i,j) \right)$
Corrélation	$f_3 = \frac{\sum_{i=1}^{N_g}\sum_{j=1}^{N_g} (i,j)\tilde{P}(i,j) - \mu_x \mu_y}{\sigma_x \sigma_y}$
Variance ou somme des carrés	$f_4 = \sum_{i=1}^{N_g}\sum_{j=1}^{N_g} (i-j)^2 \tilde{P}(i,j)$
Moment différeur inverse	$f_5 = \sum_{i=1}^{N_g}\sum_{j=1}^{N_g} \frac{1}{1+(i-j)^2}\tilde{P}(i,j)$
Moyenne des sommes	$f_6 = \sum_{i=2}^{2N_g} i\, \tilde{P}_{x+y}(i)$
Variance des sommes	$f_7 = \sum_{i=2}^{2N_g} (i - f_8)^2 \tilde{P}_{x+y}(i)$
Entropie des sommes	$f_8 = \sum_{i=2}^{2N_g} \tilde{P}_{x+y}(i) \log\left(\tilde{P}_{x+y}(i)\right)$
Entropie	$f_9 = -\sum_{i=1}^{N_g}\sum_{j=1}^{N_g} \tilde{P}(i,j) \log\left(\tilde{P}(i,j)\right)$
Variance des différences	f_{10} = variance de $\tilde{P}_{x-y}$

en vertu de la loi du droit d'auteur.

Caractéristiques de texture	Formule
Entropie des différences	$f_{11} = -\sum_{i=2}^{2N_g} \tilde{P}_{x+y}(i).\log\left(\tilde{P}_{x+y}(i)\right)$
Mesures de corrélation 1 et 2	Soient : • $HXY = -\sum_{i=1}^{N_g}\sum_{j=1}^{N_g} \tilde{P}(i,j).\log\tilde{P}(i,j)$; • HX et HY sont les entropies de $\tilde{P}_x$ et $\tilde{P}_y$, • $HXY1 = -\sum_{i=1}^{N_g}\sum_{j=1}^{N_g} \tilde{P}(i,j).\log\left(\tilde{P}_x(i).\tilde{P}_y(j)\right)$; • $HXY2 = -\sum_{i=1}^{N_g}\sum_{j=1}^{N_g} \tilde{P}_x(i).\tilde{P}_y(j).\log\left(\tilde{P}_x(i).\tilde{P}_y(j)\right)$. Alors : $f_{12} = \frac{HXY - HXY1}{\max(HX, HY)}$ et $f_{13} = \sqrt{1-e^{-2(HXY2-HXY)}}$
Coefficient maximal de corrélation	$f_{14} = \sqrt{\text{deuxième valeur propre du Q}}$, où $Q(i,j) = \sum_{k=1}^{N_g} \frac{\tilde{P}(i,k)\tilde{P}(j,k)}{\tilde{P}_x(i)\tilde{P}_y(k)}$

Copyright CLECH Jérémie et Université Lumière - Lyon 2 - 2004.Ce document est protégé

Annexe B

en vertu de la loi du droit d'auteur.

		LAT C.V.		LAT classifier applied on LM		LAT classifier applied on SDA		LM C.V.		SDA C.V.	
		3-NN	C4.5	3-NN	C4.5	3-NN	C4.5	3-NN	C4.5	3-NN	C4.5
e_1	ρ	100%	93%	100%	93%	100%	91%	100%	83%	100%	100%
	π	93%	96%	96%	100%	100%	84%	96%	87%	100%	88%
e_2	ρ	90%	97%	100%	94%	100%	97%	100%	99%	100%	96%
	π	96%	100%	100%	100%	100%	97%	100%	99%	100%	100%
e_3	ρ	100%	0%	95%	0%	97%	0%	96%	96%	100%	99%
	π	100%	0%	82%	0%	100%	0%	85%	88%	100%	99%
e_4	ρ	87%	100%	100%	84%	100%	81%	100%	11%	100%	99%
	π	100%	32%	100%	12%	100%	28%	100%	17%	100%	100%
e_5	ρ	95%	90%	100%	85%	100%	95%	90%	91%	100%	90%
	π	90%	81%	87%	78%	96%	100%	91%	100%	96%	100%
e_6	ρ	95%	70%	95%	0%	100%	17%	91%	30%	50%	0%
	π	95%	86%	85%	0%	50%	100%	88%	11%	75%	0%
e_7	ρ	77%	77%	90%	77%	100%	100%	90%	79%	96%	90%
	π	85%	80%	74%	51%	80%	41%	85%	70%	95%	88%
e_8	ρ	90%	90%	60%	84%	84%	5%	79%	60%	90%	90%
	π	81%	97%	78%	82%	95%	67%	89%	96%	96%	96%
e_9	ρ	80%	0%	95%	0%	100%	0%	95%	52%	100%	0%
	π	67%	0%	87%	0%	100%	0%	87%	11%	91%	0%
e_{10}	ρ	91%	79%	78%	29%	70%	13%	91%	80%	91%	96%
	π	96%	96%	96%	97%	100%	100%	96%	64%	84%	27%
ρ	$\rho' = \pi'$	92%	83%	92%	51%	95%	50%	95%	76%	97%	83%
M	ρ^M	91%	70%	89%	52%	92%	62%	94%	68%	90%	77%
	π^M	91%	67%	90%	50%	96%	58%	90%	71%	95%	70%

Rappel et Précision du modèle anglais (LAT) décrit par 100 mots :

en vertu de la loi du droit d'auteur.

		LAT C.V.		LAT classifier applied on LM		LAT classifier applied on SDA		LM C.V.		SDA C.V.	
		3 NN	C4.5	3 NN	C4.5	3 NN	C4.5	3 NN	C4.5	3 NN	C4.5
w_1	p	100%	96%	100%	96%	100%	100%	100%	100%	100%	96%
	π	100%	96%	97%	96%	100%	100%	100%	97%	100%	85%
w_2	p	99%	99%	100%	99%	100%	98%	100%	100%	100%	94%
	π	98%	99%	100%	98%	100%	97%	99%	100%	100%	98%
w_3	p	100%	0%	100%	0%	100%	0%	100%	96%	100%	99%
	π	100%	0%	52%	0%	100%	0%	52%	52%	100%	99%
w_4	p	94%	98%	100%	94%	100%	49%	100%	92%	100%	100%
	π	94%	94%	95%	98%	100%	79%	100%	99%	100%	100%
w_5	p	94%	94%	100%	100%	98%	96%	97%	100%	100%	96%
	π	98%	98%	92%	97%	99%	100%	97%	100%	100%	96%
w_6	p	99%	0%	98%	0%	100%	0%	100%	91%	0%	0%
	π	99%	98%	52%	0%	98%	100%	98%	98%	0%	0%
w_7	p	92%	99%	52%	97%	96%	98%	96%	92%	96%	92%
	π	92%	92%	97%	49%	99%	96%	92%	89%	88%	97%
w_8	p	99%	99%	99%	91%	99%	9%	97%	90%	96%	99%
	π	99%	99%	99%	94%	100%	90%	97%	98%	93%	97%
w_9	p	92%	0%	0%	0%	100%	0%	96%	29%	100%	0%
	π	95%	0%	95%	0%	95%	0%	9.1%	46%	100%	0%
w_{10}	p	94%	99%	91%	98%	91%	0%	90%	97%	9.1%	100%
	π	98%	93%	99%	90%	100%	0%	99%	94%	88%	96%
μ	$p^\mu - \pi^\mu$	99%	94%	91%	98%	96%	9%	95%	94%	98%	0%
M	p^M	99%	90%	91%	90%	94%	9%	96%	89%	94%	89%
	π^M	92%	90%	91%	97%	91%	0%	96%	99%	94%	98%

Rappel et Précision du modèle anglais (LAT) décrit par 200 4-grammes :

Copyright CLECH Jérémie et Université Lumière - Lyon 2 - 2004.Ce document est protégé

Bibliographie

AAS, K. et EIKVIL, L., *Text Categorization: a survey*, Norwegian Computing Center. **1999**.

ACR, **[en-ligne]** Disponible sur http://www.acr.org (consulté le octobre 2002).

AFNOR, *Information et documentation. Principes généraux pour l'indexation des documents*: 10-11. **1993**.

AIGRAIN, P., ZHANG, H. et PETKOVIC, D., Content-based representation and retrieval of visual media: A state-of-the-art review. *Multimedia Tools and Applications* 3(3): 176-202. **1996**.

AITCHISON, T. M., HALL, A. M., LAVELLE, K. H. et TRACY, J. M., *Comparative Evaluation of Index Languages*. London, England, Project INSPEC, Institute of Electrical Engineers. **1970**.

ALTAVISTA, **[en-ligne]** Disponible sur http://www.altavista.com (consulté le 27 août 2003).

AMINI, M.-R., *Apprentissage automatique et Recherche d'information : application à l'extraction d'information de surface et au résumé de texte*. Thèse de doctorat. Paris, France, Université Paris 6. **2001**

ANAES, **[en-ligne]** Disponible sur http://www.anaes.fr (consulté le octobre 2002).

APTE, C., DAMERAU, F. J. et WEISS, S. M., Towards Language-Independent Automated Learning of Text Categorization Models. *Proceedings of the 17th Annual ACM SIGIR*

en vertu de la loi du droit d'auteur.

Conference, Dublin, IE, Springer Verlag, Heidelberg, DE: 23-30. **1994**.

BARBA, D., *Traitement numérique d'images avec critère psychovisuel de qualité*. Thèse d'état, Université Paris 6. **1981**

BARTHÉLÉMY, J.-P. et GUÉNOCHE, A., *Les arbres et les représentations des proximités*. Paris, Masson. **1988**.

BARTHÉLÉMY, J.-P. et LUONG, X., Représentations arborées de mesures de dissimilarité. *Statistique et Analyse de Données* 11: 20-41. **1986**.

BARTHÉLÉMY, J.-P. et LUONG, X., Représenter les données textuelles par les arbres phylogénétiques. *4th International Conference on Textual Data Statistical Analysis*, Nice: 49-70. **1998**.

BAZSCALICZA, M. et NAÏM, P., *Data mining pour le Web - Profiling - Filtrage collaboratif - Personnalisation client*. Paris, Eyrolles. **2001**.

BELLMAN, R. E., *Adaptive Control Processes*, Princeton University Press. **1961**.

BENTLEY, J. L., K-d trees for semidynamic points set. *Proc. of ACM Sympos . Comput. Geom*: 187-197. **1990**.

BOURIGAULT, D., JACQUEMIN, C. et L'HOMME, M.-C., Eds. *Recent Advances in Computational Terminology*. Amsterdam. Les Pays-Bas, John Benjamins. **2001**.

BOWMAN, C. M., DANZIG, P. B., HARDY, D. R., MANBER, U. et SCHWARTZ, M. F., The harvest information discovery and access system. *Computer Networks and ISDN Systems* 28(1-2): 119-127. **1995**.

BREIMAN, L., Bagging predictors. *Machine Learning* 24: 123-140. **1996**.

BREIMAN, L., FRIEDMAN, J. H., OLSHEN, R. A. et STONE, C. J., *Classification and regression trees*. Belmont, CA, Wadsworth International Group. **1984**.

BRILL, E., Some Advance in Transformation Based Part of Speech Tagging. *Proceedings of the 12th National Conference on Artificial Intelligence*. **1994**.

BRIN, S. et PAGE, L., The anatomy of a large-scale hypertextual web search engine. *Computer Networks and ISDN Systems* 30(1-7): 107-117. **1998**.

BRODLEY, C. E. et FRIEDL, M. A., Identifying and Eliminating Mislabeled Training Instances. *Proceedings of the 13th National Conference on Artificial Intelligence*, Portland, AAI Press: 799-805. **1996**.

BRODLEY, C. E. et FRIEDL, M. A., Identifying mislabeled training data. *Journal of Artificial Intelligence Research* 11: 131-167. **1999**.

BROWN, P. F., DELLA PIETRA, V. J., DESOUZA, P. V., LAI, J. C. et MERCER, R. L., Class-based n-gram models of natural language. *Computational Linguistics* 18(4): 467-479. **1992**.

CALLON, M., COURTIAL, J.-P., TURNER, W. A. et BAUIN, S., From Translation to Problematic Networks: An Introduction to Co-Word Analysis. *Social Science Information* 22: 191-235. **1983**.

CAVNAR, W. B. et TRENKLE, J. M., N-Gram-Based Text Categorization. *Proceedings of SDAIR-94, 3rd Annual Symposium on Document Analysis and Information Retrieval*, Las Vegas, US: 161-175. **1994**.

CHAKRABARTI, S., DOM, B. E., GIBSON, D., KLEINBERG, J. M., KUMAR, S. R., RAGHAVAN,

en vertu de la loi du droit d'auteur.

P., RAJAGOPALAN, S. et TOMKINS, A., Mining the link structure of the World Wide Web. *IEEE Computer* 32: 60-67. **1999**.

CHANDON, J.-L. et PINSON, S., *Analyse typologique - Théories et applications*. Paris, Masson. **1981**.

CHANG, C. et HSU, C., Customizable multi-engine search tool with clustering. *Computer Network and ISDN Systems* 29(8-13): 1217-1224. **1997**.

CHAVENT, M., GUINOT, C., LECHEVALLIER, Y. et TENENHAUS, M., Méthodes divises de classification et segmentation non supervisée : recherche d'une typologie de la peau humaine saine. *Statistique Appliquée* XLVII: 87-99. **1999**.

CIAMPI, A., Constructing prediction trees from data: the RECPAM approach. *Proceedings of the Prague '91 Summer School of Computational Aspects of Model Choice*, Heidelberg, Physica-Verlag: 105-152. **1991**.

CIAMPI, A., ZIGHED, D. A. et CLECH, J., Trees and Induction Graphs form Multivariate Response. *Principles of Data Mining and Knowledge Discovery, Proceedings of the 4th European Conference on Principles and Practice of Knowledge Discovery in Databases*, Lyon, France, Springer-Verlag Berlin Heidelberg, 1910: 359-366. **2000**.

CLECH, J. et HASSAS, S., Web Mining et Système Multi-Agents. *Tutoriel EGC 2003*, Lyon, France. **2003**.

CLECH, J., RAKOTOMALALA, R. et JALAM, R., Sélection multivariée de termes. *XXXVèmes Journées de Statistiques*, Lyon, France: 933-936. **2003a**.

CLECH, J. et ZIGHED, D. A., Data Mining et Analyse des CV : Une Expérience et des Perspectives. *Conférence EGC*, Lyon, France, Editions Lavoisier, 17: 189-200. **2003**.

CLECH, J. et ZIGHED, D. A., Une technique de ré-étiquetage dans un contexte de catégorisation de textes. *Document numérique - Fouille de Textes et Organisation de Documents*: à paraître. **2004**.

CLECH, J., ZIGHED, D. A. et BREMOND, A., Apport des techniques de Text Mining pour la définition de caractéristiques clefs d'une mammographie. *Journée De la Statistique*, Lyon, France, Cépaduès, 1: 183-192. **2003b**.

CLEF, **[en-ligne]** Disponible sur http://clef.iei.pi.cnr.it (consulté le juillet 2002).

CLEVERDON, C. W., MILLS, L. et KEEN, E. M., *Factors Determining the Performance of Indexing Systems*. Cranfield, England, Aslib Cranfield Research Project. **1966**.

COHEN, W. W. et SINGER, Y., Context-sensitive learning methods for text categorization. *Proceedings of the 19th Annual ACM SIGIR Conference*: 307-315. **1996**.

COOLEY, R., STRIVASTAVA, J. et MOBASHER, B., Web Mining: Information and Pattern Discovery on the World Wide Web",. *Proceedings of the 9th IEEE International Conference on Tools with Artificial Intelligence*, Newport Beach: 558-567. **1997**.

COURTIAL, J.-P., Construction des connaissances scientifiques, construction de soi et communication sociale. *Les Sciences de l'information : bibliométrie, scientométrie, infométrie*(2). **1995**.

COVER, T. M. et HART, P. E., Nearest neighbor pattern classification. *IEEE Transactions on Information Theory* 11: 21-27. **1967**.

DAY, W. H. E., Computational complexity of inferring phylogenies from dissimilarity measures. *Bulletin of Mathematical Biology* 49: 461-467. **1987**.

Copyright CLECH Jérémie et Université Lumière - Lyon 2 - 2004.Ce document est protégé

DE LOUPY, C., *Evaluation de l'Apport de Connaissances Linguistiques en Désambiguïsation Sémantique et Recherche Documentaire*. Thèse de doctorat. Laboratoire Informatique d'Avignon. Avignon, France, Université d'Avignon et des Pays de Vaucluse. **2000**

DE LOUPY, C., L'apport de connaissances linguistiques en recherche documentaire. *TALN 01*, Tours, France, 2: 129-143. **2001**.

DUDANI, S. A., The distance-weighted *k*-nearest-neighbor rule. *IEEE Transactions on Systems, Man and Cybernetics* 6(4): 325-327. **1976**.

DUFFOUX, A., BOUSSAID, O., LALLICH, S. et BENTAYEB, F., Fouille de données à partir de la structure de documents XML. *Conférence EGC*, Clermont-Ferrand, Editions Lavoisier: à paraître. **2004**.

DUMAIS, S. T., Improving the retrieval of information from external sources. *Behavior Research Methods, Instruments &Computers* 23: 229-236. **1991**.

ETZIONI, O., The World-Wide Web: Quagmire or Gold Mine ? *Communication of the ACM* 39(11): 65-68. **1996**.

FALOUTSOS, C., BARBER, R., FLICKNER, M., HAFNER, J., NIBLACK, W., PETKOVIC, D. et EQUITZ, W., Efficient and Effective Querying by Image Content. *Journal of Intelligent Information Systems* 3(3): 231-162. **1994**.

FARRADANE, J., RUSSEL, J. M. et YATES-MERCER, A., Problems in information retrieval. Logical jumps in the expression of information. *Information Storage and Retrieval* 9: 65-77. **1973**.

FAYYAD, U. et GRINSTEIN, G. G., Brief History of Related Fields. *Information Visualization in Data Mining and Knowledge Discovery*. U. FAYYAD, G. G. GRINSTEIN and A. WIERSE. San Diego, USA, Academic Press:1-17. **2002**.

FAYYAD, U., PIATETSKY-SHAPIRO, G. et SMYTH, P., The KDD process for extracting useful knowledge from volumes data. *Communication of the ACM* 39(11): 27-34. **1996**.

FEKETE, J.-D., Concepts Fondamentaux en Visualisation d'Information. *Atelier Visualisation et Extraction Adaptatives des Connaissances, Conférence EGC*, Lyon. **2003**.

FISHER, D. H., The use of multiple measurements in taxonomic problems. *Annals of Eugenics* 7: 179-188. **1936**.

FITCH, W. M. et MARGOLIASH, E., Construction of phylogenetic trees. *Science* 155: 279-284. **1967**.

FIX, E. et HODGES, J. L., *Discriminatory analysis - nonparametric discrimination: Consistency properties*. Randolph Field, Texas, Rapport n° 21-49-004, USAF School of Aviation Medicine. **1951**.

FOLEY, J., VAN DAM, A., FEINER, S. et HUGHES, J., *Computer Graphics: Principles and Practice*, Addison Weslay. **1995**.

FREUND, Y. et SCHAPIRE, R. E., A decision theoretic generalization of online learning and an application to boosting. *Proceedings of the 2nd European Conference on Computational Learning Theory*, Springer Verlag: 137-140. **1995**.

FUHR, N. et BUCKLEY, C., A Probabilistic Learning Approach for Document Indexing. *ACM Transactions on Information Systems*, 9: 223-248. **1991**.

 Copyright CLECH Jérémie et Université Lumière - Lyon 2 - 2004.Ce document est protégé

GILLI, Y., *Texte et fréquence*. Paris, Université de Besançon. **1988**.

GOOGLE, **[en-ligne]** Disponible sur http://www.google.com (consulté le 27 août 2003).

GORDON, A. D., *Classification*, CRC Press, 2nd. **1999**.

GRINSTEIN, G. G. et WARD, M. O., Introduction to Data Visualization. *Information Visualization in Data Mining and Knowledge Discovery*. U. FAYYAD, G. G. GRINSTEIN and A. WIERSE. San Diego, USA, Academic Press:21-45. **2002**.

GUARINO, N., Some Ontological Principles for Designing Upper Level Lexical. *Proceedings of the 1st International Conference on Lexical Resources and Evaluation*, Granada, Espagne. **1998**.

HAND, D., MANNILA, H. et SMYTH, P., *Principles of Data Mining*. London, England, MIT Press. **2001**.

HARALICK, R. M., SHANMUGAN, K. et DINSTEIN, I., Texture features for image classification. *IEEE Transactions Systems, Man and Cybernetics* 3: 610-621. **1973**.

HEALEY, C., *Effective visualization of large multidimensional dataset*. Department of Computer Science, University of British Colombia, Canada. **1996**

HERMAN, I., MELANCON, G. et MARSHALL, M. S., Graph Visualization and Navigation in Information Visualisation: a Survey. *IEEE Transactions on Visualization and Computer Graphics* 6(1): 24-43. **2000**.

HERMITAGE, *Système QBIC* **[en-ligne]** Disponible sur http://www.hermitagemuseum.org/fcgi-bin/db2www/qbicSearch.mac/qbic?selLang=English (consulté le **1995.**

HULL, D. A., Stemming Algorithms: A Case Study for Detailed Evaluation. *Journal of the American Society of Information Science* 47: 70-84. **1996**.

IRANI, M. et ANANDAN, P., Video indexing based on mosaic representations. *Proceedings of the IEEE* 86(5): 905-921. **1998**.

JACQUEMIN, C., *Spotting and Discovering Terms through Natural Language Processing*. London, England, The MIT Press. **2001**.

JALAM, R., *Apprentissage automatique et catégorisation de textes multilingues*. Thèse de doctorat. Lyon, France, Université Lumière Lyon2: 172. **2003**

JALAM, R., CLECH, J. et RAKOTOMALALA, R., Un cadre pour la catégorisation de texte multilingues. *7th International Conference on the Statistical Analysis of Textual Data*, Louvain-la-Neuve, Belgique: à paraître. **2004**.

JALAM, R. et TEYTAUD, O., *Kernel-based text categorization*. Lyon, France, Laboratoire ERIC, Université Lumière Lyon 2. **2000**.

JOHN, G. H., Robust decision trees: removing outliers from data. *Proceedings of the 1st International Conference on Knowledge Discovery and Data Mining*, Montréal, Québec, AAI Press: 174-179. **1995**.

KEIM, D. A. et KRIEGEL, H.-P., Visualization techniques for mining large databases: a comparison. *IEEE Transactions on Data Engineering* 8(6): 923-938. **1996**.

KIRA, K. et RENDELL, L. A., A Practical Approach to Feature Selection. *9th International Workshop on Machine Intelligence*, Aberdeen, Scotland, Morgan-Kaufman. **1992**.

KLEINBERG, J. M., Authoritative sources in a hyperlinked environment. *Journal of the*

en vertu de la loi du droit d'auteur.

ACM 46(5): 604-632. **1999**.

KODRATOFF, Y., Technical and Scientific Issues of KDD (or: Is KDD a Science?). *Algorithmic Learning Theory*, Springer Verlag, 997: 261-265. **1994**.

KOHAVI, R., A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection. *IJCAI*, Montréal: 1137-1145. **1995**.

KONONENKO, I., Estimation attributes: Analysis and Extensions of RELIEF. *Proceedings of the European Conference on Machine Learning*, Catana, Italy, Springer Verlag: 171-182. **1994**.

KOSALA, R. et BLOCKEEL, H., Web Mining Research: A Survey. *SIGKDD Explorations: Newsletter of the Special Interest Group on Knowledge Discovery & Data Mining, ACM* 2: 1-15. **2000**.

KRUSKAL, J. B., On the shortest spanning tree of a graph and the travelling salesman problem. *Proceedings of the American Mathematical Society*, 7: 45-50. **1956**.

KUO, Y.-H. et WONG, M. H., Web document classification based in hyperlinks and document semantics. *Workshop on Text and Web Mining (PRICAI)*: 44-51. **2000**.

LALLICH, S., MUHLENBACH, F. et ZIGHED, D. A., Improving classification by removing or relabeling mislabeled instances. *Proceedings of the 13th International Symposium on Methodologies for Intelligent Systems, Foundations of Intelligent Systems*, Lyon, France, Springer-Verlag: 5-15. **2002**.

LARGERON, C., *Reconnaissance des formes par relaxation : un modèle d'aide à la décision*. Thèse de doctorat, Université Lyon1. **1991**

LEBART, L., MORINEAU, A. et PIRON, M., *Statistique exploratoire multidimensionnelle*. Paris, Dunod, 3eme édition. **2000**.

LEBART, L. et SALEM, A., *Statisque textuelle*. Paris, Dunod. **1994**.

LEFÉBURE, R. et VENTURI, G., *Data Mining*. Paris, Eyrolles. **2001**.

LEFÈVRE, P., *La recherche d'information - du texte intégral au thésaurus*. Paris, Hermès Science. **2000**.

LESPINASSE, K., TREC, une conférence pour l'évaluation des systèmes de recherche d'information. *Documentaliste* 34(2): 107-109. **1997**.

LESPINASSE, K., KREMER, P., SCHIBLER, D. et SCHMITT, L., Evaluation des outils d'accès à l'information textuelle : les expériences américaine (TREC) et française (AMARYLLIS). *Langues* 2(2): 100-109. **1999**.

LEWIS, D. D. et RINGUETTE, M., Comparison of two learning algorithms for text categorization. *Proceedings of the 3rdAnnual Symposium on Document Analysis and Information Retrieval*. **1994**.

LIU, H. et MOTODA, H., *Feature selection for knowledge discovery and data mining*, Kluwer. **1998**.

LONCARNIC, S., A survey of shape analysis techniques. *Pattern Recognition* 31(8): 983-1001. **1998**.

LUHN, H. P., The automatic creation of literature abstracts. *IBM Journal of Research and Development* 2: 159-165. **1958**.

MAES, P., Agent that reduce work and information overload. *Communication of the ACM*

en vertu de la loi du droit d'auteur.

37(7): 30-40. **1994**.

MAHDI, W., *Macro-segmentation Sémantique des Documents Audiovisuels à l'aide des Indices Spatio-temporels*. Thèse de doctorat, Eole Centrale de Lyon: 163. **2001**

MARON, M. et KUHNS, J., On relevance, probabilistic indexing and information retrieval. *Journal of the ACM* 7: 216-243. **1960**.

MENESES, C. J. et GRINSTEIN, G. G., Research Issues in the Analysis and Visualisation of Massive Data Sets. *Information Visulization in Data Mining and Knowledge Discovery*. U. FAYYAD, G. G. GRINSTEIN and A. WIERSE. San Diego, USA, Academic Press:355-359. **2002**.

MILLER, E., SHEN, D., LIU, J. et C.NICHOLAS, Performance and Scalability of a Large-Scale N-gram Based Information Retrieval System. *Journal of Digital Information* 1. **1999**.

MILLER, G. A., BECKWITH, R., FELLBAUM, C., GROSS, D. et MILLER, K. J., Introduction to WordNet: An on-line lexical database. *Journal of Lexicography* 3: 235-244. **1990**.

MITCHELL, T. M., *Machine Learning*. New York, McGraw-Hill. **1997**.

MUHLENBACH, F., *Evaluation de la qualité de la représentation en fouille de données*. Thèse de doctorat. France, Laboratoire ERIC, Université Lumière Lyon 2. **2002**

O'CONNOR, J., Answer-passage Retrieval by Text Searching. *Journal of American Society for Information Science* 26: 171-239. **1980**.

PEARSON, K., On lines and planes of closest fit to systems of points in space. *Philosophical Magazine and Journal of Science* 6(2): 559-572. **1901**.

PORTER, M. F., An algorithm for suffix stripping. *Program* 14: 130-137. **1980**.

PREPARATA, F. et SHAMOS, M., *Computational Geometry An Introduction*. New-York, Springer. **1985**.

PRIM, R. C., Shortest connection networks and some generalizations. *The Bell System Technical Journal* 36: 1389-1401. **1957**.

PRYKE, A. N., *The Haiku Visualisation System* **[en-ligne]** Disponible sur http://www.cs.bham.ac.uk/~anp/haiku/ (consulté le 8 août 2003). **1996.**

QUINLAN, J. R., *C4.5: Programs for Machine Learning*. San Mateo, CA. **1993**.

RAFIEI, D. et MENDELZON, A., What is this page is known for? Computing web page reputations. *Computer Networks* 33: 823-835. **2000**.

RAJMAN, M. et LEBART, L., Similarités pour données textuelles. *4th International Conference on Statistical Analysis of Textual Data (JADT'98)*, Nice, France: 545-555. **1998**.

RIBEIRO, A. et FRESNO, V., A Multi Criteria Function to Concept Extration in HTML Environment. *Proceedings of the International Conference on Internet Computing*, Las Vegas (USA), 1: 1-5. **2001**.

RIBOLI, E. et KAAKS, P., The EPIC Project: rationale and study design. European Prospective Investigation into Cancer and Nutrition. *International Journal of Epidemiology* 26(1): S6-14. **1997**.

RO, J. S., An Evaluation of the Applicability of Ranking Algorithms to Improve the Effictiveness of Full-Text Retrieval. *Journal of American Society for Information*

Copyright CLECH Jérémie et Université Lumière - Lyon 2 - 2004.Ce document est protégé

Science 39: 73-78. **1988**.

ROBERTSON, S. E., The probabilistic character of relevance. *Information Processing and Management* 13: 247-251. **1977**.

ROBERTSON, S. E. et SPARCK-JONES, K., Relevance weighting of search terms. *Journal of American Society for Information Science* 27(3): 129-146. **1976**.

ROCCHIO, J. J., Relevance feedback information retrieval. *The Smart Retrieval system - experiments in automatic document processing*. S. G. Englewood Cliffs, NJ, Prentice-Hall:313-323. **1971**.

RUI, Y., HUANG, T. S. et CHANG, S.-F., Image retrieval: current techniques, promising directions and open issues. *Journal of Visual Communication and Image Representation* 10(4): 39-62. **1999**.

SAITOU, N. et NEI, M., Neighbor-joining method. *Mol. Bio. Evol.* 4: 406-425. **1987**.

SALTON, G., *Automatic Information Organization and Retrieval*. New-York, USA, Mc Graw-Hill. **1968**.

SALTON, G., *Automatic Text Processing: The Transformation, Analysis, and Retrieval of Information by Computer*. Pennsylvania, Addison-Wesley. **1989**.

SALTON, G., ALLAN, J. et BUCKLEY, C., Approaches to Passage Retrieval in Full Text Information Systems. *Proceedings of the sixteenth annual international ACM SIGIR conference on Research and development in information retrieval*, ACM Press: 49-58. **1993**.

SALTON, G. et MCGILL, M., *Introduction to Modern Information Retrieval*. New York, McGraw-Hill. **1983**.

SALTON, G., SINGHAL, A., BUCKLEY, C. et MITRA, M., Automatic Text Decomposition Using Text Segments and Text Themes. *Proceedings of the Seventh ACM Conference on Hypertext*, Washington D. C. **1996**.

SAMIER, H. et SANDOVAL, V., *La veille stragique sur l'internet*. Paris, France, Hermes Sciences Publication. **2002**.

SAPORTA, G., *Probabilités, Analyse des données et Statistique*. Paris, France, Edition Technip. **1990**.

SATTAH, S. et TVERSKY, A., Additive similarity trees. *Psychometrika* 42: 319-345. **1977**.

SAVOY, J., Report on CLEF-2002 Experiements: Combining multiple sources of evidence, University of Neuchatel. **2002**.

SCHMID, H., Probabilistic Part-of-Speech Tagging Using Decision Trees. *International Conference on New Methods in Language Processing*, Manchester, UK. **1994**.

SCOTT, S. et MATWIN, S., Feature Engineering for Text Classification. *Proceedings of the 16th International Conference on Machine Learning*, San Francisco: 379-388. **1999**.

SCUTURICI, M., *Contribution aux techniques orientées objet de gestion des séquences vidéo pour les serveurs Web*. Thèse de doctorat. Laboratoire ERIC, Institut National des Sciences Appliquées de Lyon. **2002**

SCUTURICI, M., CLECH, J., SCUTURICI, V. M. et ZIGHED, D. A., Topological Representation Model for Image Database Query. *Journal of Experimental & Theoretical Artificial*

 Copyright CLECH Jérémie et Université Lumière - Lyon 2 - 2004.Ce document est protégé

Intelligence: à paraître. **2003a**.

SCUTURICI, M., CLECH, J., SCUTURICI, V. M. et ZIGHED, D. A., Modèle topologique pour l'interrogation des bases d'images. *Conférence EGC*, Clermont-Ferrand, Editions Lavoisier: à paraître. **2004**.

SCUTURICI, M., CLECH, J. et ZIGHED, D. A., Topological Query in Image Databases. *8th Iberoamerican Congress on Pattern Recognition*, Havana, Cuba: 144-151. **2003b**.

SEBASTIANI, F., Machine learning in automated text categorization. *ACM Computing Surveys* 340(1): 1-47. **2002**.

SEBBAN, M., *Modèles théoriques en reconnaissance de formes et architexture hybride pour machine perceptive*. Thèse de doctorat. France, Université Claude Bernard Lyon 1. **1996**

SHANNON, C., The Mathematical Theory of Communication. *Bell System Technical Journal* 27: 379-423 and 623-656. **1948**.

SHANNON, C. E. et WEAVER, W., *The mathematical theory of communication*, University of Illinois Press. **1949**.

SIDHOM, S., *Plate-forme d'analyse morpho-syntaxique pour l'indexation automatique et la recherche d'information : de l'écrit vers la gestion des connaissances*. Thèse de doctorat. Laboratoire SII, ENSSIB. Lyon, France, Université Claude Bernard Lyon 1. **2002**

SNOWBALL, **[en-ligne]** Disponible sur http://snowball.tartarus.org (consulté le octobre 2002).

SPARCK-JONES, K., A statistical interpretation of term specificity and its applications in retrieval. *Journal of Documentation* 28(1): 11-21. **1972**.

STONE, M., Cross-validatory choice and assessment of statistical predictions. *Journal of the Royal Statistical Society* B: 111-147. **1974**.

SYSTRAN, **[en-ligne]** Disponible sur http://babelfish.altavista.com/ (consulté le juin 2002).

TELECOMITALIALAB, *Site officiel MPEG* **[en-ligne]** Disponible sur http://mpeg.telecomitalialab.com (consulté le septembre 2002).

TEOMA, **[en-ligne]** Disponible sur http://www.teoma.com (consulté le 27 août 2003).

TEYTAUD, O. et JALAM, R., Kernel based text categorization. *12th International Joint Conference on Neural Networks*, Los Alamitos, US, IEEE Computer Society Press: 1891-1896. **2001**.

TREC, *Text REtrieval Conference* **[en-ligne]** Disponible sur http://trec.nist.gov/ (consulté le septembre 2003). **2000.**

TREND, D., **[en-ligne]** Disponible sur http://www.disktrend.com/ (consulté le 2002). **1999.**

TSAPARAS, P., *Nearest neighbor search in multidimensional spaces*, Rapport n° 31902, Dept. of Computer Science, University of Toronto. **1999**.

TUCERYAN, M., Moment-Based Texture Segmentation. *Pattern Recognition Letters* 15: 659-668. **1994**.

VAN RIJSBERGEN, C. J., *Information Retrieval, 2nd edition*, Dept. of Computer Science,

en vertu de la loi du droit d'auteur.

University of Glasgow. **1979**.

VANDENDORPE, C., Au-delà de la phrase : la grammaire du texte. *Pour un nouvellement enseignement de la grammaire*. S. CHARTRAND. Montréal, Logiques:83-105. **1995**.

WASHINGTON, *Ground Truth Database by the University of Washington* **[en-ligne]** Disponible sur http://www.cs.washington.edu/research/imagedatabase (consulté le 12 mai 2003). **1999.**

WILSON, D., Asymptotic properties of nearest neighbors rules using edited data. *IEEE Transactions on Systems, Man and Cybernetivs* 2: 408-421. **1972**.

WOLPERT, D., Stacked generalisation. *Neural Networks* 5: 241-259. **1992**.

YANG, Y., An evaluation of statistical approaches to text categorization. *Journal of Information Retrieval* 1(1/2): 69-90. **1999**.

YANG, Y. et LIU, X., A re-examination of text categorization methods. *Proceedings of the 22nd Annual ACM SIGIR Conference*: 42-49. **1999**.

YANG, Y. et PEDERSEN, J. O., A comparative Study on Feature Selection in Text Categorization. *Proceedings of the 14th International Conference on Machine Learning*: 412-420. **1997**.

YANG, Y., SLATTERY, S. et GHANI, R., A study of approaches to hypertext categorization. *Journal of Intelligent Information Systems* 18(2-3): 219-241. **2002**.

ZHANG, H., LOW, C. Y., SMOLIAR, S. W. et WU, J. H., Video Parsing, Retrieval and Browsing: An integrated and Content-Based Solution. *Proceedings ACM Multimedia*, San Francisco, US: 15-24. **1995**.

ZIGHED, D. A. et RAKOTOMALALA, R., *Graphes d'induction. Apprentissage et Data Mining*. Paris, Hermes Science Publication. **2000**.

ZIGHED, D. A. et RAKOTOMALALA, R., *Extraction de connaissances à partir de données (ECD)*. Techniques de l'Ingénieur. HA. **2002**.

ZIGHED, D. A., TOUNISSOUX, D., AURAY, J. P. et LARGERON, C., Discrimination basée sur un critère d'homogénéité locale. *T.S.I.* 3: 213-220. **1990**.

en vertu de la loi du droit d'auteur.