EXTRACTION DE CONNAISSANCES À PARTIR DES DONNEES ET DES TEXTES
(ECD ET ECT "DATA & TEXT MINING")

YVES KODRATOFF[1]

(soumis à "Revue Maths. et Sciences Humaines")

RESUME : *Cet article définit l'Extraction de Connaissances à partir de Textes (ECT, "Text Mining") comme un descendant direct de l'Extraction de Connaissances à partir de Données (ECD, "Data Mining"). Le point fondamental qui distingue ces domaines de ceux dont ils sont issus est la reconnaissance du raisonnement inductif en tant que mode normal de raisonnement, tout autant que les raisonnements plus classiques, déductifs et abductifs. Le raisonnement analogique n'est pas encore, de fait, utilisé en ECD ni en ECT, mais rien ne s'oppose en principe à ce que ce type de raisonnement soit inclus à terme dans les mécanismes utilisés par ces nouveaux domaines de recherche.*

En conséquence, les diverses mesures effectuées sur les données ne le sont plus dans le but d'appliquer ou de valider un modèle existant, mais dans celui de construire un nouveau modèle. Cette attitude a des conséquences extrêmement profondes sur la définition de la qualité d'une mesure. Un exemple frappant en est le fameux paradoxe de Hempel qui montre que l'usage de certaines implications (y compris leur contraposée) est parfaitement normal en déduction, par contre, en induction, c'est à dire lorsqu'on cherche à confirmer une hypothèse inductive à partir de données, cet usage est de nature différente et exige des précautions particulières.

MOTS-CLES : *Confirmation des hypothèses, Extraction de connaissances, Fouille de données, Fouille de textes, Induction.*

SUMMARY : *This paper, "KNOWLEDGE DISCOVERY IN DATA AND TEXTS," defines Knowledge Discovery in Texts (KDT, or Text Mining, TM) as a continuation of Knowledge Discovery in Databases (KDD, or Data Mining, DM). These two domains differ from classical Science, by their intensive use of inductive reasoning techniques, while they do use more classical methods as well, namely deductive and abductive reasoning. It is interesting to note that analogical reasoning is not yet of use in KDD, but it certainly will be in the near future.*

Induction means that measurements on data do not confirm an existing model of the data - as deductive reasoning would do - instead, they are used in order to build new, still unknown models. This in turn, profoundly changes the way implication has to be used. The most famous example of such a change is known as Hempel's paradox, showing that the normal deductive equivalence of a theorem and its contraposition can be harmful during an induction. We then stress that the inductive use of contraposition asks for special care.

KEY WORDS : *Data Mining, Knowledge Discovery in Data, Knowledge Discovery in Texts, Inductive reasoning, Theory of confirmation, Text Mining.*

INTRODUCTION

Le succès industriel et simultanément universitaire de l'ECD est une surprise. Il est amusant de constater que les universitaires américains distinguent l'ECD (qu'ils appellent Knowledge Discovery in Databases, KDD), de la fouille de données (FD, "Data Mining"), terme universellement adopté par les industriels. En effet, la majorité des universitaires pratiquent la

[1] CNRS, LRI Bât. 490 Univ. Paris-Sud, F - 91405 Orsay Cedex, yk@lri.fr

FD, qui consiste à développer des méthodes pour extraire la connaissance d'un ensemble fixé de données, souvent largement différentes des données réelles. Les industriels, au contraire, pratiquent l'ECD, c'est à dire un cycle qui comprend le processus entier de sélection des données, de nettoyage des données, leur transfert à une technique de FD, l'application de la technique de FD, la validation des résultats de la technique de FD, et finalement l'interprétation de ces résultats pour un utilisateur.
L'Apprentissage Symbolique Automatique (ASA) et l'ECD ont en commun un lien très fort: tous les deux reconnaissent l'importance de l'induction en tant que méthode normale de raisonnement, alors que les autres domaines scientifiques sont peu disposés à l'accepter, et c'est peu dire. Le lecteur peut trouver dans [Partridge 97] une discussion du pourquoi et dans [Kodratoff 92, Kodratoff 94] une discussion du comment de l'utilisation de l'induction.

Cet article analyse les problèmes posés par l'utilisation des systèmes existants : systèmes issus de la linguistique calculatoire, pour le pré-traitement des textes, et systèmes issus de la communauté de l'ECD, pour la détection inductive de relations au sein des textes.

Les problèmes de nature linguistique sont essentiellement

- les résultats quelques fois erratiques donnés par les analyseurs syntaxiques que nous avons pu tester,
- le fait que les relations qui font sens prennent place entre concepts (et non entre simples termes) et que la définition des concepts soit - en réalité - très variable en fonction du contexte,
- le fait que les méthodes d'apprentissage de concepts à partir d'exemples soient encore très rudimentaires (cet article présentera certains de nos résultats en ce domaine).

Les problèmes de nature statistique sont essentiellement liés aux fait que

- le statistiquement significatif diffère largement de l'utile ou du compréhensible,
- que la plupart des systèmes repèrent des règles à fort taux de couverture et ("donc") totalement triviales - et c'est pourquoi nous utilisons l'intensité de l'implication,
- il n'y a pas de lien clair entre les associations détectées par les systèmes existants et la notion de causalité, telle que les réseaux bayésiens peuvent le rechercher.

Nous décrivons ici une méthode inductive de découverte de modèles de relations entre concepts valides dans les textes, et nous l'illustrons par deux applications sur des données réelles. En Français, nous avons analysé les textes présents sur la toile de la revue québécoise "Bise" consacrée à l'étude de l'environnement et de ses pollutions. En Anglais, nous avons analysé les contes de Grimm. Nous montrons que les logiciels d'analyse de textes que nous utilisons sont suffisamment peu bruités pour être utilisables, mais qu'on peut les améliorer de nombreuses façons. En particulier, le contexte de l'ECD exige l'utilisation de taxonomies exprimant des relations de généralité entre concepts et leurs instances.
Pour engendrer les associations sous-tendant le modèle, nous avons utilisé l'algorithme classique "Apriori", et CHIC, un logiciel de détection de l'intensité de l'implication.

Une des nombreuses raisons de l’opposition des scientifiques à l'induction tient dans ce qu'on a appelé le paradoxe de Hempel qui "interdit" la confirmation expérimentale des inductions. Cet article va, au contraire, montrer comment le paradoxe de Hempel devrait être utilisé pour obtenir une mesure satisfaisante de la confirmation des inductions.

LE PARADOXE DE HEMPEL ET LA THEORIE DE CONFIRMATION

Ce fameux argument s'appuie sur l'existence d'une contraposition pour chaque théorème:

$$(A \Rightarrow B) \sim (\neg A \vee B) \sim (\neg A \vee \neg\neg B) \sim (\neg\neg B \vee \neg A) \sim (\neg B \Rightarrow \neg A).$$

Tout théorème est certes confirmé par l'observation simultanée de sa prémisse et de sa conclusion, mais aussi par l'observation simultanée de la négation de sa prémisse et de la négation de sa conclusion, à cause de l'existence d'une contraposition. Par exemple :

$\forall x$ (corbeau (x)$\Rightarrow$ noir (x)) [confirmé par l'observation de corbeau (A) $\wedge$ noir (A)]	~	$\forall x$ ($\neg$ noir (x)$\Rightarrow$ $\neg$ corbeau (x)) [confirmé par l'observation de $\neg$ corbeau (B) $\wedge$ $\neg$ noir (B). Exemple : blanc(A) $\wedge$ soulier (A)]

L'existence d'une contradiction nette est absolument flagrante dans un monde infini où le nombre d'objets qui confirment la contraposition peut devenir accablant. Si, pour confirmer que les corbeaux sont noirs, on est obligé de compter, entre autres, tous les grains de sable dans la mer, ceux qui sont noirs et ceux qui ne le sont pas, alors il vaut mieux abandonner de suite l'idée d'une confirmation empirique. De plus, l'absurdité d'un théorème fait que sa contradiction est confirmée universellement. Par exemple, aucun être humain n'a les cheveux verts et trois jambes, donc la totalité de l'humanité confirme la contraposition de $\forall x$ (cheveux_verts (x) $\Rightarrow$ trois_jambes (x)). Dans [Kodratoff 94; 00], se trouve un argument détaillé au sujet des limites de ce paradoxe, et nous le résumerons et simplifierons ici comme suit:

- Ou bien le théorème exprime une description (théorèmes descriptifs), comme dans $\forall$ x (corbeau (x) $\Rightarrow$ noir (x)). Dans ce cas, l'implication est bien une implication logique du point de vue déductif, puisque de $\neg$ noir (X) on déduit justement $\neg$ corbeau (B). Par contre, ce n'est pas une implication du point de vue inductif car la croyance que nous avons en $A \Rightarrow_{descriptif} B$ ne dépend pas d'éventuelles observations de (A = Faux, B = Faux). En d'autres termes, le mécanisme de la confirmation doit confirmer $A \Rightarrow_{descriptif} B$ chaque fois qu'il observe (A = Vrai, B = Vrai) et l'infirmer chaque fois qu'il observe (A = Vrai, B = Faux), sans prendre en compte le nombre de fois où (A = Faux, B = Faux).

- Ou bien le théorème exprime une causalité (théorèmes causaux), comme par exemple, $\forall$x (inhale_amiante (x) $\Rightarrow$ cancer_poumon (x)), ou encore $\forall$x ([pleut_sur (x) $\vee$ arrosé(x)] $\Rightarrow$ mouillé (x)). Dans ce cas, l'implication est une implication logique aussi bien du point de vue déductif que du point de vue inductif, et le mécanisme de la confirmation doit confirmer $A \Rightarrow_{causal} B$ chaque fois qu'il observe (A = Vrai, B = Vrai) et (A = Faux, B = Faux). Par exemple, l'absence simultanée d'humidité (= $\neg$ mouillé (x)) et de pluie confirme que la pluie et l'arrosage mouillent, c'est à dire que les instances de la contraposée $\forall$x ($\neg$ mouillé (x) $\Rightarrow$ [$\neg$ pleut_sur (x) $\wedge$ $\neg$ arrosé(x)]) confirment en effet le théorème direct. De même, l'absence simultanée de cancer et d'amiante confirme que l'amiante donne le cancer. Bien entendu, comme pour les descriptions, l'observation de (A = Vrai, B = Faux) infirme $A \Rightarrow_{causal} B$.

Notons en passant que notre intuition nous trompe un peu dans ce cas. En effet, chacun est tellement convaincu de la vérité du théorème $\forall$x ($\neg$ inhale_amiante (x) $\Rightarrow$ $\neg$ cancer_poumon (x)) qu'il fait tout pour ne pas respirer d'amiante afin d'éviter le cancer, alors que le théorème contraposé de la causalité de l'amiante sur le cancer est : $\forall$x ($\neg$ cancer_poumon (x) $\Rightarrow$ $\neg$ inhale_amiante (x)). En pratique, cette intuition n'est pas si fausse puisque l'observation de

[¬ cancer_poumon (Toto) ∧ ¬ inhale_amiante (Toto)] et de
[cancer_poumon (Zozo) ∧ inhale_amiante (Zozo)] confirment ces trois théorèmes à la fois.

C'est pourquoi, quand on calcule une confirmation, il faut faire la différence entre les théorèmes descriptifs, et ce que nous appelons des théorèmes causaux[2] où l'implication comporte une signification causale. Du fait que la Science s'est intéressée seulement à la déduction jusqu'à présent, la différence entre théorèmes descriptifs et causaux n'est pas reconnue. Nous espérons que cet article montre clairement en quoi cette différence est utile - et finalement assez simple.

LES MESURES DE SIGNIFICATION STATISTIQUE

En général, on considère que sens statistique et intérêt sont synonymes. Nous verrons plus loin des mesures d'intérêt dont certaines sont en effet une forme particulière de mesure du sens statistique, mais ce n'est pas toujours le cas, c'est pourquoi je prends soin de distinguer les deux.
Ce paragraphe va illustrer avec précision l'importance du lien entre la nature des mesures numériques et la sémantique de la relation cherchée dans les données. En fait, selon que la relation est de nature descriptive ou de nature causale, je prétends que les mesures de signification statistique devraient être différentes.

Soient A et B deux assertions, et soient X et Y leurs supports, c.-à-d., {X} = {x / A(x) = Vrai} et {Y} = {x / B(x) = Vrai}. Supposons aussi que nous traitons avec un ensemble fini d'exemples, de sorte que {X} ∪ { $\overline{X}$}= {Y} ∪ { $\overline{Y}$}= {Tot}. Le cardinal de {Tot} est le nombre total d'exemples, c'est à dire le nombre total d'enregistrements de la BD. Quand nous voulons inférer A ⇒ B à partir des données, alors les mesures suivantes fourniront une estimation de la validité de cette implication.

Ces mesures calculent plusieurs relations entre les cardinaux de {X} et {Y}. Deux mesures très classiques sont celles de support et confiance qui sont utilisées par tous les systèmes d'ECD qui découvrent des associations.
Pour illustrer notre raisonnement, considérons l'ensemble suivant de données.

[2] Deux objections sont classiquement faites à mon affirmation de l'aspect non paradoxal de la contraposition pour les théorèmes causaux.
L'une, parfaitement valide, est que les causalités fortuites introduisent des paradoxes, comme par exemple dans ∀x (fume(x) ∧ Français(x) ⇒ cancer(x)) absurdement confirmé par tous les non Français non cancéreux. Dans le cas où la vraie cause est présente dans les données, on sait facilement la distinguer des causes fortuites par un calcul numérique. Par contre, il est exact que l'ignorance d'une cause vraie, et son remplacement par une cause fortuite va mener à des paradoxes. Ceci a été illustré par l'exemple du "paradoxe français": bien entendu personne ne croyait que la nationalité française suffisait à protéger de l'infarctus! Ces paradoxes indiquent la nécessité de rechercher une cause non fortuite. En d'autres termes, le paradoxe de Hempel existe bel et bien dans le cas où on fait l'hypothèse d'une causalité fortuite.
L'autre objection, à mon sens non valide, est liée à une erreur de représentation des connaissances: l'introduction de conjonctions ou de disjonctions dans la prémisse du théorème causal serait à l'origine de contradictions en confirmation. Par exemple, ∀x (fume(x) ∧ boit_alcool(x) ⇒ cancer(x)) est absurdement confirmé par les fumeurs non-buveurs et non cancéreux. C'est tout simplement dû au fait que le "∧ médical" n'est pas un ∧ logique, et que le théorème proposé est en effet absurde avec un ∧ logique (chacun isolément peut causer le cancer, et chacun renforce l'action de l'autre, mais ceci n'est pas un ∧ logique).

Table 1. Un exemple de données confirmant le théorème seulement par sa contraposition.

	A	**¬A**	P(A,B)=1/10,P(¬A,¬B)=7/10,P(¬B,A)=1/10,P(A)=2/10,P(B)=2/10, P(¬A)=
B	2	2	8/10,P(¬B)=8/10,P(B|A)=1/2,P(¬A|¬B)=7/8,P(¬B|A)=1/2, P(A|¬B)=1/8
¬B	2	14	Confirmation (directe) = 0, Confirmation (par contraposition) = 6/10

Ces données ne confirment absolument pas $A \Rightarrow B$ puisque le nombre d'exemples tels que $A \wedge B$ est égal au nombre d'exemples tels que $A \wedge \neg B$. De plus, chacun d'eux couvre seulement 1/10 des exemples qui peuvent être dus au bruit, surtout si nous supposons un niveau de bruit de 10%. Ainsi, $A \Rightarrow_{descriptive} B$ est fausse puisqu'elle est autant confirmée qu'infirmée.
Inversement, la contraposition de $A \Rightarrow B$ est bien confirmée puisque $P(\neg A, \neg B) = 7/10$ ce qui est plus élevé que le niveau de bruit que nous avons supposé. Le fait que l'implication causale soit assez fortement confirmée, et ce, donc, seulement par la contraposition, nous pousse à penser que $A \Rightarrow_{causale} B$ a de bonnes chances d'être absurde.

Dans la suite, "l'implication" est toujours $A \Rightarrow B$, et "la contraposition" est toujours $\neg B \Rightarrow \neg A$.

MESURES DE SUPPORT ET DE CONFIANCE

Les mesures absolument classiques de support et de confiance sont définies respectivement par **P(A, B)** et: **P(B | A)** = P(A, B) / P(A).si P(A) ≠ 0.

Malgré leur application universelle, il semble évident que la mesure de support définisse parfois un concept sans intérêt puisque le support effectif d'une implication dépend aussi de son degré d'infirmation. L'intuition relative à l'intérêt d'un support est que P(A, B) doit être " largement supérieure " à P(A, ¬B) pour que l'implication puisse avoir un sens. Si P(A, B) est, disons K fois supérieure à P(A, ¬B), alors, en effet, la définition classique redevient intéressante. Autrement dit, on peut définir

Supporteffectif-de (A⇒B) =
- 0 si P(A, B) < K * P(A, ¬B),
- P(A, B) sinon.

où K est un coefficient dépendant du domaine.

De façon semblable, nous proposons une définition de la confiance effective par

Confianceffective-en (A⇒B) =
- 0 si P(A, B) < K * P(A, ¬B),
- P(B | A) sinon.
= Supporteffectif-de (A⇒B) / P(A)

Ces nouvelles mesures sont identiques aux anciennes dans le cas où l'implication est largement confirmée, mais elles sont égales à 0 si l'implication n'est pas K fois plus confirmée qu'infirmée.
Cette nouvelle définition peut paraître très proche de l'ancienne mais, en pratique, elle apporte une variation importante. En effet, il est bien évident que les implications à très forte couverture, disons supérieure à 0,5, ne peuvent qu'être que peu infirmées. Mais comme ces implications à forte couverture sont bien entendu valides pour la majorité de la population étudiée, elle décrivent des propriétés qui sont soit bien connues, soit triviales. Pour reprendre l'analogie minière de base de l'ECD, une mine d'or dans laquelle l'or est majoritaire par rapport aux déchets est certes intéressante, mais, sauf chance exceptionnelle, elle a été déjà exploitée depuis longtemps. Ces sont les mines dont l'or est caché par une grande quantité de

scories qui présentent la plus grande plus-value. Les implications à très forte couverture sont donc, sauf chance exceptionnelle, des sortes de scories de la connaissance.
Dans la mesure où P(A, B) doit être assez faible pour que l'implication soit intéressante, il est clair que rien n'interdit à P(A, ¬B) de prendre des valeurs du même ordre, surtout en présence de données bruitées.

Lorsque les implications détectées expriment des dépendances causales entre prémisse et conclusion, nous suggérons que les définitions du support et de la confiance prennent explicitement en compte la contraposition. En d'autres termes, nous suggérons d'utiliser

Supporteffectif$_{causal}$-de (A⇒B) =
- 0 si P(A, B) + P(¬A, ¬B) < 2K * P(A, ¬B)
- P(A, B) + P(¬A, ¬B) sinon.

La confiance sera alors calculée par

Confianceffective$_{causale}$-en (A⇒B) =
- 0 si P(A, B) + P(¬A, ¬B) < 2K * P(A, ¬B),
- 1/2 [P(B | A) + P(¬A | ¬B)] sinon

Dans le tableau 1, en posant K = 1, le Supporteffectif-de (A⇒B) égale 0, alors que le Supporteffectif$_{causal}$-de (A⇒B) égale 8/10.

Le vrai problème, dans ce cas, n'est pas d'analyser une table de données unique: on ne pourra jamais savoir si une implication confirmée par sa contraposée seulement est absurde ou non à partir d'une seule observation. Dans ce cas, la façon dont se manifeste dans le temps le fait que A cause B, d'un côté, et que ¬B "cause" ¬A, de l'autre, n'ont strictement aucun rapport. Par exemple, fumer cause le cancer selon une certaine fonction liée au temps et à la quantité fumée, alors que c'est de façon indépendante du temps que ne pas avoir le cancer "cause" (que la causalité semble aller maintenant en sens inverse de l'implication est un autre problème, non abordé ici) ne pas fumer.
Ainsi, en analysant des tranches d'âge de population vieillissante, on va découvrir que P(B | A) augmente avec le temps, alors que P(¬A | ¬B) ne varie pas, ou, en général, varie de façon extrêmement différente. Ainsi, une valeur élevée de P(¬A | ¬B), possiblement insignifiante sur des populations jeunes (non encore affectées par l'effet B lorsque A est vrai, et donc que P(B | A) est encore faible), est reconnue comme certainement significative en observant les populations âgées où P(B | A) devient important.

MESURES DE CORRELATION OU DE DEPENDANCE

Coefficient de corrélation linéaire

Considérons les variables x et y à valeurs réelles x_i et y_i pour chacun des N enregistrements de la base de données. Soit M_x la valeur moyenne de x, et E_x sa variance. Alors, la corrélation linéaire de x et y est définie par:

$$1 / (N-1) \Sigma_{i = 1, \ldots, N} ((x_i - M_x) / E_x) * ((y_i - M_y) / E_y)$$

Cette mesure se justifie intuitivement par le fait que lorsque x et y sont corrélés, alors x_i et y_i tendent à être ensemble "du même côté de la moyenne".

La corrélation est appelée plutôt dépendance quand on traite des valeurs discrètes. La définition classique de cette mesure est:

Dépendance(A, B) = Abs(P(B | A) - P(B))

Où Abs est la fonction valeur absolue.
Afin de conserver une définition assez proche, nous définirons la dépendance effective par

Dépendanceffective(A, B) =

- 0 si Confianceffective-en (A⇒B) = 0

- Abs(P(B | A) - P(B)) sinon.

Cette définition nous permet d'éliminer les dépendances insuffisamment confirmées, et elle conserve les propriétés de la définition classique dans les autres cas.

La dépendance causale est plus complexe à définir en ce sens que les dépendances fortuites et, éventuellement, les dépendances indirectes[3] doivent être éliminées ([Pavillon 96] n'élimine que les dépendances fortuites, mais nous ne voulons pas entrer dans ce débat ici). Par contre, cette mesure est directement inspirée de la **Confianceffective$_{causale}$**

SI A⇒B n'est pas fortuite (ni indirecte**)**

ALORS

Dépendanceffective$_{causale}$(A, B) =

- 0 si P(A, B) + P(¬A, ¬B) < 2K * P(A, ¬B),

- 1/2 [Abs(P(B | A) - P(B)) + Abs(P(¬A | ¬B) - P(¬A))] sinon

Une autre application de nos mesures se trouve dans la construction automatique de réseaux bayésiens à partir de données. Les méthodes modernes de construction utilisent le principe de description de longueur minimale [Munteanu00], [Jouffe 00], et n'évaluent jamais explicitement les valeurs des probabilités conditionnelles. Dans la mesure où la nature causale des réseaux bayésiens est quelques fois discutée, il paraît intéressant de comparer, dans le réseau obtenu par apprentissage, les valeurs de la dépendance, causale ou non causale.

LES MESURES D'INTERET

L'induction est toujours un processus intellectuellement dangereux s'il n'est pas étroitement contrôlé. Il est évident que les buts de l'utilisateur constituent les meilleurs a priori à utiliser pour éviter d'inventer des banalités. Jusqu'à maintenant, les recherches ont négligé ce point de vue du problème car la recherche essaie de développer des techniques qui sont, autant que possible, universelles au lieu d'être spécifiques: quel que soit le but de l'utilisateur, le résultat se doit d'être optimal, ce qui est la quadrature du cercle. Quelques recherches ont néanmoins été faites dans la communauté ECD sous le nom de validation des règles ou "mesures d'intérêt". Ceci a produit trois mesures d'intérêt différentes.

L'INTENSITE DE L'IMPLICATION

Cette mesure, nommée intensité de l'implication, [Gras 93] mesure la distance à l'aléatoire d'implications de faible support. Soient A' et B' deux sous-ensembles aléatoirement choisis au sein des données et tels que le cardinal de A' égale celui de A, et le cardinal de B' égale celui de B. On dira que A ⇒ B est intense lorsque P(A, ¬B) est beaucoup plus faible que P(A', ¬B'). Elle est mesurée par I = P(**A, ¬B**) - P(A', ¬B').

La mesure de I demande des calculs complexes qui sont simplifiés en choisissant une approximation à la distribution aléatoire attendue. Les mesures actuelles utilisent une approximation de Poisson. On a :

$I = 1/(\sqrt{2}\sqrt{p}) \int_{ii}^{\infty} e^{-t^2/2} dt$

où ii est donnée par les valeurs observées, comme suit :

[3] L'observation d'une dépendance entre le fait "C" et le fait "B" est appelée fortuite lorsqu'il existe un troisième fait, "A" (la causalité "réelle") tel que P(B | A, C) = P(B | A): A est causal pour B, mais C ne l'est pas en réalité. L'observation d'une dépendance entre le fait "A" et le fait "B" est appelée indirecte lorsqu'il existe un troisième fait, "C" (le "nécessaire intermédiaire" à la causalité) tel que P(B | A,C) = P(B | C): A est bien causal pour B, mais cette causalité ne se manifeste que par l'intermédiaire de C.

$n_{ab'} = |\{X\} \cap \{\bar{Y}\}|$, $n_a = |\{X\}|$, $n_{b'} = |\{\bar{Y}\}|$, $n = \{Tot\}$, alors

$ii = (n_{ab'} - n_a\, n_{b'}/n) / (\sqrt{n_a}\,\sqrt{n_{b'}})/n$.

Cela montre que, bien qu'un peu plus compliquée à calculer que les autres mesures, l'intensité de l'implication n'est pas exagérément coûteuse.

REGLES CONTRADICTOIRES

La deuxième mesure cherche les règles contradictoires issues des données. Elles sont définies comme étant intéressantes. Cette méthode cherche des couples d'assertions de la forme:

$A \Rightarrow B$

$A \wedge A' \Rightarrow \neg B$.

Un exemple d'une telle contradiction dans la vie réelle peut être:

Airbag $\Rightarrow$ augmente la sécurité

Airbag $\wedge$ (âge = bébé) $\Rightarrow$ diminue la sécurité.

Quelques combinaisons de signification statistique des deux assertions sont plus intéressantes, par exemple quand $A \Rightarrow B$ a un grand support et que $A \wedge A' \Rightarrow \neg B$ a un petit support mais une grande confirmation.

[Suzuki 97] a débuté ce travail dont la difficulté principale vient de ce qu'un trop grand nombre de contradictions sont trouvées dans les données, si bien qu'il faut rechercher (récursivement) les plus intéressantes. Dans [Suzuki 98] nous explorons la possibilité d'utiliser l'intensité d'implication pour caractériser des contradictions intéressantes.

INTERET D'UN ENSEMBLE DE REGLES

La troisième mesure ne s'attache pas à l'intérêt d'une règle individuelle, mais elle essaie d'évaluer l'intérêt d'un ensemble entier de règles: l'hétérogénéité globale de l'ensemble est alors supposée exprimer l'intérêt [Gago 98].

On définit une mesure de distance entre règles, et à chaque pas de l'algorithme, on rajoute à l'ensemble de règles existant la règle la plus distante, et ce jusqu'à ce qu'à un nombre fixé à l'avance.

Une extension évidente à leur méthode serait de chercher l'ensemble le plus hétérogène de règles qui maintient un certain % d'une mesure sur l'ensemble complet de règles.

LES NUANCES CRITIQUES (" NEAR MISSES ") ET LES REGLES TYPIQUES

Je voudrais suggérer ici une nouvelle mesure d'intérêt qui n'a jamais encore été implémentée. Elle utilise aussi une notion de distance, mais ne considère que la distance entre les prémisses de deux règles. Elle ne peut être utile que pour diminuer le nombre de règles présentées à l'utilisateur.

Soit un ensemble de règles du type Prémisse $\Rightarrow$ Conclusion obtenues par une mesure quelconque, et supposons que cet ensemble soit trop copieux pour pouvoir être présenté à l'utilisateur. On peut alors classifier (c'est à dire réunir dans une même classe) les règles ayant la même Conclusion. Les règles ayant une couverture et une confiance élevée vont exprimer les connaissances " bien connues " relatives à Conclusion.

L'idée de règle typique consiste à rechercher la prémisse la plus centrale (ou la plus typique) des règles d'une même classe. Elle résume en une seule formule ce qui est bien connu sur la façon d'obtenir une certaine conclusion : c'est une banalité typique.

L'idée de nuance critique consiste à rechercher toutes les règles ayant des prémisses très semblables, ou très proches, de la règle typique de la ième classe, mais concluant sur une classe différente. Ces règles, inversement, peuvent avoir une très faible couverture et même une assez petite confiance. Elles tracent les limites de la banalité et vont justement exprimer la façon dont on peut se distinguer de la banalité.

On ne présenterait à l'utilisateur que les règles typiques et les nuances critiques qui sont associées à chaque règle typique.

UN EXEMPLE D'ECT

Nous avons appliqué le logiciel d'analyse des langues française et anglaise, TROPES (un produit industriel dû à la compagnie Acetic) aux contes de Grimm (en Anglais - 200 contes) et à la revue environnementale québécoise Bise (10 ans de revue, produisant quelques 170 textes). Le hasard a fait que les deux expériences utilisent à peu près la même quantité de textes, chacun de l'ordre de 5 megabytes. Le travail d'extraction de connaissance prend place en 3 étapes principales:

1. Définition des concepts intéressants
2. Discrétisation des données numériques
3. Extraction de règles

DEFINITION DE CONCEPTS INTERESSANTS

Une composante importante du logiciel Tropes est la définition de classes en trois niveaux de généralité. De plus, il autorise la définition de classes propres à l'utilisateur. Il est évident qu'une bonne définition des concepts est capitale pour obtenir des résultats intelligibles. Par exemple, le concept de "mythologie" est défini dans Tropes par l'occurrence dans les textes des contes de Grimm des mots: "heaven, hell, elf et gnome". Le moins que l'on puisse dire, c'est que ces mots recouvrent mal notre intuition quant au sens de ce concept. Le logiciel détecte les occurrences de ces concepts. La table 2 ci-dessous explicite les définitions des classes que nous avons utilisées dans l'expérience décrite dans cet article.

Table 2: Liste des concepts définis pour Tropes.

Définition des concepts utilisés pour les contes de Grimm			Définition des concepts utilisés pour la revue Bise		
Niveau	Terme	Concept	Niveau	Terme	Concept
2	anger	anger	3	air	air
3	animal	animal	3	alimentation	alimentation
2	aristocracy	aristocracy	3	boisson	boisson
2	amphibian	batrachian	0	cancer	cancer
2	bird	bird	0	leucémie	cancer
1	boy	boy	0	sarcome	cancer
0	hans	boy	3	chimie	chimie
0	hansel	boy	2	combustible	combustibles
0	poor_boy	boy	0	gaz_naturel	combustibles
1	fortification	castle	0	oxygène	combustibles
1	palace	castle	0	propane	combustibles
3	child	child	0	comp_nitré	comp_nitré
3	death	death	0	nitrate	comp_nitré
0	dwarf	dwarf	0	nitrite	comp_nitré
0	ah	exclamation	0	contaminant	contamination
1	rhetorical_device	exclamation	1	contamination	contamination
3	feeling	feeling	2	contrôle	contrôle
3	food	food	2	cours_d_eau	cours_d_eau
2	folly	fool	1	déchet	déchet
3	friend	friend	2	boisson_alcoolisée	drogue
0	Cinderella	girl	2	drogue	drogue

1	female_child	girl	2	tabac	drogue
1	girl	girl	3	eau	eau
0	gretel	girl	0	puits	eau
0	grethel	girl	3	environnement	environnement
0	maid	girl	1	gaz	gaz
0	One-Eye	girl	0	ammoniac	gaz_polluants
0	Three-Eyes	girl	0	chlore	gaz_polluants
0	Two-Eyes	girl	0	co2	gaz_polluants
2	happiness	happiness	0	dioxine	gaz_polluants
3	hunting	hunt	0	fluor	gaz_polluants
0	blood	inside_body	0	furane	gaz_polluants
0	bone	inside_body	0	furanne	gaz_polluants
0	brain	inside_body	0	monoxyde	gaz_polluants
0	entrail	inside_body	0	oxyde_de_soufre	gaz_polluants
0	heart	inside_body	0	radon	gaz_polluants
0	lung	inside_body	0	Dupont_de_nemours	industrie_chimique
0	rib	inside_body	1	industrie_chimique	industrie_chimique
0	stomach	inside_body	3	informatique	informatique
0	tongue	inside_body	2	intérieur_du_corps	intérieur_du_corps
0	tooth	inside_body	1	intoxication	intoxication
0	vein	inside_body	0	acétone	liqde_polluant
3	man	man	0	benzaldéhyde	liqde_polluant
3	money	money	0	benzène	liqde_polluant
3	mountain	mountain	0	bromure	liqde_polluant
0	dragon	myth_animal	0	chloroéthylène	liqde_polluant
0	frog	myth_animal	0	chloroforme	liqde_polluant
1	mythical_monster	myth_animal	0	hydrocarbure	liqde_polluant
3	mythology	mythology	0	hydrocarbures	liqde_polluant
3	occultism	occultism	0	hypochlorite	liqde_polluant
3	piece	part_of	0	phénol	liqde_polluant
3	plant	plant	0	styrène	liqde_polluant
0	ferdinand	servant	0	tétrachlorure	liqde_polluant
2	servant	servant	0	toluène	liqde_polluant
0	waiting-maid	servant	3	mer	mer
1	tree	tree	2	métal	métal
3	woman	woman	3	nocivité	nocivité
				etc.	

Lorsque le niveau est compris entre 1 et 3, les concepts sont définis en fonction de la taxonomie à trois niveaux fournie par Tropes. Par exemple, le concept de "servant" est défini comme la classe de niveau 2 définie dans Tropes. Nous y ajoutons le mot "Ferdinand" pour prendre en compte qu'un personnage nommé Ferdinand joue le rôle de serviteur dans un conte. Lorsque le niveau indiqué est 0, alors nous n'utilisons pas la taxonomie de Tropes et nous définissons nos propres concepts. Par exemple, les concepts de "inside_body" et de "liqde_polluant" ne sont pas définis dans Tropes, ils le sont par les mots dans la colonne "termes". On constate que nous avons très mal défini les concepts de "boy" et "girl" puisque la liste de prénoms fournis est loin d'être complète. Ceci nous fournit un exemple de concept mal défini par l'utilisateur, et dont les relations avec d'autres concepts ne sont pas significatives.

Pour chaque texte, le rapport nombre d'occurrences d'un concept, divisé par le nombre total de mots du texte, définit la fréquence d'occurrence du concept dans ce texte. On obtient ainsi une table d'occurrence des concepts dans chaque texte.

DISCRETISATION DES DONNEES NUMERIQUES

Les méthodes statistiques classiques, et en particulier le logiciel CHIC que nous avons utilisé, sont destinées à détecter des relations entre variables, à partir d'un tableau des corrélations, définies plus haut. Dans l'application qui nous occupe ici, le fait que deux concepts soient corrélés ou non peut être intéressant, mais, de fait, nous avons besoin d'une analyse plus fine des interactions. En particulier, il faut savoir si les concepts sont corrélés via leur absence ou via leur présence. Par exemple, dire que le concept de "death" est fortement corrélé à celui de "aristocracy" peut conduire à une interprétation absurde quand on s'aperçoit (comme c'est le cas dans les contes de Grimm) que c'est l'absence du concept de mort dans un conte qui implique l'absence du concept d'aristocratie.
Une étude détaillée comporte donc une phase de discrétisation des données numériques qui dépend du niveau de finesse recherché. En pratique, la taille des donnés est aussi importante. Par exemple, le nombre de contes de Grimm étant limité à 200, il est absurde de créer un nombre excessif de variables qui auront presque toutes une couverture quasi nulle et dont les interactions les unes avec les autres n'auront plus aucun sens.
C'est pourquoi, pour cette première phase d'expériences, nous avons adopté une discrétisation uniforme, qui mériterait d'être raffinée dans une application réelle. Nous conservons les valeurs 0 sans modification (elles représentent un taux d'apparition nul du concept dans le texte considéré), et nous divisons les valeurs supérieures à 0 selon une équipartition en deux. Les valeurs de la moitié inférieure sont dites représenter un taux d'apparition "faible" du concept dans le texte considéré, et celles de la moitié supérieure, un taux d'apparition "fort" du concept.
C'est ainsi que nous avons obtenu, par exemple, que 112 des 200 contes de Grimm ne parlent pas de la mort, que 44 en parlent "un peu" et que 44 en parlent "beaucoup". Du fait de notre choix relatif à l'équipartition des valeurs non nulles, il se trouve que, pour ce concept, "un peu" correspond à une fréquence comprise entre 0,00028 et 0,00091, et que "beaucoup" correspond à une fréquence comprise entre 0,00103 et 0,02326. Ces fréquences sont donc, en principe, différentes pour chaque concept.
Le tableau 3 ci-dessous montre les 19 premières valeurs discrétisées en face des valeurs continues dont elles sont issues.

Table 3.

death_0	death_1	death_2	death
1	0	0	0
1	0	0	0
0	0	1	0,00165
0	1	0	0,00028
1	0	0	0
0	0	1	0,00273
1	0	0	0
1	0	0	0
0	0	1	0,00323
1	0	0	0
1	0	0	0
1	0	0	0
0	0	1	0,00103
1	0	0	0
0	1	0	0,00036
0	0	1	0,0074
0	1	0	0,00067
0	0	1	0,00413
1	0	0	0

EXTRACTION DE REGLES

L'extraction de règles à partir de ces valeurs peut se faire de nombreuses manières différentes. Nous avons testé deux types de logiciels différents.
Un logiciel classique dit de détection des associations utilisant les mesures de support et de confiance. Le nombre d'associations ainsi détectées est tellement important que nous avons abandonné immédiatement cette approche.
Le logiciel CHIC mesurant l'intensité de l'implication. CHIC peut aussi saturer le champ visuel de l'utilisateur, mais présente l'avantage que l'on peut éliminer à la volée des concepts et ainsi se concentrer sur des relations intéressantes, comme nous allons l'illustrer sur nos exemples.
La figure 1 ci-dessous présente les implications d'intensité supérieure ou égale à 99% pour les concepts de la revue Bise. Bien entendu, si on demandait le graphique correspondant à une intensité de, disons, 85%, on obtiendrait un graphique totalement illisible, tout comme les logiciels de détection d'associations.

Afin d'éviter ce phénomène, il est utile de diminuer progressivement le niveau requis pour accepter une implication, tout en se concentrant sur des concepts jugés intéressants. On ne sait pas d'avance quelles relations seront possibles, si bien qu'aucun concept ne peut être éliminé a priori

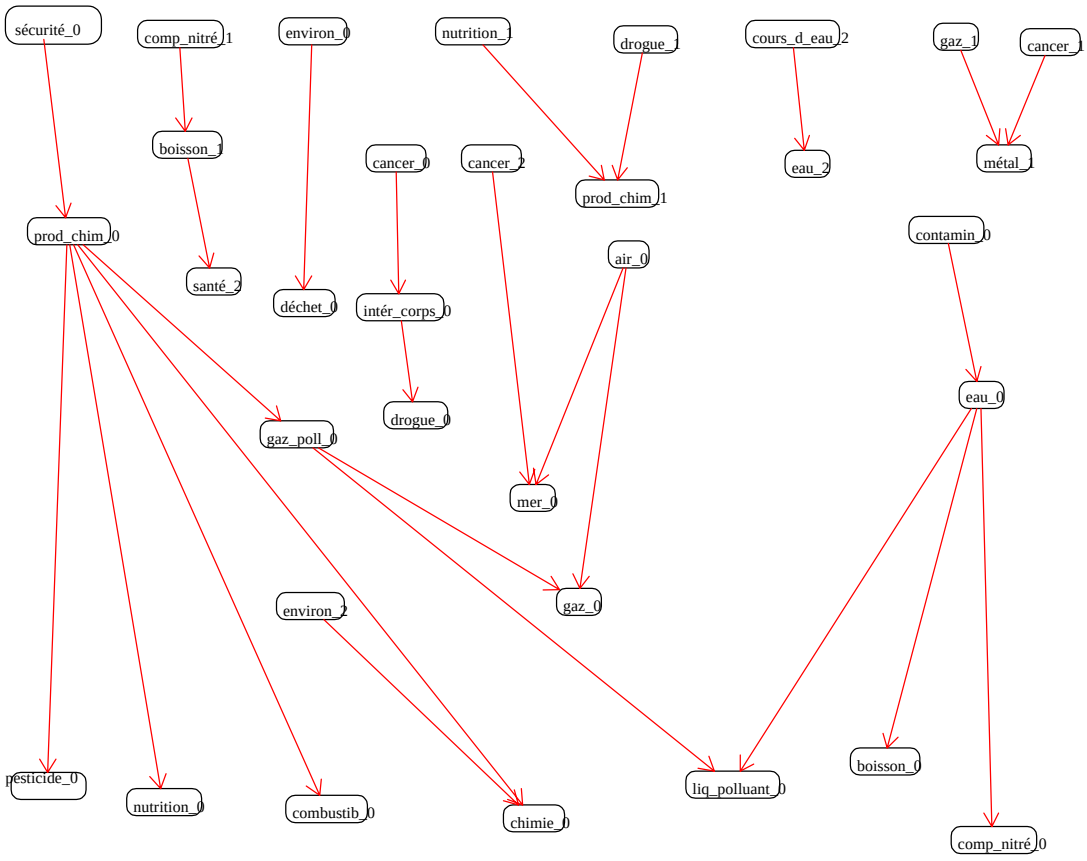

Figure 1. Relations très intenses entre concepts de la revue Bise.

Ces relations peuvent être triviales, comme le fait que l'on parle beaucoup d'eau quand on parle de cours d'eau.
Par contre, le fait que de parler un peu de composés nitrés provoque le fait que l'on parle un peu de boissons est un trait caractéristique de cette revue, et qui peut permettre de la distinguer d'autres revues semblables. De même, le fait que de parler beaucoup de cancer implique que l'on ne parle pas de la mer n'est certainement pas une trivialité de langage, mais une caractéristique de cette revue.

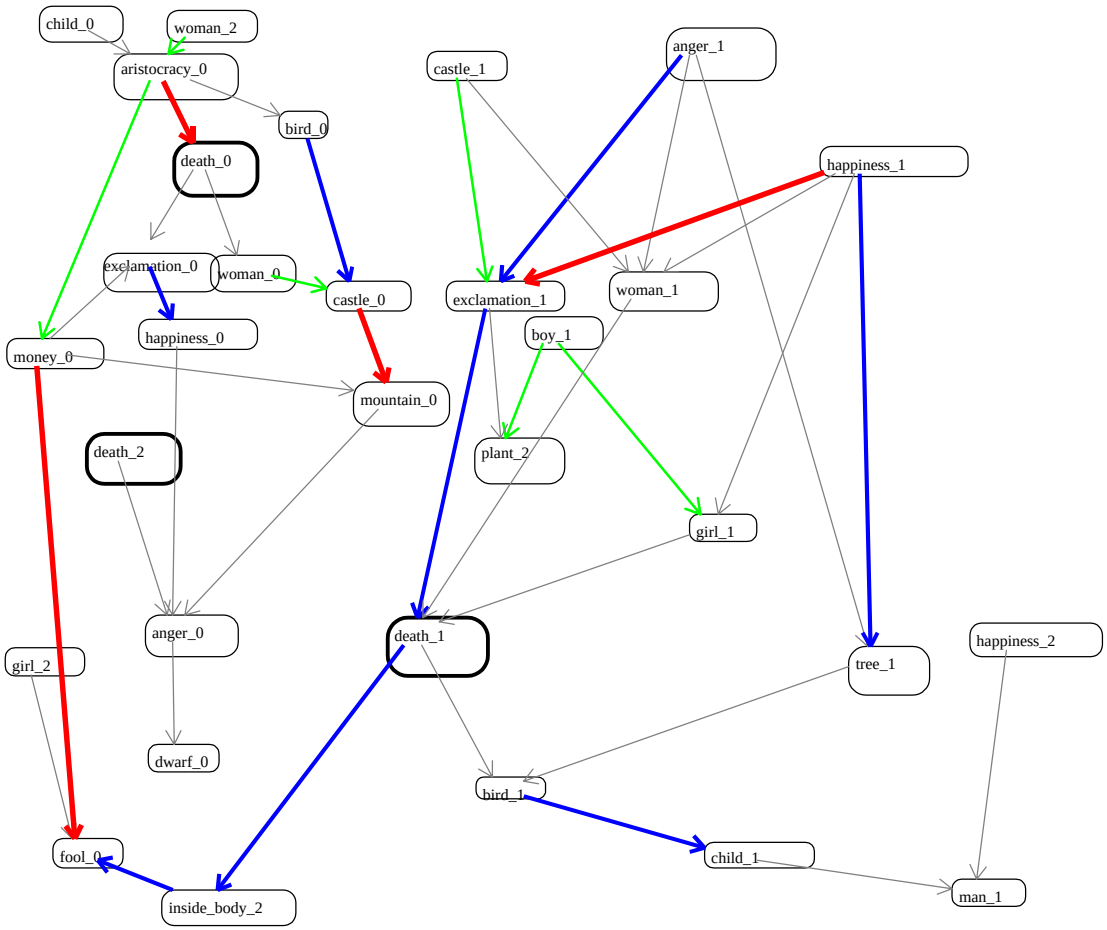

Figure 2. Relations d'intensité variant entre 99% (traits les plus épais) et 85% (traits les plus fins) entre les concepts liés à "death" dans les contes de Grimm.

Pour obtenir ce graphe, il faut isoler les trois concepts death_0, death_1, death_2 puis éliminer les concepts distants de plus de trois liens de ces trois concepts.
Notre choix d'équipartition entre death_1 et death_2 permet d'être certain que, dans les contes de Grimm, le thème de la mort, s'il devient très important, élimine de la même façon tout autre thème, excepté celui de l'absence de colère qui est complètement éliminé. Par contre, le fait qu'on parle un peu de la mort est directement "causé" par le fait qu'on parle un peu de

femmes, de filles, et que le texte présente quelques exclamations. Le fait que l'on parle un peu de la mort "cause" aussi que l'on parlera beaucoup de l'intérieur du corps (cette relation est facile à comprendre, elle montre que les contes de Grimm décrivent les circonstances de la mort avec détail) et un peu d'oiseaux. Bien entendu, n'étant pas folkloriste, je ne peux interpréter toutes ces constatations. De nombreuses expériences intéressantes pour le folkloriste sont possibles en utilisant cet outil.

Ce type de recherche s'applique directement à des textes d'un intérêt économique plus immédiat, comme les courriels, le e-commerce, les articles sur un thème etc.

CONCLUSION: UNE DEFINITION DE L'ECT

La plupart des auteurs définissent l'ECT ("Text Mining") comme étant soit de la recherche documentaire, soit de l'extraction d'information (c'est à dire l'instanciation de patrons prédéfinis par des textes). Je la définirais plutôt [Kodratoff 99] comme étant **la recherche inductive de connaissances nouvelles, intelligibles et utiles (pour un certain but) dans un corpus de nombreux textes**. L'utilité se manifeste par le fait que cette connaissance doit modifier le comportement d'un agent humain ou robotique. Cette connaissance peut aussi être, éventuellement, contradictoire, c'est à dire qu'elle contient des contradictions comme expliqué plus haut. Elle peut aussi contredire des connaissances existant sur le domaine. Par exemple, un texte s'opposant aux connaissances existantes doit évidemment contenir des connaissances contredisant la connaissance du domaine.
On notera aussi que je n'ai pas introduit de condition de précision, au contraire de toutes les méthodes expérimentales qui justifient leurs résultats sur un principe de précision empirique. La raison en est qu'une connaissance trop imprécise est inutile, et que c'est le critère d'utilité qui domine.
On notera encore que je n'introduis pas le critère de compréhension des textes eux-mêmes car il est possible d'extraire des connaissances utiles et compréhensibles de textes qui ne sont pas compris. Par exemple, on peut connaître le langage dans lequel un texte est rédigé par une analyse de n-grammes, sans compréhension de chaque texte.
Dire que cette recherche est inductive signifie que la connaissance recherchée est exprimée ni explicitement ni implicitement dans aucun des textes. Un exemple typique montrant bien la difficulté de cette démarche est celui des relations entre concepts ***n'existant pas*** dans les textes: leur nombre est toujours infini, et le but est de trouver les relations ***intéressantes*** qui sont absentes des textes. Par exemple, l'absence du concept de "mort" dans un conte de Grimm a des conséquences sur l'absence ou la présence d'autres concepts dans ces contes.
C'est donc un domaine entièrement nouveau qui n'a pas été encore souvent abordé, excepté par ceux qui ont déjà appliqué des techniques d'analyse de données à leurs textes, voir par exemple [Hogenraad 95].

BIBLIOGRAPHIE

[1] Gago P., Bento C., “A Metric for Selection of the Most Promising Rules,” in *Principles of Data Mining and Knowledge Discovery*, Zytkow J. & Quafafou M. (Eds.), pp. 19-27, LNAI 1510, Springer, Berlin 1998.

[2] Gras R., Lahrer A., “L’implication statistique: une nouvelle méthode d’analyse des données,” Mathématiques Informatique et Sciences Humaines 120,:5-31, 1993.

[3] Hogenraad, R., Bestgen, Y., Nysten, J. L."Terrorist Rhetoric: Texture and Architecture, in Ephraim Nissan, Klaus Schmidt (Eds.)"From Information to Knowledge Intellect, Oxford, GB, 1995, pp. 48-59.

[4] L. Jouffe, P. Munteanu, Smart-Greedy+ : Apprentissage hybride de réseaux bayésiens, Colloque francophone sur l'apprentissage (CAP), St. Etienne, juin 2000.

[5] Kodratoff Y, Bisson G. “The epistemology of conceptual clustering: KBG, an implementation”, *Journal of Intelligent Information System*, 1:57-84, 1992.

[6] Kodratoff Y., “Induction and the Organization of Knowledge”, *Machine Learning: A Multistrategy Approach*, volume 4, Tecuci G. et Michalski R. S. (Eds.), pages 85-106. Morgan-Kaufmann, San Francisco CA, 1994.

[7] Kodratoff Y., “Knowledge Discovery in Texts: A Definition, and Applications,” Proc. ISMIS’99, Warsaw, June 1999.

[8] Kodratoff Y. "L'induction symbolique et numérique en ECD", dans *Induction numérique-symbolique à partir de données complexes:des outils pour le "data mining*" Diday, Kodratoff, Brito, Moulet (Eds.) Cepadues, 2000.

[9] P. Munteanu, D. Cau, "Efficient Learning of Equivalence Classes of Bayesian Networks", 4th European Conference on Principles and Practice of Knowledge Discovery in Databases (PKDD), p. 96-105, Lyon, septembre 2000.

[10] Partridge D. “The Case for Inductive Programming,” *IEEE Computer 30*, 1, 36-41, 1997. Une version plus complète se trouve dans: “The Case for Inductive Computing Science,” in Computational Intelligence and Software Engineering, Pedrycz & Peters (Eds.) World Scientific, in press.

[11] Pavillon G. "ARC II: a System for Inducing and Simplifying Dependence and Causal Relationships”, *Cybernetics and Systems'96*, R. Trappl (Ed.), Austrian Soc. for Cyber. Studies, Vienna, Austria, pp. 985-990, 1996.

[12] Suzuki E. “Autonomous Discovery of Reliable Exception Rules,” Proc. KDD-97, 259-262, 1997.

[13] Suzuki E., Kodratoff Y., “Discovery of Surprising Exception Rules Based on Intensity of Implication”, in *Principles of Data Mining and Knowledge Discovery*, Zytkow J. & Quafafou M. (Eds.), pp. 10-18, LNAI 1510, Springer, Berlin 1998.