Extraction de l'information implicite par analyse textuelle de sites internet en UNICODE

Bernard Dousset

EXTRACTION DE L'INFORMATION IMPLICITE PAR ANALYSE TEXTUELLE DE SITES WEB EN UNICODE

Bernard DOUSSET, dousset@irit.fr
Institut de Recherche en Informatique de Toulouse
IRIT-SIG-EVI, Université Paul Sabatier, 118, route de Narbonne, 31062 Toulouse cedex 9 (France)

Mots-clés :
UNICODE, langues étrangères, information implicite, signaux faibles, information semi structurée, fouille de données, fouille de texte.

Keywords :
UNICODE, foreign languages, implicit information, weak signals, semi-structured information, data mining, text mining.

Palabras clave :
UNICODE, lenguas extranjeras, información implícita, señales débiles, semi-estructuradas información, minería de datos, minería de texto.

Résumé

Afin de diversifier les sources d'information que permet de traiter notre plate-forme "Tétralogie" dédiée à la veille stratégique, nous nous proposons de l'adapter, dans un premier temps, au traitement de la langue chinoise en nous basant sur son codage informatique en UNICODE. Il est alors possible de travailler sur des segments de texte repérés dans les différents dictionnaires disponibles (généraux ou spécifiques d'un domaine) et que nous pouvons alimenter par le traitement de bases d'information semi structurées et balisées en champs sémantiques : auteurs, mots-clés, organismes, journaux, … Par la suite, des langues comme le japonais, le coréen, l'arabe seront abordées suivant le même principe, une exploitation multilingue pouvant même être envisagée via l'anglais ou un autre langage pivot.

Abstract

In order to diversify the sources of information that can handle our platform "Tetralogie" dedicated to the business intelligence, we will adapt, as a first step, treatment of the Chinese language by using computer coding UNICODE. It is then possible to work on text segments identified in the various available dictionaries (general or specific domain) and we can feed the processor databases and semi-structured semantic marked fields: authors, words key agencies, newspapers, ... As a result, languages such as Japanese, Korean, Arabic will be addressed on the same principle, a multilingual exploitation can be seen via the English language or another pivot.

1. Problématique

Nous proposons une adaptation de notre plate-forme d'analyse stratégique ''Tétralogie'' au traitement des langues non européennes comme l'arabe ou le chinois dont le codage peut être ramené au système UNICODE UTF-8 en utilisant les capacités de transcodage d'outils comme *MS-Word*.
En se basant sur le codage informatique unifié de ces langues, nous pouvons détecter les entités nommées via le traitement de bases semi structurées comme les publications scientifiques, les brevets et la presse. Il est alors possible d'établir des dictionnaires assez complets de termes scientifiques et techniques (mots-clés, classifications, …), de noms d'acteurs (auteurs, personnalités, entreprises, organismes publics, laboratoires, …), de lieux géographiques, de dates et de les traduire dans la langue de l'utilisateur. Le croisement de ces briques signifiantes de texte doit nous conduire à l'établissement des réseaux sociaux et des réseaux sémantiques qui expliquent en grande partie les stratégies sous-jacentes du domaine. Si, de plus, le facteur temps est pris en compte (évolution des réseaux), nous pouvons avoir une vision prospective (tendances, trajectoires des acteurs, type d'évolution, ruptures technologiques, …). Nous pouvons ensuite analyser le texte libre (titre, résumé, texte intégral) afin d'en extraire les nouveaux termes et donc les nouvelles tendances. Un premier traitement des textes à analyser permet de baliser l'ensemble des termes déjà identifiés (conscient collectif) et de les marquer comme entités indissociables (mots composés ou segments). Un second traitement, de type statistique, permet de découvrir de nouvelles séquences émergentes non présentes dans les dictionnaires initiaux mais représentant une nouvelle terminologie qui, si elle s'organise, permet de détecter les fameux signaux faibles (concepts émergents dans le domaine étudié). Nous appelons cela l'inconscient collectif, car il ne s'agit ni de mots-clés, ni de termes proposés par les auteurs eux-mêmes (free-terms), mais d'un vocabulaire nouveau souvent très précis et qui n'est utilisé que par une infime partie de la communauté (les personnes qui ont lu ou entendu des informations dignes d'intérêt et qu'elles reprennent à leur compte). Il se peut même que ces personnes ne se soient jamais concertées, et nous savons pourtant qu'elles valident, par leurs écrits, une idée émise dans un document (scientifique, technique ou d'actualité) ou lors d'une conférence. Il s'agit d'un consensus tacite, qui pourra être validé par croisement avec les signaux forts (connexions aux autres domaines) ou par identification de l'origine (qualité des acteurs concernés).
Pour illustrer notre propos, nous allons présenter une analyse d'un corpus d'information extrait de la base scientifique chinoise CQVIP. Nous donnerons aussi quelques indications pour traiter de la même manière d'autres sources comme la base chinoise CNKI, le site d'information en arabe Al Jazeera ou une base coréenne en ligne : e-koreanstudies.com.

2. Préparation de la base scientifique chinoise cqvip.com

2.1. Principe général de la méthode

Cette base, accessible par l'URL http://www.cqvip.com, consigne une grande partie de la production scientifique chinoise. Citons pour exemple les 3 000 notices bibliographiques récupérées par mes étudiants de M1 de l'IUP SID, sur le domaine de l'aquaculture, pour la période 2004-2007. En fait, sans limite de date, plus de 8500 notices étaient disponibles sur cette base. Le codage des informations en UNICODE, nous permet de travailler sur des séquences d'idéogrammes qui correspondent soit à des mots-clés, soit à des acteurs du domaine (journaux, colloques, organismes, laboratoires, auteurs, …). Les passages en texte libre peuvent à leur tour être traités (Titre et Résumé) afin d'y détecter de nouvelles séquences souvent inconnues des experts du domaine. Leur traitement par agrégation dans des réseaux sémantiques permet d'en déduire la consistance (nombre de documents y faisant référence), l'actualité (apparition récente), la pertinence (domaine connexe et qualité des acteurs), la structure sémantique (nombre de segments émergents et typologie de leurs relations : verbes). On peut ainsi, sans parler la langue, extraire tous les éléments statistiques utiles à la prise de décision et même détecter les signaux faibles, qui représentent l'essentiel de la demande en terme d'information à haute valeur stratégique.

Dès la page d'accueil du site cqvip.com, un moteur de recherche est disponible, il doit être alimenté par des requêtes en chinois qui peuvent être traduites depuis l'anglais par les outils linguistiques de *Google*. Il suffit ensuite de cliquer sur la loupe pour obtenir les premières notices (certains champs sont masqués). Plusieurs options sont alors possibles : soit récupérer les notices visualisées par un copier-coller vers *MS-Word*, soit demander un téléchargement de l'ensemble des champs, soit encore demander à un robot de faire le travail.
Voici le protocole de travail que nous préconisons pour réaliser des analyses stratégiques depuis cette base documentaire.

- établir une équation de recherche en anglais et la traduire en chinois [*Google*]. Par exemple : ''aquaculture'' devient ''水产养殖''
- soumettre cette équation au moteur de recherche de la base cqvip.com,
- pour chaque page de résultat, la copier-coller dans *MS-Word 2003* ou dans *Open Office* [1],
- enregistrer le fichier obtenu au format html,
- l'ouvrir avec *Internet Explorer,*
- afficher le code source de la page,
- l'UNICODE apparaît en mode texte ASCII sous la forme « 〹 »,
- réenregistrer le code source en format texte.

Ce dernier fichier, après un reformatage par un programme *Perl* (détection des balises et des idéogrammes codés UNICODE voir figure 2), peut être entièrement analysé par notre plate-forme « *Tétralogie* » :

- création des dictionnaires (auteurs, mots-clés, journaux, organismes, villes, dates et chaînes répétées d'idéogrammes issues du texte libre),
- croisement des variables (tableaux 2D ou 3D binaires, de contingences ou de cooccurrences simples ou multiples),
- visualisation des résultats (tris de matrices, cartes factorielles, arbres de classification, dessins de graphes, cartes géostratégiques),
- détection des signaux faibles, …

Une fois l'analyse stratégique effectuée, il est possible d'exporter les croisements dans une base de données *MySql* qui peut être interrogée via le Web [Ghalamallah 2008]. Si tous les segments de texte (nom des auteurs, mots-clés, journaux, …) sont traduits du chinois à l'anglais (par *Google* ou *Systran*), l'utilisateur distant travaillera de façon transparente dans cette langue. Il lui sera alors possible de se focaliser, par filtrage multi critères, sur un point de détail qui l'intéresse particulièrement et d'obtenir toutes les statistiques possibles sur cet extrait des données sources (tops n, histogrammes, secteurs, réseaux d'acteurs, réseaux sémantique, évolutions, cartes géographiques, classifications, …).

[1] On se sert, ici, des capacités de transcodage de ces outils pour passer de polices chinoises diverses à l'UNICODE. *MS-WORD 2007* n'est pas recommandé car il sait travailler avec plus de polices, il ne fait donc plus le transcodage voulu : « qui peut le plus peut le moins ».

Figure 1 *: le site cqvip.com (en haut le moteur de recherche, à droite un extrait du résultat)*

2.2. Reformatage de la base

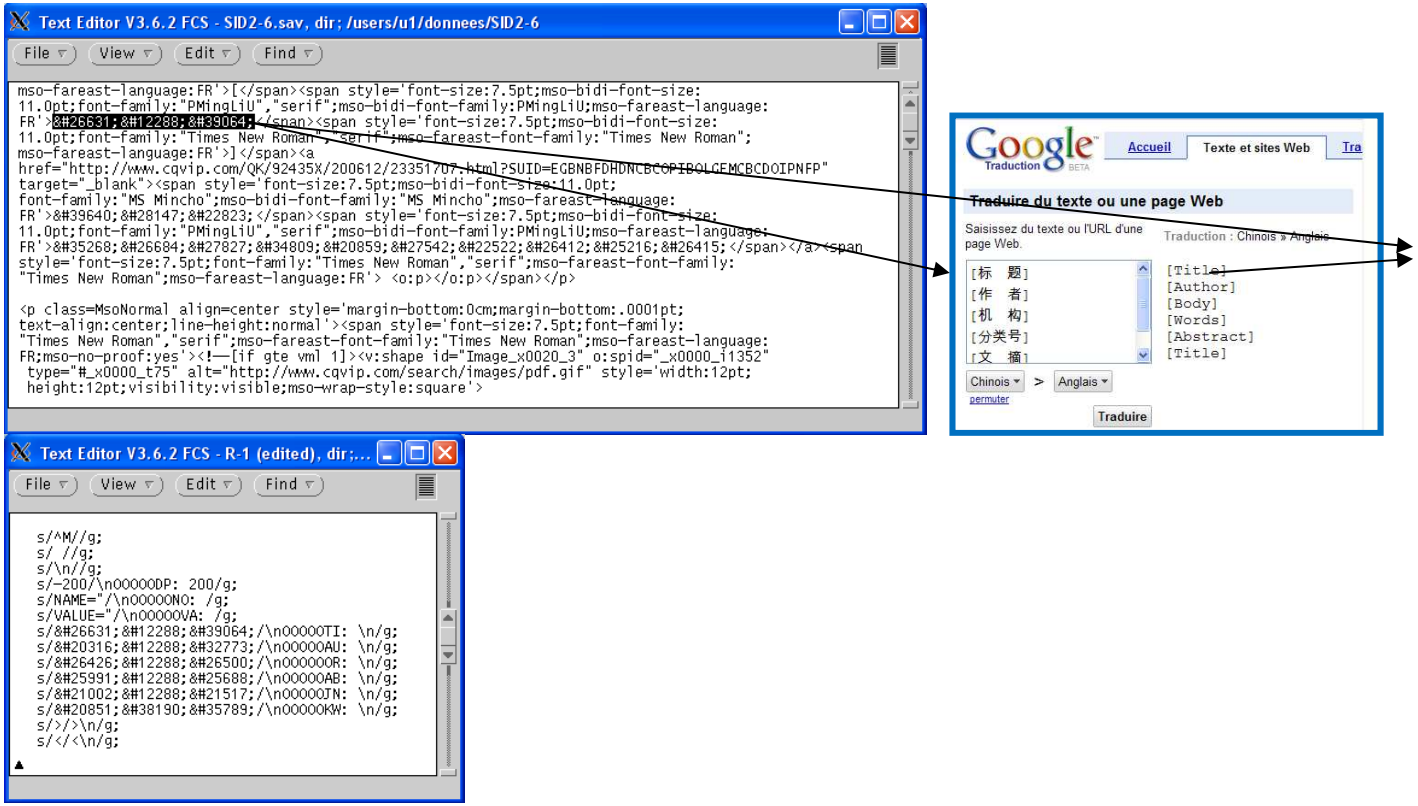


Figure 2 *: Reformatage de la base CQVIP : traduction par Google et transformation des balises par un utilitaire Perl*

Le but de cette étape est multiple :

- éliminer la mise en forme du texte (html) qui n'apporte rien au contenu, mais qui représente 90% du poids du fichier,
- reconstituer des chaînes de texte qui sont séparées en tronçons par la mise en forme. Cette étape est nécessaire car de nombreux mots clés, par exemple, sont coupés (changement de police au niveau html). Est-ce volontaire pour géner l'analyse automatique et l'indexation ?
- rebaliser le texte par des balises ASCII (du type : TI:, AU:, OR:, …) traduites des balises existantes en chinois,
- certaines balises et leur contenu ne sont pas visibles sur la page web, mais sont présentes dans le code source en ASCII (ex : KW:)
- rajouter des balises inexistantes dans le texte en les créant à partir de la mise en forme html : DP:, NO:, VA:,

- conserver les informations qui sont codées en caractères latins ou en chiffres arabes comme les dates, les numéros ou certains termes comme les auteurs occidentaux, les formules ou les éléments techniques,
- éviter les doublons dus à la présence simultanée de deux versions du même terme (ex : noms des auteurs cliquables sur la page web).

Une fois le reformatage/rebalisage terminé, le corpus se trouve sous une forme analysable par « *Tétralogie* » au même titre qu'une base initialement codée en ASCII comme *PASCAL, FRANCIS, INSPEC* ou *SCI*. Des méta données compatibles avec ce balisage vont permettre ensuite de piloter les automates qui vont traiter le texte : extraction des items de tous les champs balisés, puis croisement des champs entre eux afin d'établir des tableaux de cooccurrences à 2 ou 3 dimensions.

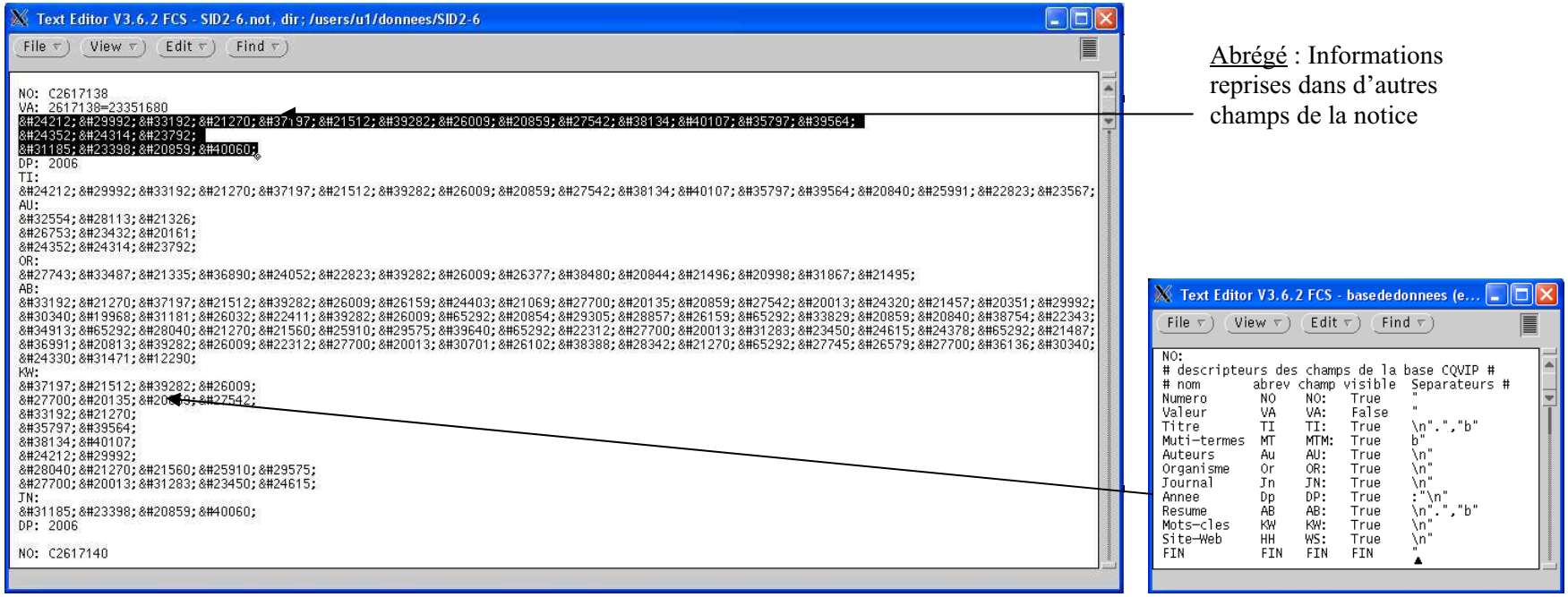
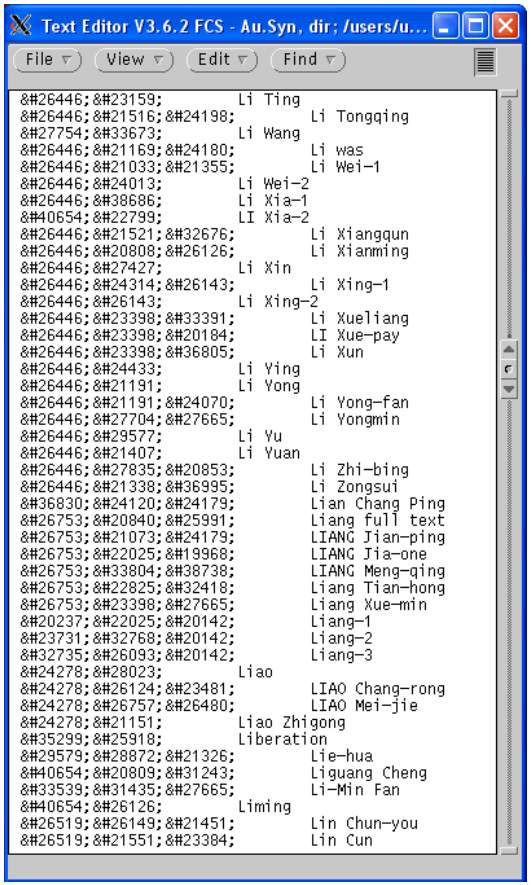


Figure 3 *: Une notice bibliographique reformatée (balises en ASCII, texte chinois en UNICODE) et ses méta données*

Dans la figure 3, nous pouvons remarquer les balises de chaque champ en ASCII et le texte (contenu) en UNICODE. Dans la notice C2617138, le titre, le premier auteur, le journal et la date forment un abrégé en début de document, ces informations sont ensuite reprises dans les champs correspondants : TI:, AU:, JN:, DP:. Le champ VA: ne sera donc pas utilisé dans l'analyse.

Si nous analysons visuellement le document, il comporte 3 auteurs (de 3 idéogrammes chinois = 3 codes), un seul organisme, 8 mots clés (ici de 2 à 5 idéogrammes) plus le journal et la date : 2006. Nous allons voir par la suite que le titre et l'abstract feront l'objet d'un traitement sémantique particulier : détection des chaînes répétées d'idéogrammes ne correspondant à aucun mot clé, d'où possibilité d'une terminologie qui n'apparaît pas dans l'indexation proposée.
Les méta données (à gauche) décrivent le format obtenu : nom complet de chaque champ, abréviation de ce même champ (sert à nommer les fichiers et répertoires associés à son analyse), identifiant exact du champ dans la notice (ex : TI: pour le champ Titre), TRUE pour l'utiliser dans l'analyse, suite des séparateurs utilisés pour découper le texte (chaîne de caractère, ''b '' pour l'espace, ''\n'' pour la fin de ligne, ''ORD2'' pour le 2° segment, …). Dès que nous disposons du corpus reformaté et de ses méta données, il est possible de l'analyser par les mêmes techniques que pour une source en alphabet latin, y compris réaliser son traitement sémantique et proposer une indexation plus fine et récente.

2.3. Problèmes de traduction



Afin de décoder l'UNICODE (et donc, ici, le chinois), nous avons établi des dictionnaires de correspondances entre les noms d'auteurs en chinois et leur traduction en phonétique (Pinyin) grace au traducteur de *Google*. Mais dans ce cas, se posent deux problèmes :

- *Google* n'arrive pas à traduire tous les noms et restitue alors l'UNICODE (voir 7° auteur)
- Plusieurs auteurs de codes différents peuvent avoir la même correspondance, d'où une ambiguïté très néfaste à l'établissement de réseaux sociaux pertinents et la nécessité impérative de corriger ce problème.

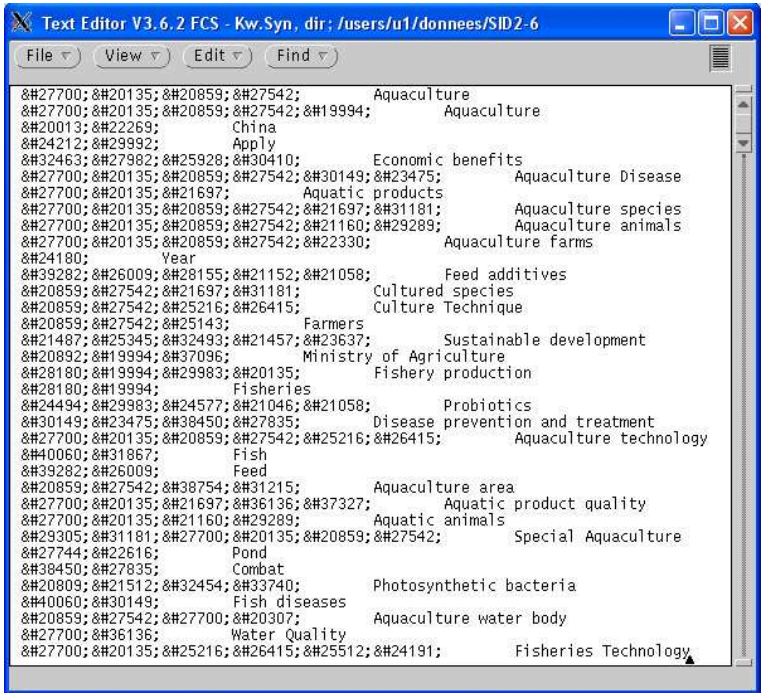
Nous avons choisi, dans le premier cas, de garder les codes et, dans le second, d'ajouter un différenciateur numérique aux noms traduits en totalité ou partiellement (ici, LI-3 et XUE GANG-1, par exemple).
La lecture des dictionnaires, des cartes, des graphes, des dendogrammes est ainsi plus parlante, tout en évitant d'introduire des artéfacts dus à la traduction approximative souvent proposée par *Google*. Dans ces conditions, il est possible de faire la traduction avant l'analyse, sinon il faut tout analyser en chinois (donc en UNICODE, c'est-à-dire en aveugle) et ne traduire qu'à la fin, afin de restituer un résultats fiable et intelligible (traduit) pour les utilisateurs. Deux noms identiques peuvent alors cohabiter dans la même visualisation du moment qu'ils ont été différenciés par leurs codes respectifs tout au long de l'analyse. Les deux Li-Xing (-1 et -2) seront alors alors différenciés non par leur nom (homonymie complète après traduction) mais par leur position dans les cartes ou leur appartenance à des classes différentes. On retrouve le problème bien connu de

l'ambiguïté des noms asiatiques, mais ,ici, il est résolu presqu'entièrement à la différence des bases occidentales où il perdure.

Figure 4 *: Correspondance UNICODE-Pinyin et occurrences*

Un autre problème de traduction peut se poser pour la terminologie technique (mots clés, indexation additionnelle, texte intégral) car *Google* ou *Systran* ont des difficultés sur ce point là. Absence dans leurs dictionnaires (termes trop techniques ou trop récents), problème de contexte, complexité des phrases, ambiguïté du discours. La plupart du temps, cette incertitude est levée, car le nombre de termes bien traduits, dans par exemple un cluster sémantique, est suffisant pour comprendre la nature du concept mis en valeur et sa pertinence. Bien entendu le recours à un expert connaissant bien la langue est recommandé.

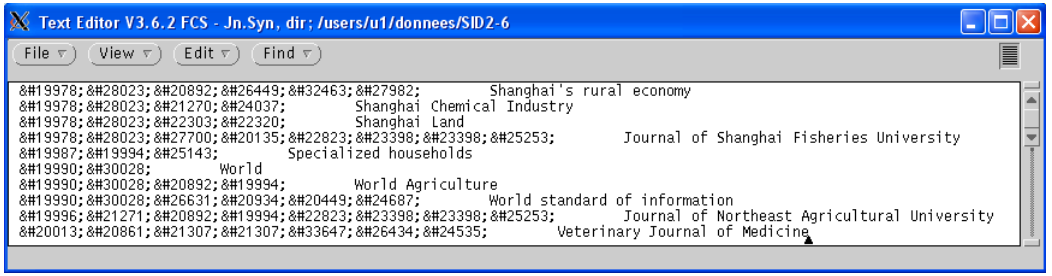
- Pour les mots-clés, le problème est similaire. En effet, certains mots-clés, qui en UNICODE sont différents, se voient attribuer la même traduction par *Google*. Ce phénomène est heureusement assez rare, ce qui ne compromet pas trop l'interprétation des experts.



Ci-contre, le début du dictionnaire de synonymes donnant la correspondance entre mots clés chinois en UNICODE et leur traduction *Google* en anglais. Les occurrences des mots clés sont ensuite calculées pour l'anglais, d'où fusion des termes traduits de la même façon. Ici, dans notre exemple, le terme générique le plus fréquent « aquaculture » cumule les occurrences de plusieurs formes, c'est beaucoup moins génant que dans le cas des auteurs, mais on risque de perdre certaines nuances ou faire des approximations peu judicieuses. A souligner que dans l'affichage html des documents dans *Internet Explorer* (figure 1), le champ mot-clé n'apparaît pas, il est simplement masqué par la mise en forme, mais bien présent dans le code source et donc restitué à l'issue du reformatage (figure 3). Ces mots clés seront très utiles dans la phase de traitement sémantique du titre et du résumé, car ils imposeront des chaînes d'idéogrammes ayant un sens dans le contexte étudié.

Figure 5 *: traduction des mots clés et leurs occurrences calculées après traduction.*

- Pour les journaux, la traduction est sans souci bien que parfois un peu folklorique.



- Par contre, pour les organismes, plusieurs formes peuvent être rencontrées (problème inhérant à la morphologie des adresses). D'où la nécessité de mettre au point un dictionnaire de synonymes permettant de corriger et donc de regrouper toutes les formes équivalentes rencontrées.

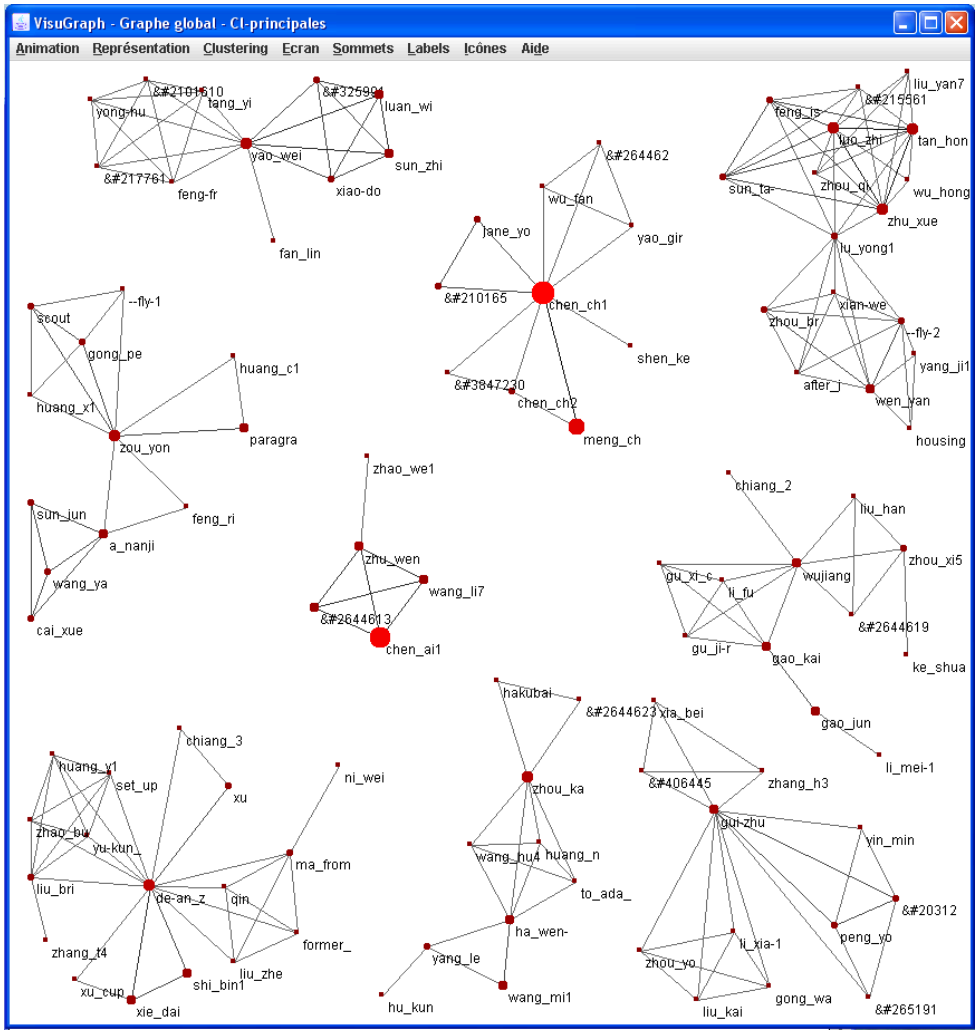
Figure 6 *: traduction des journaux.*

3. Analyse de l'aquaculture en Chine

3.1. Les réseaux d'acteurs

Après avoir traduit les noms des auteurs en anglais et avoir différencié ceux dont les noms sont différents en chinois et identiques en anglais, nous avons croisé les auteurs ayant au moins deux publications (connecteurs potentiels). La figure 7 présente la topologie des principales équipes. Première constatation : très peu de collaborations affichées dans les publications scientifiques chinoises. Seconde constatation, les équipes sont le plus souvent dirigées par un mandarin qui contrôle 2, 3 ou 4 sous équipes distinctes. Certains noms ne sont pas traduits, d'autres ont des traductions folkloriques, car leur nom a une autre signification en chinois :

- 古群红 Ancient group of red
- 金彩杏 Apricot Jincai
- 吴早保 As early as Paul Wu
- 孟和平 Bangladesh peace
- 蓝正升 Blue is up
- 商德章 Business ethics chapter
- 商万成 Business Wancheng
- 蔡秀丽 Cai beautiful
- 蔡建堤 Cai embankment
- 陈国兔 Chan Kwok-rabbit
- 章秋虎 Chapter autumn tiger
- 陈权军 Chen the right to military
- 邓正营 Deng Zhenglai business
- 瘐莉萍 Die in a prison Liping
- 别文群 Do not text-qun



- 董在杰 Dong in the kit
- …

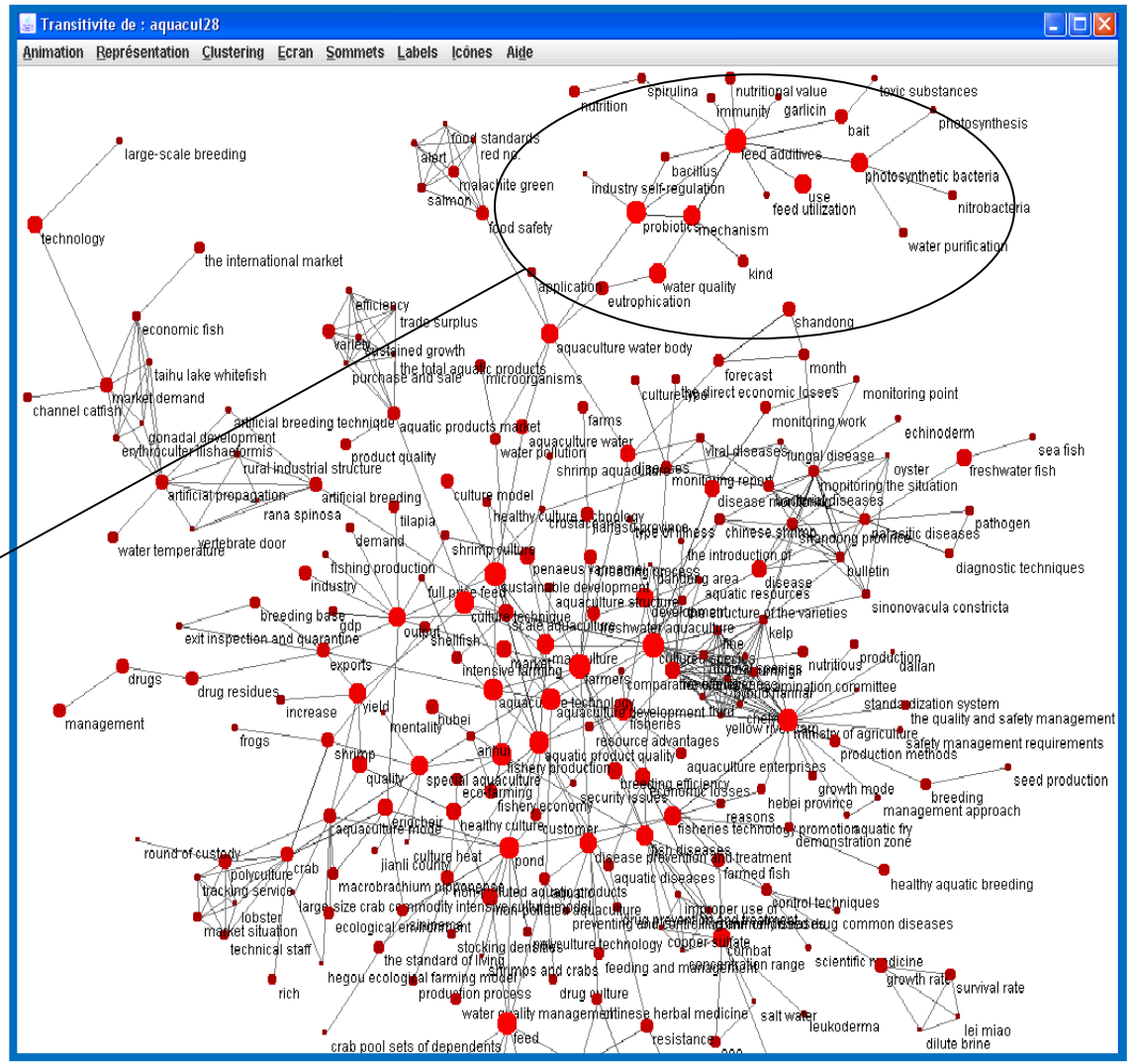
3.2. Les réseaux sémantiques

Dans le même esprit, il est possible de croiser les mots clés proposés dans les notices et ainsi faire apparaître une organisation de la « terminologie contrôlée » en fonction des différents domaines abordés en aquaculture. Nous ne pouvons pas, ici, détecter de signaux faibles, car le fait d'appartenir à un thésaurus de mots clés ne plaide pas en faveur de l'emergence d'un concept. Par contre, les signaux forts et la diversité des domaines représentés doivent ressortir de ce type de structuration d'éléments à forte sémantique.

Voici l'exemple d'un des clusters trouvés :

- Feed additives
- Nutrition
- Spirulina
- Nutritional value
- Immunity
- Garlicin
- Bait
- Toxic substances
- Photosynthetic bacteria
- Photosynthesis
- Nitrobacteria
- Water purification
- Feed utilization
- Bacilius
- Probiotic
- Industry self-regulation
- Mechanism

Figure 7 *: réseaux des auteurs des principales équipes de recherche.*



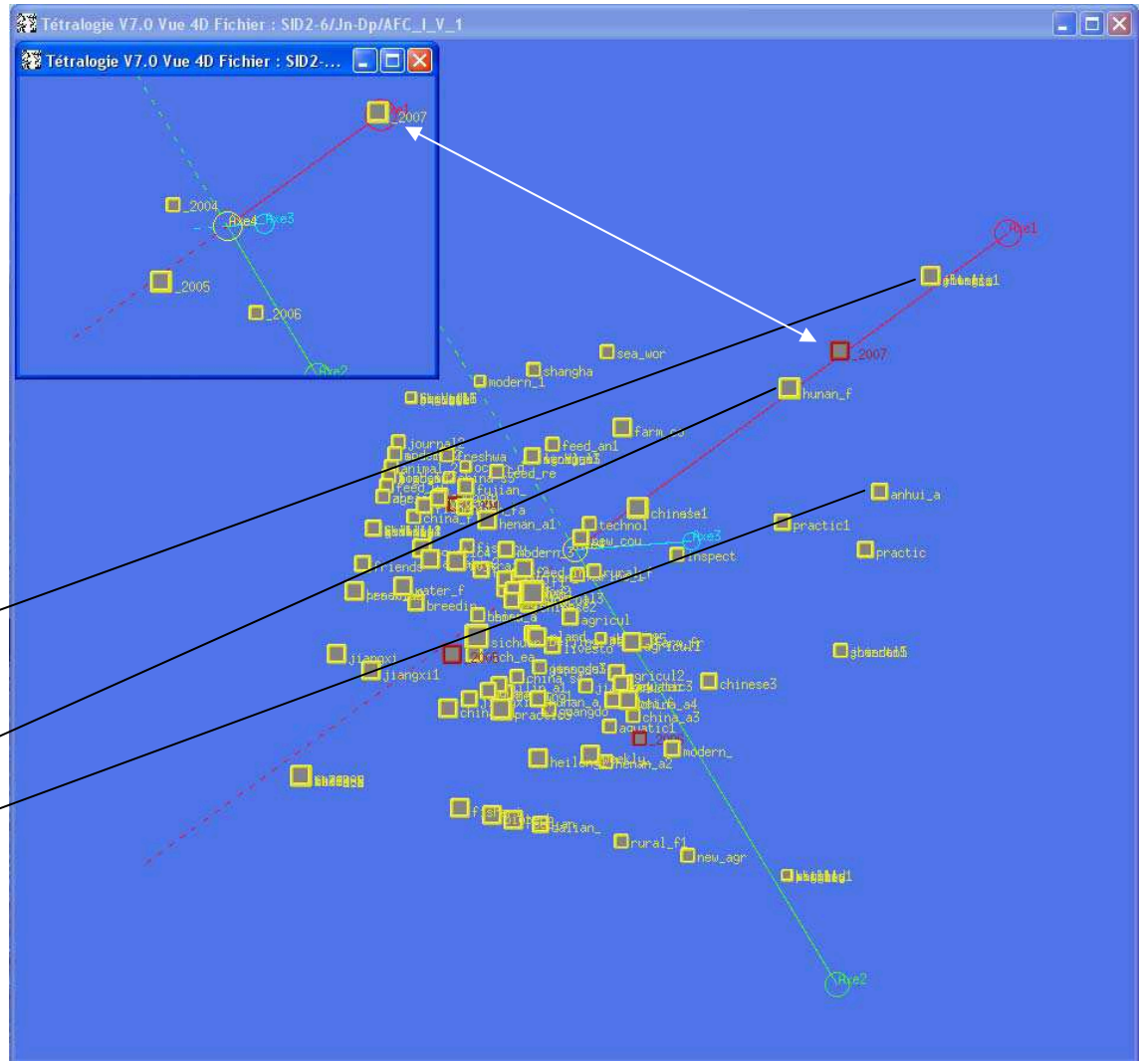
- Kind
- Water quality …

Figure 8 *: réseau sémantique des mots clés proposés dans CQVIP.*

3.3. Analyse de l'évolution

Pour illustrer une première méthode, nous avons choisi d'analyser le profil d'évolution des journaux sur les quatre années de notre étude, à savoir [2004 ; 2007]. Une analyse factorielle des correspondances sur la matrice de cooccurrences Journaux X Dates (Jn X Dp) nous permet de visualiser facilement (figure 9) les différents profils sur un tétraèdre régulier représenté, ici, en 3 D. Au niveau de chaque sommet les journaux qui ne sont présents qu'un an : les nouveaux journaux sont donc en haut à droite. Sur une arête ceux qui ne sont présents que sur 2 années : ceux de 2006 et 2007 sont sur l'arête de droite. Ceux présents 3 ans sont sur une face, ceux présents tout le temps sont à l'intérieur du tétraèdre et attirés par l'année où ils sont le plus présents. La carte en haut à gauche ne visualise que les années, l'autre y ajoute les journaux. Ces deux cartes peuvent être orientées dans tous les sens, l'azimut de visualisation choisi peut être exporté sur l'autre carte ou vers toute carte visualisée sur l'ordinateur d'un utilisateur distant. On peut ainsi lui communiquer un point de vue significatif comme, ici, celui servant à identifier les nouveaux journaux abordant l'aquaculture.

- Chemical friends
- China education information
- Jiangsu health service management
- China electrical industry
- Intelligent buildings and urban information
- Electric locomotive and mass transit vehicles
- China metallurgical
- Anhui agricultural science
- Hunan feed
- Shanxi agricultural: the village committee director



- China inspection and quarantine

Figure 9 *: analyse factorielle sur la matrice Journaux X Dates*

Une seconde méthode consiste à réaliser un tableau de cooccurrences 3D qui croise, par exemple, les auteurs entre eux et le temps. Nous pouvons ensuite visualiser le réseau évolutif de ces auteurs sur un graphe adapté. Les périodes y sont distribuées chronologiquement sur une couronne comme les heures sur une horloge. Les auteurs sont attirés par ces sommets fictifs et viennent se placer vers le centre du graphe s'ils sont présents tout le temps, vers le repère correspondant s'ils ne sont présents qu'une fois, ou dans une position intermédiaire s'ils sont présents au cours de plusieurs périodes consécutives ou non. Cette analogie espace-temps fonctionne un peu comme la carte factorielle précédente, mais on peut mixer cette approche avec le dessin classique de graphe (sans repère additionnel). Nous obtenons alors, comme ci-contre, un graphe qui fait apparaître les principales équipes avec leurs évolutions respectives. L'histogramme coloré attaché à chaque sommet nous indique son évolution quantitative. La position vis-à-vis de ses collaborateurs indique l'époque de son implication dans l'équipe. Ses liens précisent avec qui et combien de fois il a collaboré. Dans la figure 10, nous avons rassemblé les dynamiques des principales équipes chinoises dans le domaine de l'aquaculture. Certaine sont persistantes, d'autres émergentes, d'autres ont cessé momentanément ou définitivement de publier dans le domaine. Il est facile de repérer les leaders (taille des icônes), de détecter les nouveaux auteurs (en vert), les auteurs qui ne sont plus présents (pas de vert), les auteurs clés, connecteurs entre sous équipes. Par exemple, dans l'équipe du haut, seul le leader (Chen Changfu) reste encore actif, il avait un gros collaborateur jusqu'en 2006 (Meng Chang-Ming), il a dirigé deux équipes distinctes en 2004, il a travaillé avec Shen Ke-Ray en 2005 et une équipe de 2 personnes en 2006. Par contre, les trois équipes de gauche ont de nombreux auteurs émergents et des leaders persistants. Les autres équipes ont disparu, les quatre de droite dès 2006. Enfin, il est possible de réaliser une analyse factorielle multiple (AFM) sur ce même tableau 3D (Au-Au-Dp) et faire aparaître les trajectoires des auteurs s'ils changent d'équipe. Ce n'est pas le cas ici, donc pas de mobilité constatée sur les quatre ans de notre étude.

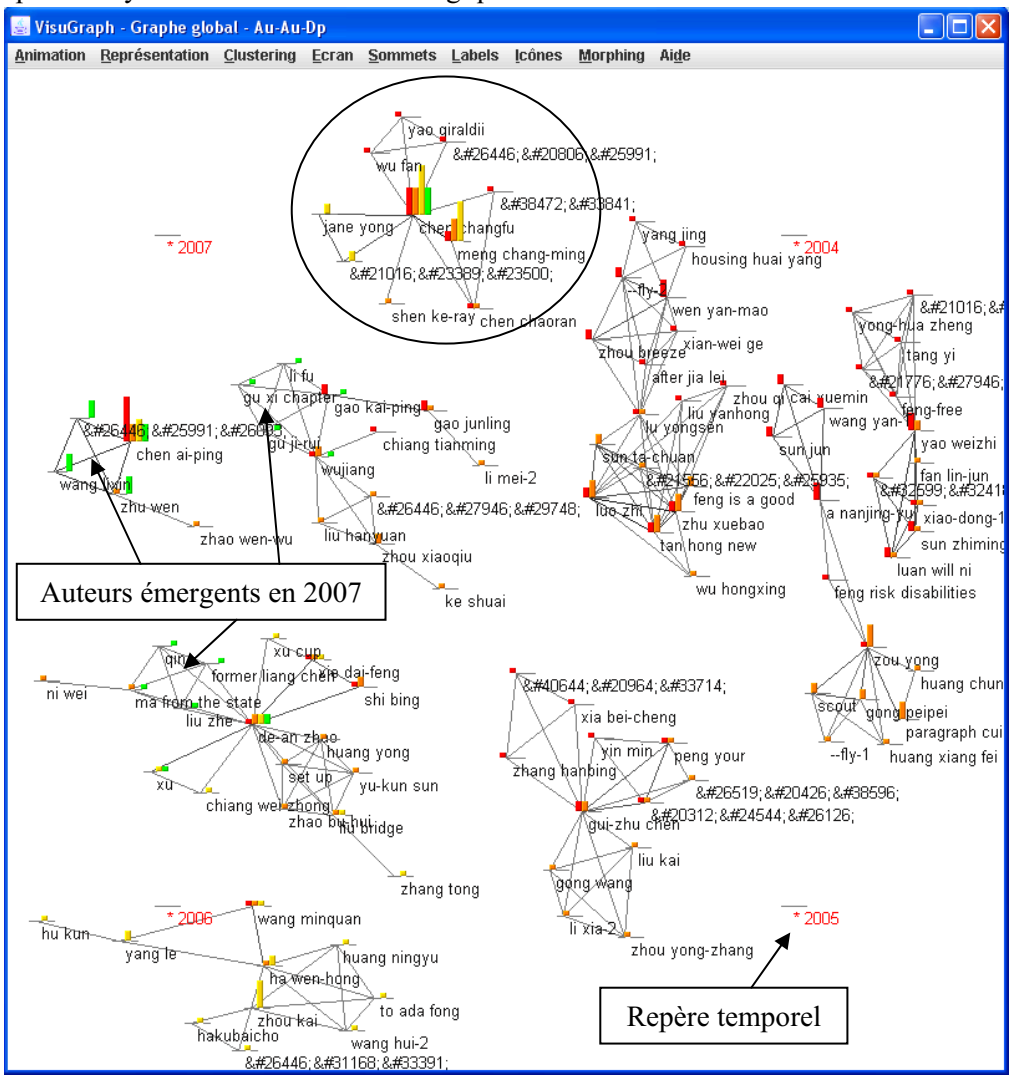


Figure 10 *: réseau montrant l'évolution des principales équipes.*

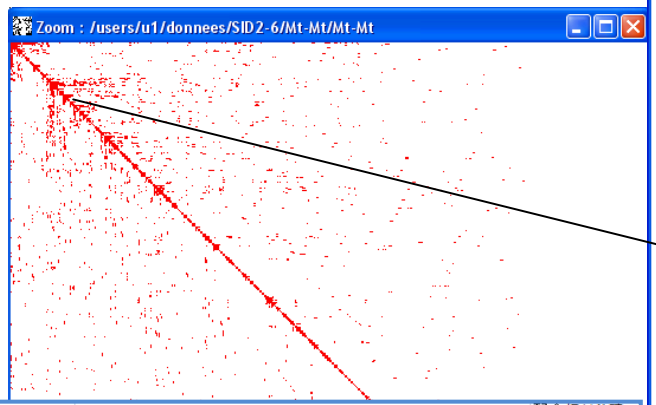
3.4. Analyse sémantique du texte libre

Pour analyser le texte libre, nous utilisons le dictionnaire de mots-clés de la figure 5 (termes techniques, mots-clés), plus un dictionnaire de mots vides (les mots outils de la langue : of, the en anglais), et éventuellement un dictionnaire de synonymes (équivalences terme à terme ou terme avec concept générique : car, vehicle). Le texte libre est, dans un premier temps, découpé en phrases grâce aux ponctuations présentes dans le titre et surtout dans le résumé. Les chaînes d'idéogrammes correspondant aux mots-clés sont ensuite recherchées dans le texte et elles sont complétées par la détection de nouvelles chaînes donc le nombre d'occurrences dépasse un seuil defini par l'utilisateur en fonction de la taille du corpus. Ces nouvelles chaînes d'idéogrammes qui intègrent éventuellement des mots-clés existants sont traduites afin d'essayer d'en comprendre le sens. Il y a alors deux possibilités, ou *Google* en donne une traduction acceptable (mot-clé manquant dans l'indexation de la base, ici CQVIP), ou la traduction ne veut rien dire (nouveau concept absent du dictionnaire de *Google* et donc certainement émergent). Dans ce second cas, il faut revenir au contexte d'utilisation de cette nouvelle chaîne d'idéogrammes et recourir à un spécialiste du domaine, ici l'aquaculture, et parlant le chinois. Une autre possibilité, est de voir si ces nouveaux segments forment des clusters [ROUX 2009] (cooccurrences de plusieurs d'entre eux dans quelques documents). Il est alors fort possible que nous ayons détecté un signal faible (cohérence, simultanéité, concensus). Pour le valider, il suffit de le croiser avec tous les autres champs de la base (Auteurs, Organismes, Mots-clés, Journaux, Dates). S'il est récent, dans un bon journal, dans une bonne équipe et connexe à des domaines connus (mots-clés), il a certainement du potentiel (ce qui reste à valider auprès d'un expert en lui donnant à analyser tous les documents contribuant à ce signal). Nous pouvons ainsi, sans connaître la langue, détecter une information implicite présente dans le corpus et inaccessible par la simple lecture. Ce travail de détection des signaux faibles est en fait très demandé par les décideurs car il correspond à un besoin naturel de détecter précocément l'innovation afin d'évaluer la pertinence de nouvelles pistes et de réagir en conséquence. Ci-dessous, une liste de termes pertinents détectés (nouvelles chaînes d'idéogrammes) et un cluster sémantique émergent.

养殖塘	Breeding pond	养殖塘
养殖可持续发展	Sustainable development of aquaculture	养殖可持续发展
养殖持续健康发展	Sustained and healthy development of aquaculture	养殖持续健康发展
养殖河蟹	Breeding crab	养殖河蟹
养殖船	Culture vessel	养殖船
养殖良种	Breeding improved varieties	养殖良种
养殖大菱鲆	Cultured turbot	养殖大菱鲆
养殖农户	Aquaculture farmers	养殖农户
养殖病原体	Breeding of pathogens	养殖病原体
养殖工作座谈会	Work culture forum	养殖工作座谈会
养殖息	Farming income	养殖息
养殖高产高效	Breeding high yield and high efficiency	养殖高产高效
养殖经济效益	Economic benefits of aquaculture	养殖经济效益
养殖罗非鱼	Tilapia culture	养殖罗非鱼
养殖螃蟹	Breeding crabs	养殖螃蟹
大水产养殖户	Large aquaculture households	大水产养殖户
水产品消费	Consumption of aquatic products	水产品消费

3.5. Détection des signaux faibles

Pour détecter les signaux faibles, nous recherchons tout d'abord les mots-clés et les termes techniques connus et nous les codons dans le texte (Titre et Résumé). Puis, nous détectons les nouvelles séquences dépassant un seuil d'occurrence fixé par l'utilisateur. Nous croisons ensuite ces nouvelles séquences (les multi-termes) avec le temps et nous ne gardons que celles qui émergent fortement dans la dernière période (ici 2007). Nous croisons enfin ces termes entre eux et nous trions la matrice obtenue par blocs diagonaux. Chaque bloc représente un concept émergent identifié par une terminologie nouvelle non présente dans les thésaurus du domaine et issue seulement de quelques documents récents. Ce sont les fameux « signaux faibles » que nous devons ensuite valider en les croisant avec tous les autres champs de la base et notamment les mots-clés. Dans la figure 11, nous remarquons, le long de la diagonale de la matrice, un certain nombre de clusters constitués de nouveaux termes n'appartenant pas aux mots-clés et qui présentent une organisation sémantique très marquée. Chacun d'eux est alors extrait dans une sous-matrice carrée et visualisé sous forme de graphe sémantique (figure 12). Cette information est alors soumise à l'expert du domaine pour validation.



配合饲料养殖	Aquaculture feed	配合饲料养殖
养殖塘	Breeding pond	养殖塘
养殖品	Aquaculture products	养殖品
养殖自身污染	Breeding self-pollution	养殖自身污染
养殖可持续发&#23	Sustainable development of aquaculture	养殖可持续发展
养殖鲍	Abalone aquaculture	养殖鲍
养殖异	Culture differences	养殖异
养殖情	Culture conditions	养殖情
养殖持续健康&#21	Sustained and healthy development of aquacu	养殖持续健康发
养殖质量安全	The quality and safety culture	养殖质量安全
养殖造	Culture building	养殖造
养殖河蟹	Breeding crab	养殖河蟹
养殖营养	Aquaculture Nutrition	养殖营养

Figure 11 *: traduction Google de segments répétés et clusters émergents*

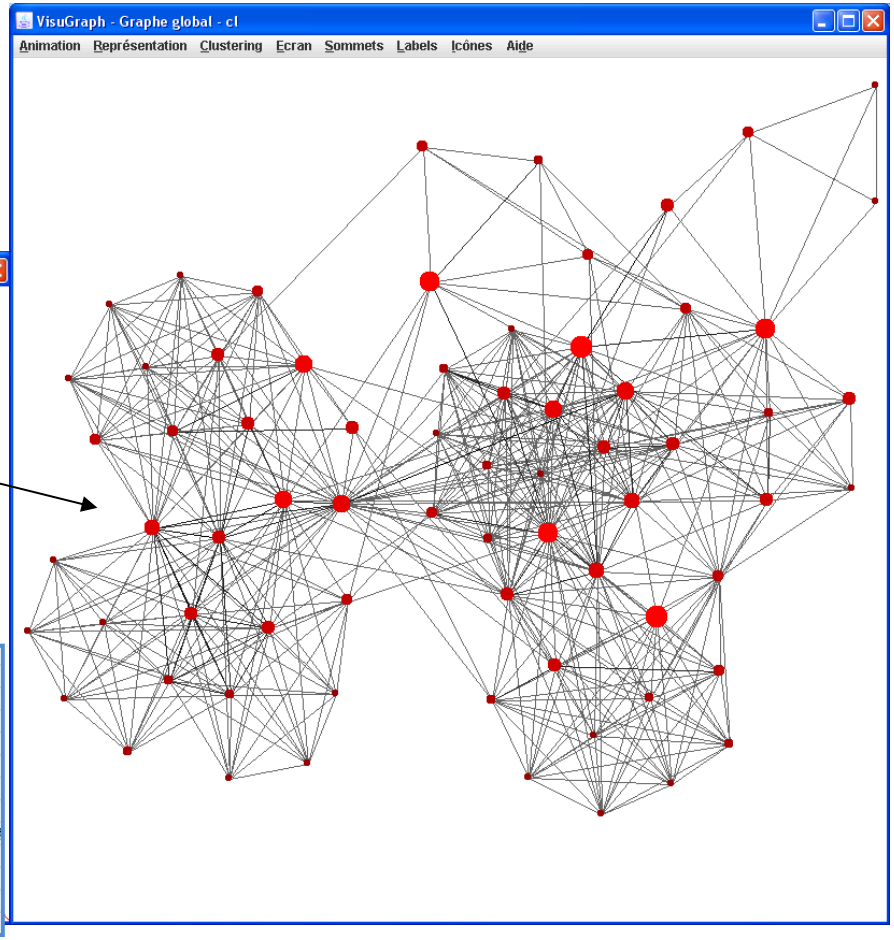


Figure 12 *: exemple d'un réseau sémantique de multi-termes*

4. Autres possibilités d'analyse

Voici deux autres exemples de sources dont l'analyse peut être menée par une méthode analogue à celle suivie pour le chinois. L'UNICODE UTF-8 peut être extrait depuis le code source des pages html. Pour Al Jazeera, l'originalité est de pouvoir analyser les réactions des internautes aux articles via le blog. Pour la base en coréen, nous voyons que la plage des caractères réservée à cette langue est différente, mais le principe d'analyse reste le même. A chaque fois, la difficulté est de trouver un balisage suffisant permettant de catégoriser au mieux les informations avant analyse (acteurs, sémantique, dates, …). Des dictionnaires de mots-clés et d'expressions sont aussi très utiles pour traiter le texte libre et y détecter l'innovation.

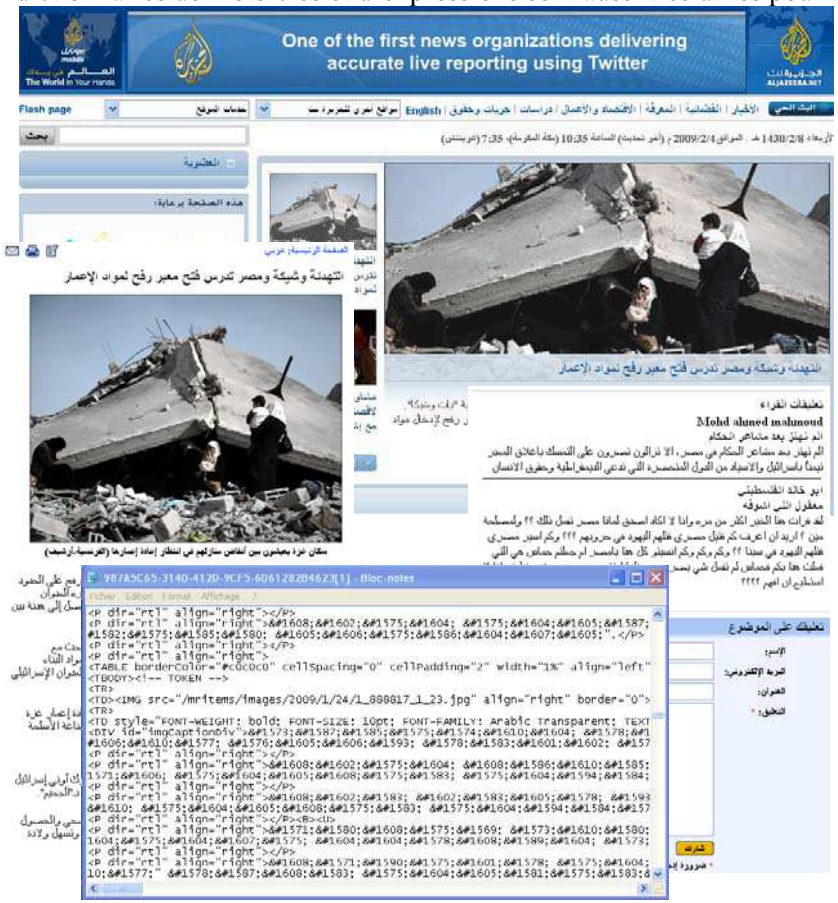


Figure 13 *: le site d'Aljazeera.net (article bref et blog correspondant)*

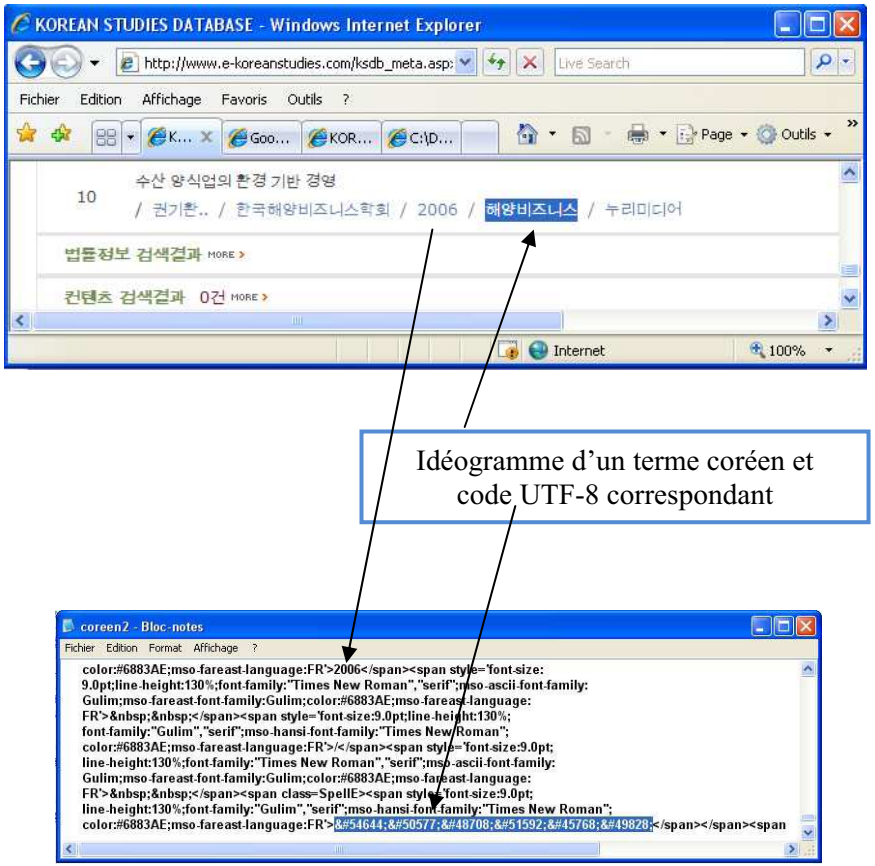


Figure 14 *: cas du site www.e-koreanstudies.com*

5. Conclusion

La base documentaire CQVIP, sur laquelle nous avons réalisé nos tests, ne représente qu'un exemple des multiples sources qu'il est possible de traiter par cette méthode, aussi bien pour le chinois que pour d'autres langues comme le japonais, le coréen, l'arabe. Il reste cependant de nombreux problèmes à résoudre, pour que cette démarche devienne parfaitement opérationnelle.

- Etablir des dictionnaires thématiques et leur traduction en anglais (et/ou en français).
- Traiter les entités nommées (pour les auteurs ou certains organismes ou journaux la traduction automatique est suffisante), mais il reste de nombreuses ambiguïtés à lever (influence des accents, de la prononciation, du contexte).
- La traduction de la nouvelle terminologie détectée de façon statistique ne sera pas nécessairement présente ni dans les dictionnaires thématiques car trop récente ou tout simplement inconnue des non initiés, ni dans la traduction proposée par *Google* pour les mêmes raisons.

Il devient donc nécessaire à terme d'envisager une collaboration entre experts :

- de la fouille de texte
- du traitement automatique des langues (sémantique, morphosyntaxique, ontologies, …)
- des langues (chinois, coréen, japonais, arabe, …)
- des domaines à analyser (scientifique, technologique, économique, géopolitique, …)

Cette collaboration peut intervenir à deux niveaux : la préparation des données (homogénéisation du vocabulaire, choix des granularités, traductions, désambiguïsation, …) et l'interprétation des résultats, car bien souvent nous sommes obligés de revenir aux documents sources qui comportent des passages assez importants en texte libre, qu'il vaut mieux appréhender dans la langue d'origine.

6. Bibliographie

[1] **DOUSSET B.,** *Extraction of strategic information through analysis of major components* . Datametrics Journal, volume 2, april 2008, issue 1.

[2] **GHALAMALLAH I., GRIMEH A., DOUSSET B.,** *Processing data stream by relational analysis*. MODULAD, INRIA, n°36, july 2007.

[3] **GHALAMALLAH I., LOUBIER E., DOUSSET B.,** *Business intelligence a proposal for a tool dedicated to the analysis relational*. International Journal of Competitive Intelligence, Strategic, Scientific and Technology Watch, SciWatch Journal, hexalog, Barcelona - Spain, Vol. 3, august 2008.

[4] **GUENEC N., LOUBIER E., GHALAMALLAH I., DOUSSET B.,** *Management and analysis of chinese database extracted knowledge*. Flexible Query Answering (FQAS 2008), Londres , september 22, 2008.

[5] **LOUBIER E., DOUSSET B.,** *Visualisation and analysis of relational data by considering temporal dimension*. International Conference on Enterprise Information Systems (ICEIS 2007), Funchal, Madeira - Portugal, Vol. ISAS, INSTICC Press, p. 550-553, june 12-16 2007.

[6] **ROUX C.,** *Methodes of sort for extracting weak signals*. International Journal of Competitive Intelligence, Strategic, Scientific and Technology Watch, SciWatch Journal, hexalog, Barcelona - Spain, Vol. 2, april 2009, issue 1, pp 23-29.