# Approche Fouille de Texte pour la détection précoce de tendances économiques

Marilyne Latour*,** Antoine Sigwalt**,***

*Université Grenoble-Alpes, GRESEC, F-38040 Grenoble
marilyne.latour@ac-grenoble.fr,
http://gresec.u-grenoble3.fr/
**ReportLinker, 4 Rue Montrochet, 69002 Lyon
marilyne.latour@reportlinker.com
antoine.sigwalt@reportlinker.com
http://www.reportlinker.com
***Ecole Centrale Lyon
antoine.sigwalt@ecl2014.ec-lyon.fr
http://www.ec-lyon.fr

**Résumé.** Cet article présente un retour d'expérience sur de la fouille de données complexes dans un processus d'extraction des connaissances dans un contexte industriel. Á partir de données volumineuses non structurées issues de dépêches d'actualités économiques et selon certains traitements linguistiques et économétriques, notre objectif est de prédire des tendances économiques dans des séquences d'évènements d'actualités. Pour cela, trois étapes sont primordiales : *(i)* l'extraction d'indicateurs économiques par des techniques linguistiques (comme les indices boursiers, les taux de change, les noms des monnaies ou encore les cours des matières premières...), *(ii)* l'annotation, par le recours à des terminologies externes, de ces indicateurs économiques : les données extraites portent alors des étiquettes permettant de les identifier, *(iii)* leur superposition à des modèles statistiques. Á la suite de ce traitement, nous pouvons vérifier si il existe une corrélation entre des indicateurs économiques relevés par l'étude linguistique pour un secteur d'activité donné et sur un territoire donné (la production d'un élément *A* sur le prix d'un élément *B* par exemple). L'intérêt de cette méthode est d'apporter des outils linguistiques en complément des méthodes statistiques utilisées habituellement pour faire émerger des données cointégrées. L'article décrit ensuite les expérimentations effectuées et tire les premières conclusions sur divers aspects de cette méthode.

## 1 Introduction

Dans les ensembles volumineux et non structurés des données, l'accès facilité et enrichi à l'information est devenu un véritable marché économique : la compréhension d'un secteur d'activité - notamment ses prévisions de marché - est un angle de recherche particulièrement

attrayant, spécialement pour les sociétés et organismes faisant de la veille stratégique. En prenant en compte le comportement du marché boursier mais aussi en le contextualisant à travers un secteur d'activité spécifique, les analyses économiques sont plus précises, et les prévisions économiques peuvent être ainsi améliorées. L'exploration de données notamment pour le domaine industriel représente un patrimoine immensément riche dont les applications restent encore à développer. Dans ce contexte applicatif intéressant, notre objectif est de développer une méthode capable de prédire des tendances économiques à partir d'ensembles spécifiques dans les séquences d'évènements de l'actualité.

Nous présenterons dans cet article une expérience menée au sein du moteur de recherche ReportLinker qui a consisté à mettre en relation des données économiques extraites par des techniques linguistiques, annotées par des terminologies externes puis superposées à des modèles statistiques. Notre objectif est de vérifier s'il existe une corrélation entre deux indicateurs économiques relevés par étude linguistique pour un secteur d'activité donné et sur un territoire donné. Pour cela, nous avons traité dans un premier temps des données volumineuses non structurées à partir de dépêches d'actualités sur un secteur d'activité bien particulier (le marché du caoutchouc dans son exploitation des pneumatiques) pour faire émerger du vocabulaire intrinsèque au marché mais aussi des indicateurs boursiers, des prix de matières premières, etc. Dans un deuxième temps, nous avons confronté ces données à des terminologies externes et enfin nous avons corrélé ces données avec d'autres indicateurs économiques afin de vérifier s'il existait des liens et des effets de causalité. L'intérêt de cette méthode est d'apporter des outils linguistiques en complément de méthodes statistiques utilisées habituellement pour faire émerger des données cointégrées. L'article décrit ensuite les expérimentations effectuées et tire les premières conclusions sur divers aspects de cette méthode.

## 2 Traitements linguistiques et économétriques

Pour créer de la connaissance économique et émettre des prévisions qui seront utiles aux analystes pour décider d'un investissement futur, il convient de mettre en relation de nombreuses données économiques actualisées permettant une analyse rapide et efficace de l'impact de certains évènements non-quantifiables sur le marché par exemple. Or, pendant des décennies, l'analyse de marché boursier a été fondée sur l'historique des prix du marché. Algorithme génétique (Goldberg, 1989), raisonnement à partir de cas (Aamodt et Plaza, 1994), réseaux de neurones (Rumelhart et McClelland, 1986), machines à vecteurs de support (SVM) (Boser et al., 1992) et autres techniques ont examiné le comportement des marchés boursiers. Le problème, en voulant effectuer des prévisions précises basées sur ces approches, est qu'elles ne prennent que trop peu en compte la modélisation du comportement aléatoire du marché comme l'arrivée de nouveaux facteurs pouvant influencer le marché. Á l'heure où la récupération et le stockage des données n'est plus un problème, les possibilités pour construire des séries temporelles de fréquence raisonnablement importantes sont désormais offertes : elles permettent de représenter avec précision l'évolution d'un indicateur économique ou d'un indice financier sur plusieurs années. Afin d'établir des relations d'équilibre de long terme entre les différentes variables économiques, nous cherchons à estimer de manière automatique la corrélation et la causalité pouvant exister entre ces indicateurs, dans un secteur d'activité et sur un territoire donnés. Notre approche se situe donc au croisement de plusieurs disciplines :

– linguistique et terminologique : parce qu'il s'agit de repérer les termes porteurs de sens pour un secteur et un territoire donné : indicateurs boursiers, noms de monnaie, valeurs des matières premières, etc, que nous identifions et annotons par des ressources externes,
– statistique et économétrique : dans la mesure où nous vérifions si des indicateurs économiques, relevés par analyse linguistique, présentent des liens et des effets de causalité.

Plusieurs travaux ont été effectués sur la façon de détecter des nouveaux évènements dans un flux de diffusion des actualités notamment appelé TDT *(Topic Detection and Tracking)* par une initiative DARPA [1], (Allan et al., 1998; Yang et al., 1998; Kleinberg, 2002; Guralnik et Srivastava, 1999; Drury et Almeida, 2011). Des travaux d'exploration de texte permettent de distinguer l'apparition d'un sujet « en vogue » dans un flux de données avec certaines caractéristiques en forte hausse quand le sujet émerge (Radinsky et Horvitz, 2013). Ces études combinent généralement une analyse de contenu et une modélisation des séries chronologiques. D'autres travaux en traitements linguistiques peuvent faire écho à notre travail notamment sur le repérage d'entités nommées comme ceux de (Nadeau et Sekine, 2007) pour la reconnaissance et l'étiquetage des indicateurs, des noms de personne ou de sociétés. Également, plus récemment, des travaux sur la reconnaissance d'évènements avec les travaux de (Serrano et al., 2012) et ceux liés aux systèmes de détection de crise avec (Capet et al., 2012). Il n'existe pas -à notre connaissance- de méthode hybride combinant une analyse fouille de textes et une analyse économétrique pour prédire des tendances économiques dans des séquences d'évènements d'actualité.

L'intérêt de la mise en relation des données linguistiques et économétriques dans cette étude est qu'elle résulte d'une approche par corpus de textes permettant de faire ressortir la spécificité terminologique d'un secteur d'activité. Cette approche nous permet en effet de mettre en évidence des propriétés d'un secteur d'activité à partir du nombre d'occurrences. Pour ce faire, la chaîne de traitements mise en place va se faire en deux temps : un traitement linguistique des termes ou expressions contenus dans les dépêches d'actualité et une mise en relation économétrique.

## 2.1 Mise en place du traitement linguistique

Le traitement linguistique s'opère en cinq étapes :

**(i) La première étape concerne le découpage du corpus en une liste de termes** : notre objectif est d'obtenir dans un premier temps tous les termes issus des dépêches d'actualité. Pour cela, nous considérons le texte comme un « sac de mots » pour obtenir une liste de termes en fonction de leur fréquence d'apparition. Plusieurs longueurs de séquence de termes (appelés *n-grammes*) sont retenues : uni-grammes, bi-grammes, tri-grammes voire quadri-grammes. Pour effectuer ce traitement, nous avons tout simplement utilisé un concordancier [2]. La visualisation en est ascendante : des n-grammes qui occurrent le plus (*i.e.* ceux qui apparaissent le plus souvent dans un corpus donné) jusqu'à ceux qui n'ont pas beaucoup d'occurrences (le ratio nombre d'occurrences / fréquence d'apparitions est alors à définir selon les objectifs de récupération). Nous obtenons à la suite de ce traitement une liste de n-grammes classés par nombre décroissant d'occurrences pour le corpus donné.

1. Defense Advanced Research Projects Agency
2. Dans cette étude, nous avons utilisé le concordancier libre de droit *AntConc*, développé par Laurence Antony et disponible à l'adresse suivante : http ://www.antlab.sci.waseda.ac.jp/software.html.

**(ii) La seconde étape concerne l'appariement entre les termes relevés lors de l'étape précédente et les termes des thésaurus** : ces thésaurus sont développés en interne de façon semi-automatique à partir de lexiques de spécialité et d'analyses manuelles ; ils sont au nombre de trois :

- **un thésaurus sectoriel** : contenant les secteurs d'activités organisés autour des six axes principaux : Agro-alimentaire, Biens et Services de Consommation, Industrie lourde, Technologies de l'Information et Médias, Sciences de la Vie et Services. Ce thésaurus présente une structure hiérarchique pour chaque axe principal et est composé de millions de termes avec des types d'entités différenciés.
- **un thésaurus géographique** : contenant tous les noms de pays (« France »), les formes adjectivales (« French »), les monnaies (« Euros »), les indices boursiers (« CAC 40 »), les codes ISO 3166-1 (« FRA », « FR »), les noms de certaines personnalités comme le président («François Hollande »), le président du Sénat (« Jean-Pierre Bel ») , la capitale (« Paris ») ainsi que les plus grandes villes (« Paris »,« Lyon ») etc.
- **un thésaurus d'Entités Nommées de noms de société** : composé de noms propres de noms de société comme « Cooper Tire », « Goodyear Tire ».

Nous obtenons alors une segmentation des termes en fonction de leur appartenance au thésaurus (une étiquette est alors attribuée) ainsi que leur fréquence d'apparition dans le corpus. Á noter que pour les termes du thésaurus sectoriel, une lemmatisation [3] s'effectue permettant de recouper un maximum de formes (« rubber products » pour « rubber product »). Á cette étape, environ 70% des n-grammes de notre corpus sont traités.

**(iii) La troisième étape concerne la suppression des mots vides** : pour les termes qui n'ont pas trouvé de correspondance avec les thésaurus, une comparaison est établie avec une liste pré-définie de mots vides de sens (*stop-list*) pour être supprimés. Cette étape ne s'applique pas aux mots vides intégrés dans un syntagme plus longs relevés lors de l'analyse avec le concordancier comme les bi-grammes, tri-grammes et quadri-grammes. Environ 10% de notre corpus est alors concerné par cette suppression.

**(iv) La quatrième étape concerne l'analyse morpho-syntaxique** : à la quatrième étape, environ 20% des termes de notre corpus n'ont été ni segmentés en entité par nos thésaurus, ni supprimés par la liste de mots vides. Il peut s'agir d'entités nommées de noms de personne, d'entreprises, de lieux non répertoriés dans nos thésaurus, de noms communs, de sigles inconnus, etc. Á cette étape, une analyse morpho-syntaxique est alors effectuée avec l'analyseur *xelda* [4] permettant d'obtenir toutes les catégories grammaticales des n-grammes ainsi relevés. Des étiquettes (appelés *'tags'*) sont alors attribuées : nom (NOUN), nom propre (PROP), nom adjectif (NOUN ADJ) pour les plus communément utilisées dans le cadre de cette analyse.

**(v) La cinquième étape concerne l'analyse manuelle des étiquettes morpho-syntaxiques** : beaucoup de termes sont étiquetés comme des PROP, nous permettant d'obtenir plus d'entités nommées sur les noms d'entreprise ou noms de personne. En fonction du ratio du nombre d'apparitions relevées sur le nombre de dépêches d'actualité, nous décidons ou non de les inclure dans le thésaurus approprié (ce ratio varie en fonction du domaine traité). Les termes jugés non discriminants et n'apportant pas d'informations pertinentes pour l'analyse d'un secteur d'activité spécifique ne sont pas retenus. L'étiquetage morpho-syntaxique sert donc uniquement dans ce processus à faciliter le travail humain de sélection des termes lors de la cinquième étape.

3. Opération qui consiste à assigner à chaque mot d'un texte son lemme (*i.e.* sa forme de base).
4. *xelda* a été développé par *xerox*.

Une fois cette analyse linguistique terminée, nous obtenons donc une liste de termes avec un type entité appartenant à l'un des trois thésaurus. Des rapprochements purement économiques peuvent alors s'opérer : il s'agit de tester les termes qui ont un plus fort ratio entre la taille du corpus et le nombre d'occurrences.

## 2.2 Mise en relations des données économiques : présentation du modèle statistique utilisé

Pour créer de la connaissance économique, plusieurs recherches et traitements vont s'opérer sur les termes retenus lors de la phase précédente afin d'étudier si il existe des inter-relations d'ordre économique ou financier. Ces recherches peuvent être utiles pour établir des relations de long terme entre les variables économiques temporelles ou pour refléter des fluctuations de production de marché, des hausses de prix, des chutes d'indices boursiers…; tous signes précoces d'activités inhabituelles qui pourraient avoir des futures répercussions sur un marché en particulier. Certains modèles de lissage exponentiel permettent en outre de récupérer des valeurs manquantes dans les séries obtenues, voire d'effectuer des prévisions à court terme (Holt, 1957; Winters, 1960). Mais, pour affiner leur étude, les analystes ont besoin d'établir des relations d'équilibre de long terme entre les variables économiques temporelles, pour savoir dans quelle mesure les indicateurs sont liés entre eux et au marché. C'est dans cette optique que nous appliquons des modèles statistiques pour estimer la corrélation et la causalité entre divers indicateurs, dans un secteur d'activité et sur un territoire donnés.

Les séries chronologiques économiques et financières ne sont généralement pas stationnaires, c'est-à-dire que la loi qui les gouverne dépend du temps. Ce sont bien souvent des processus autorégressifs intégrés d'ordre 1, c'est-à-dire que leur différence première est stationnaire, ou encore que la différence entre deux valeurs consécutives suit une loi constante dans le temps. La propriété de deux séries intégrées d'ordre 1 qui permet d'étudier une relation de long terme stable entre elles est la *cointégration*. Deux séries $x_t$ et $y_t$ intégrées d'ordre 1 sont dites cointégrées si l'aléa $\epsilon$ de la régression de l'une sur l'autre des séries est stationnaire, *i.e.* s'il existe un réel $\beta$ tel que $\epsilon = y - \beta x$ est stationnaire.

Le modèle statistique que nous utilisons est la *méthode d'Engle et Granger (1987) en deux étapes* qui consiste dans un premier temps à estimer [5] par moindres carrés ordinaires (MCO) le modèle 1

$$y = \beta x + \epsilon \tag{1}$$

puis à tester par le *test de Dickey et Fuller (1979) augmenté* la stationnarité du résidu estimé $\hat{\epsilon}$. La dernière étape de cette procédure consiste à estimer le modèle 2 par MCO [6]

$$\Delta\hat{\epsilon}_t = \rho y_{t-1} + \sum_{i=1}^{p-1} \phi_i \Delta y_{t-i-1} + u_t \tag{2}$$

5. Estimer le modèle par MCO signifie que l'on cherche la valeur des paramètres (ici, $\beta$ est le seul paramètre du modèle) qui permet de minimiser la somme des carrés des aléas $\epsilon_t = y_t - \beta x_t$. Cette estimation nous donne deux résultats : l'estimateur de $\beta$, noté $\hat{\beta}$, qui permet de calculer le résidu estimé $\epsilon_t = y_t - \hat{\beta} x_t$, et l'estimateur de l'écart-type de $\hat{\beta}$, noté $\hat{\sigma}_{\hat{\beta}}$, qui n'est autre ici que la moyenne des carrés des résidus estimés : $\hat{\sigma}_{\hat{\beta}} = \sum_{t=1}^{T} \hat{\epsilon}_t^2 = \sum_{t=1}^{T} (y_t - \hat{\beta} x_t)^2$

6. Comme précédemment, on peut obtenir $\hat{\rho}$ l'estimateur de $\rho$, et $\hat{\sigma}_{\hat{\rho}}$ l'estimateur de l'écart-type de $\hat{\rho}$.

où $\hat{\epsilon}_t$ est le résidu estimé du modèle 1 et $p$ est le nombre de retards du processus autorégressif [7], déterminé par les critères d'information usuels AIC et BIC. La statistique-test $\tau = \hat{\rho}/\hat{\sigma}_{\hat{\rho}}$ suit la loi décrite dans les tables de Fuller (1976) ou MacKinnon (1991). Si l'hypothèse nulle $\rho = 0$ est rejetée (*i.e.* si la statistique-test est supérieure à la valeur critique à 10% donnée par les tables de Dickey-Fuller ou MacKinnon, ou encore, si la p-value du test, approximée à partir de ces valeurs critiques [8], est inférieure à 10%), alors la série $\hat{\epsilon}_t$ est stationnaire, ce qui implique que les séries $x_t$ et $y_t$ sont cointégrées. Cette méthode sera testée et les résultats présentés dans la partie 4.

# 3 Présentation de l'expérimentation

L'expérimentation s'est déroulée dans un cadre applicatif bien particulier : celui d'un moteur de recherche, ReportLinker, agrégateur de données économiques.

## 3.1 Cadre de l'expérience : le moteur de recherche ReportLinker

ReportLinker [9] est un moteur de recherche spécialisé en études économiques qui a été lancé en 2007 par la société française UBIQUICK. Ce moteur agrège 1,2 million d'études de marché en anglais. Il fournit un accès direct et organisé aux documents économiques édités par 200 000 sources d'information différentes (ministères, syndicats professionnels, ambassades) sous formes de rapports de marché, de statistiques, d'études sur les tendances d'un marché, ou encore de profils d'entreprises. Ces études économiques sont organisées autour de six axes principaux : Agro-alimentaire, Biens et Services de Consommation, Industrie lourde, Technologies de l'Information et Médias, Sciences de la Vie et Services. Ils couvrent environ 450 secteurs d'activités économiques. La plateforme ReportLinker.com analyse et indexe chaque jour plusieurs milliers de documents pour fournir à ses clients un puissant outil de recherche et d'exploitation du « Web des documents » répondant plus rapidement et plus spécifiquement à leurs besoins informationnels. Un des objectifs du moteur de recherche est de corréler rapidement les informations entre elles, de les organiser, les hiérarchiser pour mieux les appréhender voire de détecter des nouvelles tendances ainsi que de nouveaux comportements émergents (il peut s'agir de comportements face à des produits récemment mis sur le marché ou de phénomène de crise dans un pays : crise sociale, politique ou géo-politique par exemple).

## 3.2 Présentation du corpus

Nous avons réalisé pour cette expérimentation un corpus composé de 10 000 dépêches d'actualité en anglais recouvrant les thèmes du caoutchouc (*rubber*) et du pneumatique (*tire*) sur les années 2011 et 2012. Ces dépêches d'actualité économiques proviennent de plusieurs

7. Un processus autorégressif $X_t$ est défini à la date t par ses réalisations aux dates $t-1, t-2, ..., t-i, ..., t-p$ pondérées par des coefficients $\phi_1, \phi_2, ..., \phi_i, ..., \phi_p$, *i.e.* $X_t = \phi_1 X_{t-1} + \phi_2 X_{t-2} + ... + \phi_i X_{t-i} + ... + \phi_p X_{t-p} + e_t$ où $p$ est appelé le nombre de retards.

8. Il est important de souligner que les p-values présentées dans la suite de ce papier sont calculées à partir des valeurs critiques de MacKinnon (1991)

9. http ://www.reportlinker.com

sources spécialisées en économie : Associated Press, AFP, Xinhua News. Elles ne sont ni structurées, ni catégorisées, ni harmonisées d'un point de vue de leur longueur ; elles sont simplement stockées en format xml en base de données après un script de pré-traitement pour le nettoyage des caractères spéciaux. Nous sélectionnons des dépêches d'actualité qui mentionnent au moins une fois les termes *rubber* et *tire* dans le même texte. L'objectif est en effet de contextualiser le marché du caoutchouc dans le secteur des pneumatiques. Nous décidons volontairement de garder les dépêches d'actualité ainsi constituées et de mettre à plat les informations sans les pondérer : elles gardent ainsi toute la richesse du vocabulaire (sans prioriser celui issu du titre ou des méta-données par exemple) afin d'observer l'ensemble du vocabulaire issu de ces dépêches. Nous excluons de ce corpus les doublons (même information, même source) [10].

# 4 Résultats

Nous présentons dans les deux sous-parties suivantes les résultats obtenus lors des analyses linguistiques puis les différents traitements statistiques réalisés pour cointégrer les indicateurs économiques ainsi obtenus.

## 4.1 Résultats des traitements linguistiques

Á la première étape (découpage du corpus en une liste de termes) : nous obtenons des termes qui n'ont pas encore d'étiquettes comme « BSE Sensex », « rubber products », « Goodyear Tire », « Rs crores », etc.

Á la seconde étape (appariement entre les termes relevés lors de l'étape précédente et les termes des thésaurus) : à partir du thésaurus sectoriel, nous relevons pour le secteur du pneumatique les termes suivants : « tire », également des synonymes « tyre », les applications « Motorcycle tyre », les composants « synthetic rubber », « natural rubber », etc. Nous obtenons également à partir du thésaurus géographique des mentions de noms de pays comme « India », la devise comme la roupie indienne ou encore son code ISO 417 : INR. Enfin, à partir du thésaurus d'Entités Nommées de noms de société, nous arrivons à obtenir des noms de société comme « Cooper Tire », « Goodyear Tire ». Nous présentons quelques-uns des résultats dans le Tableau 1.

La troisième étape supprime les mots vides et la quatrième étape permet l'étiquetage morpho-syntaxique comme illustré dans le Tableau 2.

Les PROP étant les noms propres et les NOUN noms communs.

Á la cinquième étape (analyse manuelle des résultats de l'analyse morpho-syntaxique) : les termes comme « million » ou « indicator » ne sont pas retenus car jugés peu discriminants. D'autres PROP comme RTA qui est le sigle de *Rubber Trade Association* ou des NOUN comme « storm » (*i.e.* tempête) peuvent être conservés en vue d'analyses complémentaires comme la détection d'évènements.

10. Une information est jugée « doublon » si elle présente les mêmes entités nommées de nom de personne, de lieu, le même secteur d'activité et publiée dans un intervalle de temps égal ou inférieur à 7 jours. Le secteur d'activité est attribué par un algorithme de catégorisation interne ; il n'est pas développé ici car n'est pas utilisé dans la suite du processus.

| Thésaurus | Type-Entité | N-grammes | Fréquence |
|---|---|---|---|
| Thésaurus Géographique | Indice Boursier | BSE Sensex | 11598 |
| Thésaurus Géographique | Monnaie | Rs crores | 9947 |
| Thésaurus Géographique | Indice Boursier | BSE Sensex Index | 7470 |
| Thésaurus Sectoriel | Entité sectorielle | Rubber | 7241 |
| Thésaurus Sectoriel | Entité sectorielle | Rubber products | 6928 |
| Thésaurus Sectoriel | Entité sectorielle | Natural rubber | 6523 |
| Thésaurus Sectoriel | Entité sectorielle | Auto Tyres | 6215 |
| Thésaurus Sectoriel | Indicateur sectoriel | tyres sales | 3753 |
| Thésaurus Entité Nommée | Entité nommée société | Goodyear Tire | 795 |

TAB. 1 – *Etude linguistique du terme pneumatique*

| N-grammes | Fréquence | Tags Xelda |
|---|---|---|
| NanKang Rubber | 492 | PROP NOUN |
| million | 388 | NOUN |
| indicator | 362 | NOUN |
| RTA | 109 | PROP |
| storm | 58 | NOUN |

TAB. 2 – *Analyse morpho-syntaxique*

Une fois l'analyse linguistique terminée, nous procédons à l'analyse économétrique pour tenter de caractériser statistiquement les relations entre les indicateurs que la fouille de texte a fait émerger.

## 4.2 Récupérations des indicateurs et obtention des données d'étude

L'évolution des prix (ou des cours de la Bourse) étant une mesure universelle de la santé d'un secteur d'activité (caractérisation de l'offre et de la demande sur le marché des matières premières), de la santé d'une entreprises ou de toute une région économique, nous faisons le choix de cette mesure (type entité Indice Boursier du thésaurus Géographique) dans toutes nos études. Le BSE Sensex étant le terme qui a le plus d'occurrences parmi les termes relatifs à l'économie sud-asiatique obtenus, nous le préférons aux autres indicateurs possibles comme, par exemple, le taux de change de la roupie. L'étude statistique consiste ainsi à répondre aux questions :

- Le prix du caoutchouc naturel établi sur le marché international est-il corrélé au marché financier sud-asiatique, *i.e.* au BSE Sensex ?
- Y-a-t-il une relation de causalité entre ces deux indicateurs ? Si oui, dans quel sens s'effectue-t-elle ?

Pour notre exemple, nous avons récupéré les données mensuelles du prix international du caoutchouc (base 100 en 2000) sur le site de l'INSEE, et les données mensuelles du cours ajusté du BSE Sensex sur le site Yahoo ! Finance. Les séries sont reproduites en figure 1. Il s'agit de données numériques open-data ; se présentant sous la forme de séries temporelles à fréquence

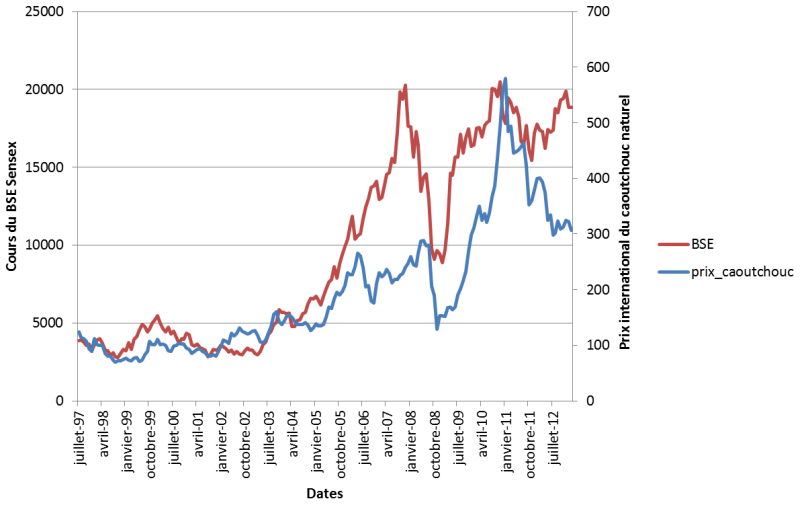


FIG. 1 – *Evolution du cours du BSE Sensex et du prix international du caoutchouc naturel entre 1997 et 2013.*

constante (généralement mensuelle ou trimestrielle) et retraçant l'évolution chronologique des indicateurs considérés, sur un intervalle de plusieurs années.

Si, à première vue, la régression linéaire présentée en figure 2 laisse apparaître que les deux variables évoluent de la même façon, Granger et Newbold (1974) montrent que cette régression peut être factice et nécessite une étude approfondie.

## 4.3 Résultats de la recherche de relation de long terme : cointégration au sens d'Engle et Granger

### 4.3.1 Application de la méthode

Un test de stationnarité de Dickey-Fuller augmenté, appliqué à chacune des deux séries considérées ici et à leur différence première, est présenté dans le tableau 3. Le test consiste ici à estimer le modèle 2 présenté en section 2.2, et tester la significativité du coefficient $\phi$ à l'aide des seuils de MacKinnon (1991).

| | Niveau | | Différence première | |
|---|---|---|---|---|
| | Statistique | p-value | Statistique | p-value |
| Prix du caoutchouc | -2,6042 | 0,3234 | -5,7679 | < 0,01 |
| BSE | -2,9619 | 0,1738 | -5,2807 | < 0,01 |

TAB. 3 – *Test de Dickey-Fuller augmenté appliqué aux séries du prix du caoutchouc et du BSE Sensex et à leur différence première.*

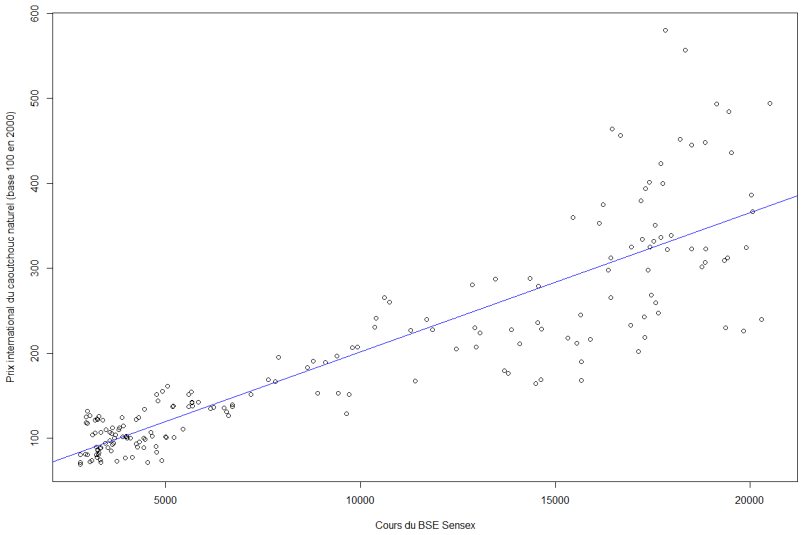


FIG. 2 – *Evolution du prix réel international du caoutchouc en fonction de l'indice du marché indien.*

Le test conclut dans notre cas à la non-stationnarité des séries (la p-value approximée est supérieure à 0.10) et à la stationnarité des différences premières (p-value < 0.10). Les deux séries sont donc bien intégrées d'ordre 1.

Les résidus estimés par les MCO (première étape de la procédure d'Engler et Granger) sont représentés en figure 3.

Les résultats de la deuxième étape de la méthode d'Engle et Granger sont présentés dans le tableau 4. La p-value est inférieure à 10%, on conclut donc à la cointégration des séries.

| Statistique $\tau$ | p-value |
|---|---|
| -18,0181 | 0,08302 |

TAB. 4 – *Résultat du test de cointégration du* BSE *Sensex et du prix du caoutchouc par la méthode d'Engle et Granger.*

### 4.3.2 Analyse des résultats

Nous avons ainsi montré par une analyse statistique qu'il existe une relation de long terme entre le prix du caoutchouc sur le marché international et le cours de l'indice boursier indien, ce que la fouille de texte présentée en section 4.1 avait suggéré. Nos deux outils semblent donc se prononcer en faveur de l'existence d'une relation économique très forte entre le prix du caoutchouc et la santé économique de l'Asie du Sud-Est. Mais rien ne dit dans quel sens s'établit cette relation. Est-ce le marché sud-asiatique qui influence le cours du caoutchouc ou une évolution du cours du caoutchouc qui impacte le Sensex ? Si, dans notre exemple, le bon

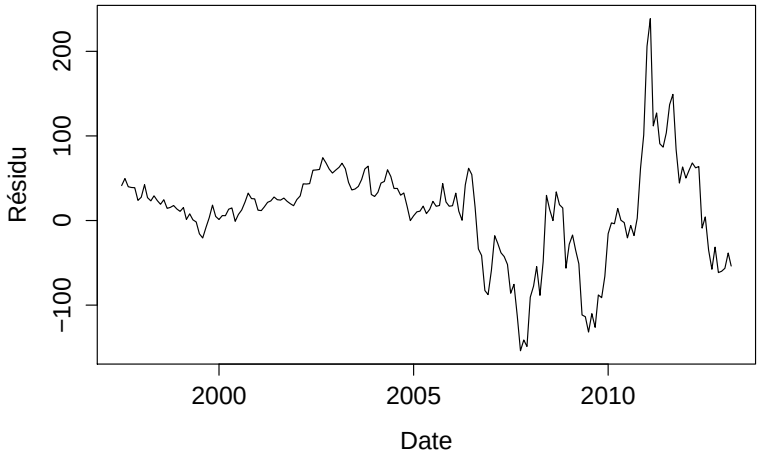


FIG. 3 – *Evolution du résidu de la régression linéaire du prix réel international du caoutchouc en fonction du BSE Sensex.*

sens nous fait supposer que le marché du caoutchouc n'est pas assez important pour influencer à lui seul l'ensemble du marché sud-asiatique et que la causalité est plutôt dans l'autre sens, d'autres cas d'étude peuvent nécessiter des recherches supplémentaires. Le test de causalité de Granger (1969) apporte alors des éléments de réponse.

#### 4.3.3 Orienter la relation économique : la causalité au sens de Granger

Lorsque deux séries $x_t$ et $y_t$ sont cointégrées, on dit que « $x$ cause $y$ au sens de Granger » si les valeurs antérieures de $x$ permettent de prédire les valeurs futures de $y$. L'hypothèse nulle de non-causalité au sens de Granger est rejetée si et seulement si l'une des variables $x_{t-1}$, $x_{t-2}$, ..., $x_{t-q}$ est significative dans le modèle 3.

$$y_t = a_0 + a_1 y_{t-1} + a_2 y_{t-2} + ... + a_p y_{t-p} + b_1 x_{t-1} + b_2 x_{t-2} + b_q x_{t-q} + v_t \qquad (3)$$

où $p$ et $q$ sont le nombre de retards des processus $x$ et $y$, déterminés par les critères d'information AIC et BIC.

Les résultats du test de causalité de Granger appliqué aux séries étudiées ici sont présentés dans le tableau 5. La p-value du test est très inférieure à 0.10 dans le modèle d'explication du prix du caoutchouc par le BSE Sensex, donc l'hypothèse nulle de non-causalité au sens de Granger peut être refusée. Autrement dit, les valeurs antérieures du BSE Sensex permettent d'expliquer les valeurs actuelles du prix du caoutchouc, ou encore, la série du BSE Sensex permet de prévoir les prix futurs du caoutchouc. L'hypothèse de non-causalité dans le sens réciproque est en revanche acceptée, la p-value étant supérieure à 0.10.

| | Statistique F | p-value |
|---|---|---|
| BSE Sensex -> Prix du caoutchouc | 4.26550 | 0.0005 |
| Prix du caoutchouc -> BSE Sensex | 0.42275 | 0.8631 |

TAB. 5 – *Test de causalité de Granger entre les séries du BSE Sensex et du prix du caoutchouc.*

# 5 Discussion : Validité des résultats et prévision économique

## 5.1 Confrontation des résultats obtenus à la réalité économique

La fouille de texte nous a indiqué un lien probable entre les évolutions du marché indien et du prix du caoutchouc, qui a été confirmé par une étude statistique. Ce lien a été orienté par un second test statistique, indiquant qu'une variation du BSE Sensex cause *au sens de Granger* une variation du prix du caoutchouc sur le marché international. Ce résultat est cohérent avec la réalité économique puisque plus de 85% de la production de caoutchouc naturel se trouve dans les pays du sud-est asiatique, comme l'indique le tableau 6 (source FAOSTAT [11]).

| Pays | Production caoutchouc naturel (tonnes) | Production mondiale (en %) |
|---|---|---|
| Thaïlande | 3 348 897 | 29,69 % |
| Indonésie | 3 088 400 | 27,38 % |
| Malaisie | 996 673 | 8,84 % |
| Inde | 891 344 | 7,90 % |
| Viêt Nam | 798635 | 7,00 % |
| Sri Lanka | 158 198 | 1,40 % |
| Birmanie | 149 627 | 1,33 % |
| Philippines | 140 500 | 1,25 % |
| Cambodge | 43 471 | 0,39 % |
| Asie du Sud-Est | 9 606 745 | 85,17 % |
| Total | 11 279 525 | 100 % |

TAB. 6 – *Production de caoutchouc naturel en 2011 (source FAOSTAT)*

Les outils développés ici ne nous permettent cependant pas d'affirmer que la bonne santé du marché de l'Asie du Sud-Est est favorable à des prix du caoutchouc hauts ou bas. En effet, l'estimateur des MCO de $\beta$ dans le modèle 1 présenté en section 2.2 est biaisé. Il suit une loi complexe dont on ne connait pas la distribution, on ne peut donc pas tester sa significativité. En revanche, on sait que les résidus estimés $\hat{\epsilon}_t$ mesurent le déséquilibre à la date t dans la relation de long terme. On peut alors quantifier la vitesse d'ajustement à l'équilibre en estimant un modèle à correction d'erreur.

De la même manière, Granger (1986) démontre statistiquement la cointégration de certaines paires d'indicateurs économiques aux Etats-Unis. Cependant, il ne trouve aucune cointégration entre des indicateurs pourtant théoriquement liés (les salaires et les prix, la quantité de monnaie et les prix), invoquant le rôle d'autres variables dans ces relations et donc l'existence de relations de cointégration de plus de deux variables. Une analyse linguistique sur un

11. Food and Agriculture Organization Corporate Statistical Database : faostat3.fao.org

corpus de dépêches d'actualité montre en effet que le taux de change apparaît fréquemment autour de l'inflation, en troisième position derrière la politique monétaire et les taux d'intérêt. Il convient alors de s'intéresser à un modèle liant ces différentes variables.

Pour étudier un tel modèle systémique, les approches les plus abouties sont celles de Johansen (1988) et Johansen et Juselius (1990), connues sous le nom de « test de la valeur propre » et « test de la trace », basées sur l'estimation d'une autorégression vectorielle (modèle VAR).

## 5.2 Quelle stabilité de la méthode sur d'autres ressources ?

Nous avons appliqué notre méthode à une série de tests supplémentaires (une dizaine) afin d'assurer la stabilité de notre méthode. Nous présentons dans cette sous section le résultat de ces tests supplémentaires portant sur des matières premières et des métaux.

Nous testons la cointégration entre la matière première et les cinq principaux indicateurs (indices boursiers, cours d'entreprises) qui se démarquent lors de l'analyse linguistique. Lorsque différents indices font référence à un même marché (par exemple, China + CNYuan + SZSE Index font référence au marché chinois), on les regroupe et l'on teste la cointégration entre la série des prix de la matière première et la série de l'indice boursier (ici, le SZSE Index). On rappelle en outre qu'une relation de cointégration existe lorsque la p-value du test de cointégration est inférieure à 10%.

Les sept premiers tests présentés dans les tableaux 7, 8, 9, 10, 11, 12 et 13 permettent d'identifier les relations par ressource testée.

| Leather | China + CNYuan SZSE Index | Pretto Leather Industries Ltd | Bombay Stock Exchange |
|---|---|---|---|
| Occurences /1000 news | 773 + 647 + 365 | 450 | 462 |
| p-value du test | <0,01 | 0,063 | 0,0181 |

TAB. 7 – *Résultats des tests linguistique et statistique mettant en évidence la relation économique entre le cuir, le marché chinois, le marché indien, et Pretto Leather Industries Ltd*

| Palm oil | Bursa Malaysia (KLSE) | Negri Sembilan Oil Palms |
|---|---|---|
| Occurences /1000 news | 373 | 205 |
| p-value du test | >0,10 | 0,0971 |

TAB. 8 – *Résultats des tests linguistique et statistique mettant en évidence la relation économique entre l'huile de palme et Negri Sembilan Oil Palms, et rejetant un lien direct avec le marché malaisien*

Chacune des relations identifiées lors de ces tests repose sur une réalité économique avérée : l'entreprise est un gros producteur de la matière première, ou le pays représente une partie majoritaire de la production mondiale de la matière première.

Cependant, d'autres tests linguistiques faisant apparaître des relations économiques probables ne sont pas confirmés par le test de cointégration statistique, ce qui soulève la limite de

| Tea | India + Bombay Stock Exchange | Joonktollee Tea Industries Ltd | Dhunseri Petrochem and Tea Ltd |
|---|---|---|---|
| Occurences /1000 news | 1372 + 1305 | 585 | 504 |
| p-value du test | 0,0228 | >0,10 | >0,10 |

TAB. 9 – *Résultats des tests linguistique et statistique mettant en évidence la relation économique entre le thé et le marché indien, et rejetant un lien direct avec Joonktollee Tea Industries Ltd et Dhunseri Petrochem and Tea Ltd*

| Aluminium | Hind Aluminium Industries | Ess Dee Aluminium | Bombay Stock Exchange |
|---|---|---|---|
| Occurences /1000 news | 977 | 865 | 433 |
| p-value du test | <0,01 | 0,0973 | 0,0469 |

TAB. 10 – *Résultats des tests linguistique et statistique mettant en évidence la relation économique entre l'aluminium, Hind Aluminium Industries, Ess Dee Aluminium et le marché indien*

| Platinium | Anglo American Platinum Ltd | South Africa + FTSE JSE + JSE Africa Top Index | Impala Platinum Holdings |
|---|---|---|---|
| Occurences /1000 news | 1168 | 777 + 590 + 381 | 412 |
| p-value du test | <0,01 | >0,10 | 0,079 |

TAB. 11 – *Résultats des tests linguistique et statistique mettant en évidence la relation économique entre le platinium, Anglo American Platinum Ltd et Impala Platinum Holdings, et rejetant un lien direct avec le marché sud-africain*

| Iron | BSE Sensex | Kumba | Vale | Nikkei 225 | Nippon Steel Corporation |
|---|---|---|---|---|---|
| Occurences /1000 news | 570 | 294 | 243 | 186 | 174 |
| p-value du test | <0,01 | 0,0176 | <0,01 | >0,10 | 0,0526 |

TAB. 12 – *Résultats des tests linguistique et statistique mettant en évidence la relation économique entre le fer, le marché indien et les entreprises Kumba, Vale et Nippon Steel Corporation, et rejetant un lien direct avec le marché japonais*

cette méthode. En effet, il peut y avoir plusieurs justifications à cette absence de cointégration. Tout d'abord, la possible relation économique soulevée par l'analyse linguistique peut être factice, auquel cas le test statistique permet d'identifier cette illusion. Par ailleurs, comme tout

| Sugar | India + rupees |
|---|---|
| Occurences /1000 news | 293 + 311 |
| p-value du test | 0,06 |

TAB. 13 – *Résultats des tests linguistique et statistique mettant en évidence la relation économique entre le sucre et le marché indien*

test statistique réalisé au seuil de confiance de 10%, il reste 10% de risque d'erreur. Enfin, la fiabilité des données peut être trop faible. C'est donc ensuite à l'économiste averti d'expliquer ces résultats et de tirer les conclusions les plus probables. De tels exemples sont donnés dans les tableaux 14, 15 et 16.

| Zinc | Hindustan Zinc Ltd | CNY Exchange Rate | Yunnan Chihong Zin | Huludao Zinc Industry |
|---|---|---|---|---|
| Occurences /1000 news | 340 | 239 | 200 | 158 |
| p-value du test | >0,10 | >0,10 | >0,10 | >0,10 |

TAB. 14 – *Résultats des tests linguistique et statistique rejetant un lien direct entre le zinc, Hindustan Zinc Ltd, le marché chinois, Yunnan Chihong Zinc et Huludao Zinc Industry*

| Titanium | CNY Exchange Rate | Japanese Yen + Nikkei Index | Anhui Annada Titanium Industry |
|---|---|---|---|
| Occurences /1000 news | 878 | 498 + 263 | 368 |
| p-value du test | >0,10 | >0,10 | >0,10 |

TAB. 15 – *Résultats des tests linguistique et statistique rejetant un lien direct entre le titanium, le marché chinois, le marché japonais, et Anhui Annada Titanium Industry*

| Coffee | Tata Coffee Ltd | Ten Peaks Coffee Company Inc | Bombay Stock Exchange | Vietnam |
|---|---|---|---|---|
| Occurences /1000 news | 494 | 237 | 201 | 156 |
| p-value du test | >0,10 | >0,10 | >0,10 | >0,10 |

TAB. 16 – *Résultats des tests linguistique et statistique rejetant un lien direct entre le café, Tata Coffee Ltd, Ten Peaks Coffee Company Inc, le marché indien et le marché vietnamien*

## 6 Conclusion et perspectives

L'intérêt de la méthode décrite est d'apporter des outils linguistiques en complément des méthodes statistiques utilisées en fouille de texte, afin de valider la corrélation des indicateurs étudiés sur un secteur. Ces résultats sont plutôt satisfaisants puisque nous avons réussi à corréler des données économiques à partir d'une étude linguistique sur un corpus de dépêches d'actualité. Le couplage des approches linguistique et statistique permet d'identifier d'éventuels faux positifs statistiques. Cette méthode devrait donc permettre de renforcer l'établissement de relations économiques, en utilisant deux outils d'apparence très différents mais en réalité finement complémentaires. Connaissant alors les indicateurs fortement corrélés, il est enfin possible de réaliser des prévisions sur le prix du caoutchouc, avec un niveau de confiance qui peut être calculé en fonction du seuil des tests statistiques (p-value) et de la fréquence d'apparition du mot « BSE Sensex » autour du mot « rubber ». Des améliorations sont envisagées. La première étant de sémantiquement expliciter les termes sélectionnés notamment grâce à des règles d'association ou des ontologies. La seconde étant d'analyser plus finement (*i.e.* par analyse syntaxique) les phrases donnant des informations sur des évènements d'ordre économique, politique ou climatique. Ce niveau conceptuel concerne en effet l'articulation des séquences de mots et de séquences grammaticales afin d'en valider leur formation : le texte ne serait plus alors un « sac de mots » mais des structures grammaticales complexes où, comme le proposent Drury et Almeida (2011), l'interprétation des évènements pourrait alors être envisagée. Ce procédé pourrait ainsi venir accompagner les résultats déjà obtenus et éventuellement éclaircir certaines relations encore inexpliquées.

## Références

Aamodt, A. et E. Plaza (1994). Case-based reasoning: Foundational issues, methodological variations, and system approaches. *Artificial Intelligence Communications 7*, 39–52.

Allan, J., R. Papka, et V. Lavrenko (1998). On-line new event detection and tracking. *SIGIR'98, Melbourne (Australia) 1*, 37–45.

Boser, B. E., I. M. Guyon, et V. N. Vapnik (1992). A training algorithm for optimal margin classifiers. In *Fifth Annual Workshop on Computational Learning Theory*, pp. 144–152.

Capet, P., T. Delavallade, M. Genereux, M. Poibeau, T. Sandor, et S. Voyatzi (2012). Un systeme de detection de crise base sur l'extraction automatique d'evenements. *Semantique et multimodalite en analyse de l'information 1*, 292–313.

Dickey, D. A. et W. A. Fuller (1979). Distribution of the estimators for autoregressive time series with a unit root. *Journal of the American Statistical Association 84*, 427–431.

Drury, B. et J. J. Almeida (2011). Identification of fine grained feature based event and sentiment phrases from business news stories. In *WIMS*, pp. 27.

Engle, R. F. et C. W. J. Granger (1987). Cointegration and error correction: Representation, estimation and testing. *Econometrica 55*, 251–276.

Fuller, W. A. (1976). Introduction to statistical time series. *John Wiley and Sons 373*.

Goldberg, D. E. (1989). *Genetic Algorithms in Search, Optimization, and Machine Learning*. Addison-Wesley.

Granger, C. et P. Newbold (1974). Spurious regressions in econometrics. *Journal of Econometrics 2*, 111–120.

Granger, C. W. J. (1969). Investigating causal relations by econometric models and cross-spectral methods. *Econometrica 37*, 424–438.

Granger, C. W. J. (1986). Developments in the study of cointegrated economic variables. *Oxford Bulletin of Economics and Statistics 68*, 213–228.

Guralnik, V. et J. Srivastava (1999). Event detection from time series data. *Proceeding KDD '99 Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining*, 33–42.

Holt, C. C. (1957). Forecasting trends and seasonals by exponentially weighted moving averages. *ONR Research Memorandum, Carnegie Institute of Technology 52*.

Johansen, S. (1988). Statistical analysis of cointegration vectors. *Journal of Economic Dynamics and Control 12*, 231–254.

Johansen, S. et K. Juselius (1990). Maximum likelihood estimation and inference on cointegration–with applications to the demand for money. *Oxford Bulletin of Economics and Statistics 52*, 169–210.

Kleinberg, J. (2002). Bursty and hierarchical structure in streams. *Proceeding KDD '02 Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, 91–101.

MacKinnon, J. G. (1991). Critical values for cointegration tests. *Long-run Economic Relationships: Readings in Cointegration*, 267–276.

Nadeau, D. et S. Sekine (2007). A survey of named entity recognition in english and classification. *Lingvisticae Investigationes 30*, 239–245.

Radinsky, K. et E. Horvitz (2013). Mining the web to predict future events. *Proceedings of the sixth ACM international conference on Web search and data mining*, 255–264.

Rumelhart, D. et J. McClelland (1986). Parallel distributed processing: Explorations in the microstructure of cognition. *Cambridge: MIT Press*, 576 p.

Serrano, L., T. Charnois, S. Brunessaux, B. Grilheres, et M. Bouzid (2012). Combinaison d'approches pour l'extraction automatique d'evenements. *JEP-TAL-RECITAL 2*, 423–430.

Winters, P. R. (1960). Forecasting sales by exponentially weighted moving averages. *Management Science 6*, 324–342.

Yang, Y., T. Pierce, et C. J. (1998). A study of retrospective and on-line event detection. *Proceeding SIGIR '98 Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*, 28–36.

## Summary

This paper describes a feedback on complex data-mining in a knowledge extraction process in an industrial context. The experience has been to treat initially a large, complex and unstructured data sets from news on a particular industry from wich we tried to extract information and to 'qualify' it (*i.e.* to give it a meaning from economic context). We present here

the method used by a search engine: ReportLinker. Then, the article describes the experiments performed and points out the first conclusions on this method.