KNOWLEDGE DISCOVERY FROM PEOPLE AND SEMANTIC NETWORKS -
ANALYZING TEXTS IN LANGUAGES WE DO NOT UNDERSTAND

Bernard Dousset, Anass Elhaddadi, Josiane Mothe
bernard.dousset@irit.fr, anass.elhaddadi@irit.fr, josiane.mothe@irit.fr
Institut de Recherche en Informatique de Toulouse, IRIT UMR 5505
Université de Toulouse, Université Paul Sabatier,
118, route de Narbonne, 31062 Toulouse cedex 9 (France)

Abstract **- In this paper we present a methodology that makes it possible to mine a document collection from a domain without knowing the language in which the documents are written. This is a challenging problem for enterprises who want to analyze resources they cannot read. We describe in detail a method, tools and results that can be used within a digital library context for science watch and competitive intelligence. We consider a collection associated with the aquaculture domain written in Chinese and extracted from a digital library. Based on the original coding (UNICODE) of the data and the tag marking the structure of the documents, we extract key elements (authors, phrases, etc.) from within the domain and analyse them. The results are displayed in the form of graphs and networks. We extract people networks and semantic networks before examining their evolution over a period of several years. The principles developed in this paper can be applied to any language.**

ACM taxonomy: **H.3.1Content Analysis and Indexing; H.4.2.aDecision support**

Keywords: **text mining, semantic network, social network, knowledge discovery, science watch, competitive intelligence.**

I. INTRODUCTION

In competitive intelligence and science watch activities, it is important to consider the documents that are written by foreigner organisations. This implies to access documents that are written in different languages. Accessing information generally implies that the user understands the language that a document is written in. To counter the problem of reading documents in a language with which the user is not familiar, online translators can be of assistance. Indeed such translations are available, for example, from Google © or Systran ©. However, reading an entire document translated using a machine is not entirely satisfactory:

-Some sentences can be difficult to understand, particularly when the original document is written using long sentences or a language which is rich,

-Some tasks involve reading many documents, particularly in relation to decision tasks or scientific monitoring.

In this paper we consider a related problem, that is, the analysis of a large collection of documents extracted from a digital library where the documents focus on a particular domain. In specific terms, the problem we tackle is the analysis of semantic and people networks from documents written in a foreign language that the user does not understand. These networks are first created by considering the entire set in a homogeneous way; then we suggest a method to analyse partitioned sets - the information is broken down according to the period of time in which it occurs and several periods are fused together so that development of people networking activities can be easily observed.

In order to analyse these documents and extract these networks when the language used in the documents cannot be understood, we set forth a method based on the extraction of n-grams. In the case of Chinese, for example, the analysis is based on n-grams of ideograms which correspond to key elements from within the domain (authors, journals, keywords, etc.). More specifically, we take advantage of the structure of some resources to extract the key elements such as phrases taken from editor keywords and in addition we build dictionaries. These dictionaries are then used to analyse free text, either directly or by cross referencing these reliable elements with other extracted elements using statistically-based automatic methods.

To illustrate our method, we describe the analysis of a document set extracted from the scientific digital library in the Chinese Scientific Journals Database (CQVIP). We also give some clues on how to manage other resources in a similar fashion, such as the Al Jazeera information site in Arabic and an on-line Korean collection, e-koreanstudies.com.

This paper is set out as follows: we first present some related work in section 2, then we present the

method for Chinese. Section 3 presents the raw data and the pre-processing before analysis can take place. The analysis is presented in section 4. Section 5 presents other examples with Arabic and Korean. Section 6 concludes the article.

II. RELATED WORK

Many articles take into account the problem of document access when documents are written in a language that the user is not familiar with or does not use as a primary language. In Cross-Lingual Retrieval for example, users query information corresponding to their information needs using their own language and the system retrieves documents written in a foreign language (Peters, 2009). Many approaches are employed to resolve this problem and query translation is one of them (He et al., 2003) (Chengye et al., 2008). Reading documents that are not written in a language the user is familiar with is an issue. (Li et al., 2003) present an *English reading-assistance system* which suggests translations of words and phrases based on mining techniques. (Fang et al., 2006) promote a method to predict possible English meanings according to each component of a Chinese term.

The second aspect we study in this paper refers to the automatic extraction of people and semantic networks based on the mining of scientific publications. Analysing scientific publications to discover trends and to understand the structure of a scientific field and the evolution of scientific communities or topics has been widely explored in literature, in particular, but not exclusively, in scientometrics (Leydesdorff, 1995). Different types of analysis can be undertaken. In information science, citation and co-citation analysis have been studied in the past as a means of monitoring scientific activities (White and McCain, 1998), (White, 2003). Citation analysis is used to identify core groups of publications, authors and journals. Conversely, co-citation analysis1 is used to detect networks of authors or to map topics and authors or journals (Zitt and Bassecoulard, 1994) (White, 2003). Groupings other than authors can be used for the purposes of correlation when mining scientific publications such as keywords, journals, etc. as presented in (Mothe and Dkaki, 1998). Digital libraries usually deliver results in the form of lists of related elements (lists of related publications or authors) even though it has been shown that graphical interfaces play an important role in displaying the results of analysis to users (Chen, 2002), (Geroimenko, 2002). In this context, graphs or networks are powerful methods of visualisation, mainly because linking concepts or elements together is a very common mining technique. Another reason is that a network is easy to understand, even by a naïve user. In (Mothe et al., 2006) scientific publications are mined in order to highlight groups of authors and their geographic relationships.

This paper extends an earlier work from Dousset (2009). This new version aims at spreading the results to an international audience.

[1] A co-citation can be extracted when two references appear in the same published paper.

III. CHINESE AS A CASE STUDY

A. Raw data

We considered the scientific digital library (DL) http://www.cqvip.com. This DL brings together a large number of Chinese scientific publications (see figure 1). A search engine is available on the main page of the site to retrieve documents in response to a query in Chinese (see figure 2). Since queries can be just a few words, it is easy to write a query in Chinese corresponding to the field of interest by simply taking any dictionary or translator. For example, “aquaculture” in French corresponds to “aquiculture” in English and “水产养殖” in Chinese.

Next we can click on the relevant button to obtain the first references (some of the fields are hidden). Several options are then possible: gather the references as visualised by copy-pasting to an editor such as MS Word ©, download all the fields, or ask an engine to download everything. For example, we managed to select 3,000 references in the aquiculture field from 2004-2007. Since the information is coded in UNICODE format (in the form “〹”) it is possible to extract n-grams or sequences of ideograms which correspond either to keywords or to actors in the field (newspapers, conferences, organisations, laboratories and authors). Free text (title and summary) can also be used in order to detect new sequences of terms that may be unknown to domain experts.

B. Re-encoding the data

There are several goals for this phase:

-To eliminate text formatting and corresponding tags (HTML in our case) which do not bring any content, but which correspond to 90% of the file size.

-To rebuild text strings that are split because of formatting. This is necessary because many terms or phrases are reduced (e.g. because of font change in HTML).

-To tag the texts again using ASCII tags (in our case we use tags in a similar way to many digital libraries: TI for Title, AU for authors, etc.). Such tags may exist in the original version and in this case they are translated from Chinese to English. Some tags are not visible on the internet browser, but occur in the texts; these should be kept.

-To add new tags to the text by analysing the

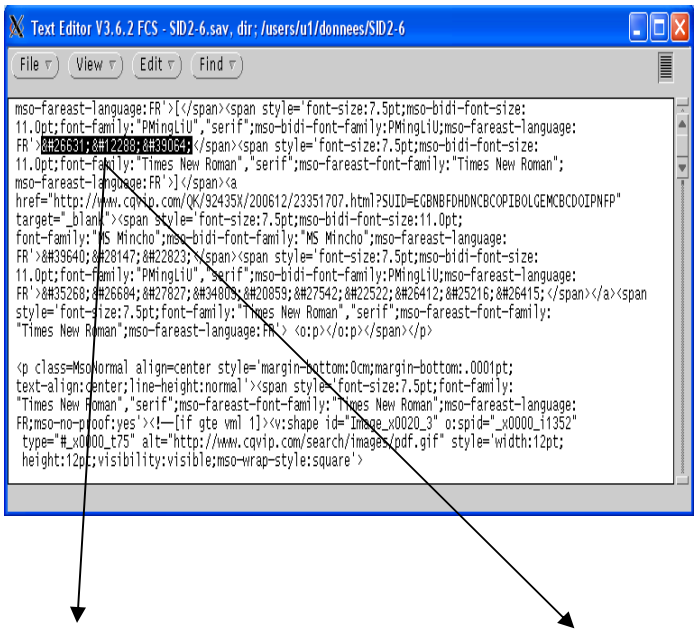
initial HTML tags.

-To retain the information which is coded in Latin characters or Arabic numerals such as dates, numbers or Western names (authors, technical formulae or elements).

This re-encoding is based on a parser and some re-writing rules as illustrated in figure 3.

Fig. 1. cqvip.com interface - the search engine is at the top of the figure.

Fig. 2. cqvip.com interface – the results are displayed.



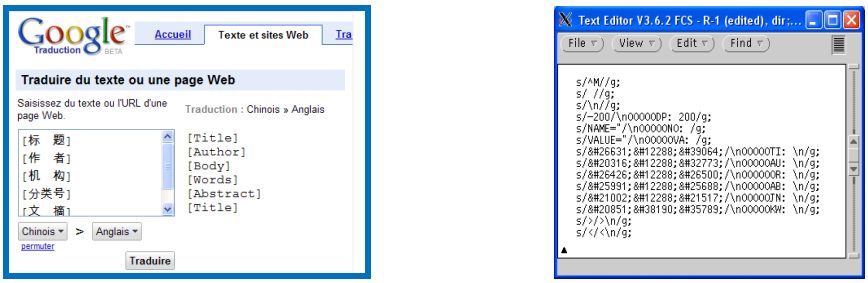


Fig. 3. Re-encoding CQVIP data - Google © translation followed up by information extraction using a Perl parser.

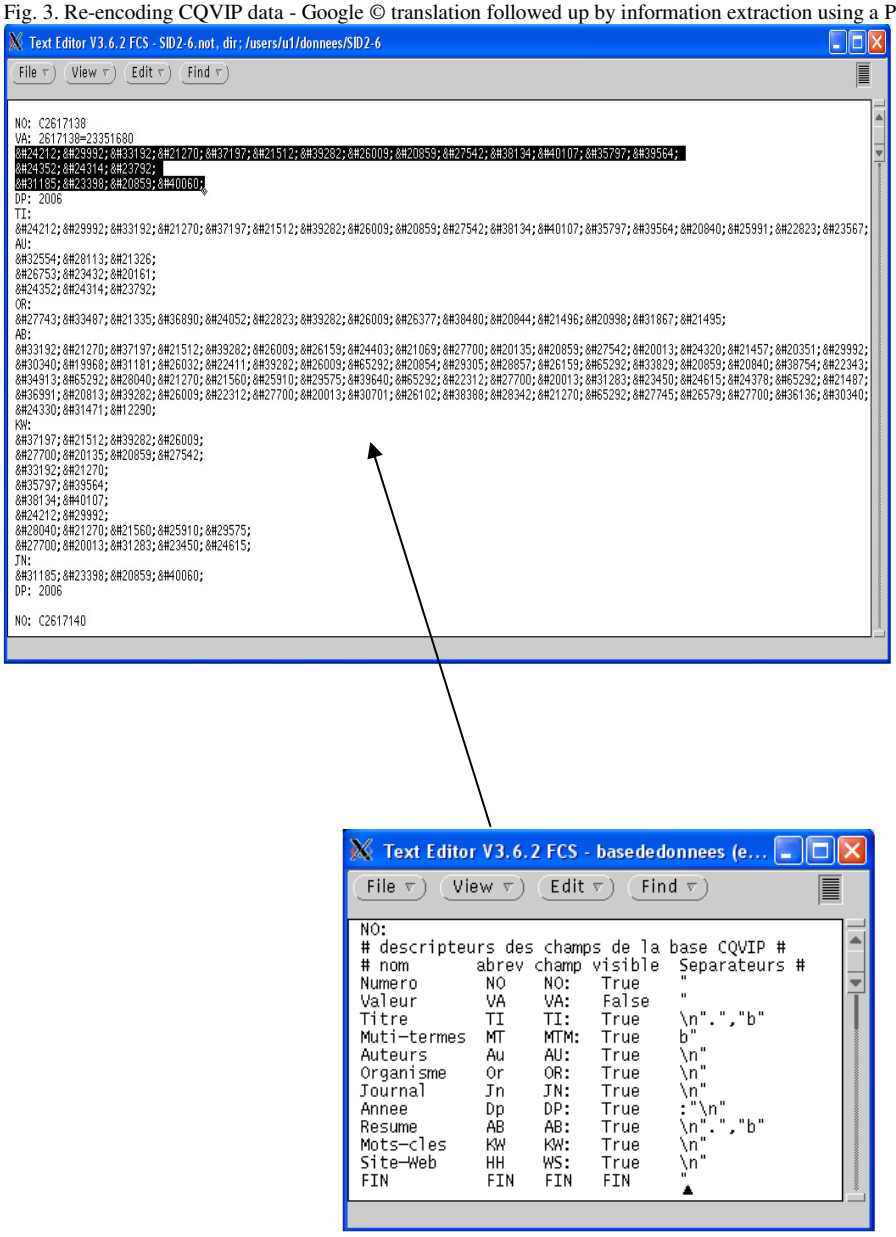


Fig. 4. A bibliographical reference that has been re-coded (tags in ASCII and content in Chinese UNICODE) and the corresponding metadata.

Figure 4 illustrates the results. Tags are written in ASCII whereas text (content) is in UNICODE. For example, in the C2617138 reference from figure 4, the publication title, first author of the publication, the journal in which it has been published, and the publication date constitute the beginning of the document. These information elements are tagged using the following field tags: TI:, AU:, JN:, DP.

When analysing the document visually, we can see that it consists of 3 authors (3 Chinese ideograms = 3 codes), only one organisation, 8 keywords (here each keyword is composed of 2 to 5

ideograms), the journal and one date (2006). We will see thereafter that the title and the abstract are analysed using a specific semantic process in order to detect repeated n-grams of ideograms that in fact do not correspond to any of the keywords. This adheres to a terminology that is not included in the initially provided indexes.

Metadata (at the bottom of figure 4) describe the new format of references: complete name for each field and its abbreviation, exact identifier of the field in the reference (ex: TI: for the field Title), TRUE means that this field will be used in the analysis, separators used to cut out text (character string, “\n” for carriage return, etc.).

C. *Translation problems*

Authors' names

To understand UNICODE (and hence Chinese), we list dictionaries that gather the correspondences between the names of authors in Chinese and their translation into phonetics (Pinyin) using the translator from Google ©. But in so doing, two difficulties arise:

-Google © fails when translating some of the names and in this case keeps the UNICODE (see 7th author figure 5),

-Several authors with different codes can be translated to give the same name. The ambiguity has to be corrected before any analysis takes place in order to avoid analysis mistakes.

In this case there is a failure in the translation process. We chose to keep the codes, but where there was ambiguity we added a code that helped to differentiate the names (e.g. LI-1, LI-2 and LI-3 refer to different translations that led to LI).

Keywords

Another translation problem can arise in relation to technical terminology (keywords, additional indexing, full text) because automatic translators struggle when the terms do not appear in their dictionaries (terms that are too technical or too recent), the context or the sentences are too complex or there is some ambiguity. Most of the time this uncertainty is resolved during the analysis itself: term clusters, for example, help in understanding a term because they occur with some terms that have been correctly translated. The problem is very similar for keywords associated with a particular publication. Indeed, some keywords, which are different in UNICODE, are translated similarly by translation engines. This phenomenon is fortunately rather rare and hence does not overly compromise the interpretation of the analysis. Of course, at the final stage, the views of an expert in the language are welcome.

Figure 6 presents the first phrases of the synonym dictionary based on the keyword field of the documents; it gives the correspondence between Chinese terms in UNICODE and their Google © translation in English. The number of occurrences of the terms is then calculated for English, thus the occurrences of a term may correspond to the sum of the occurrences of different Chinese terms.
In the example of figure 6, the most frequent term is "aquaculture"; it combines the occurrences of several Chinese forms. Even if the fusion is less problematic than in the case of homonyms found in particular authors, there is a risk here of losing some of the differences between the terms.

Other Problems

-For journal names there are no real problems.

-However, for the names of organisations the problem is that several forms can exist in different documents. This is mainly due to the way addresses are written. We therefore constructed a dictionary that brings together the different versions of the name of any given organisation.

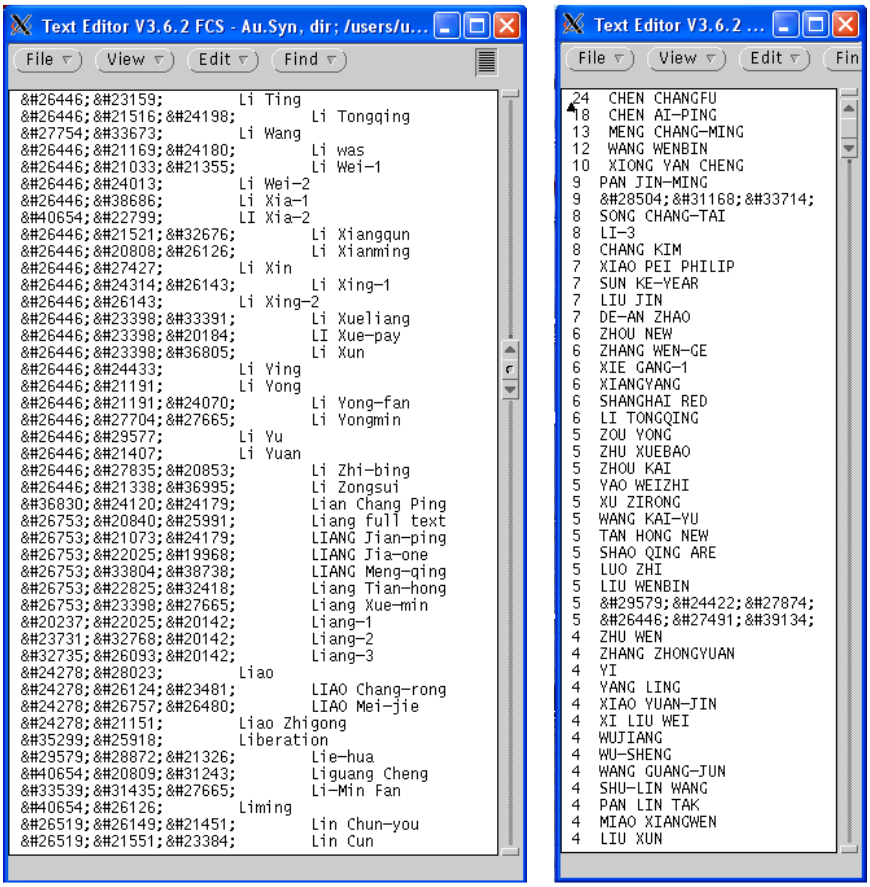


Fig. 5. UNICODE and corresponding Pinyin in addition to occurrences – Authors.

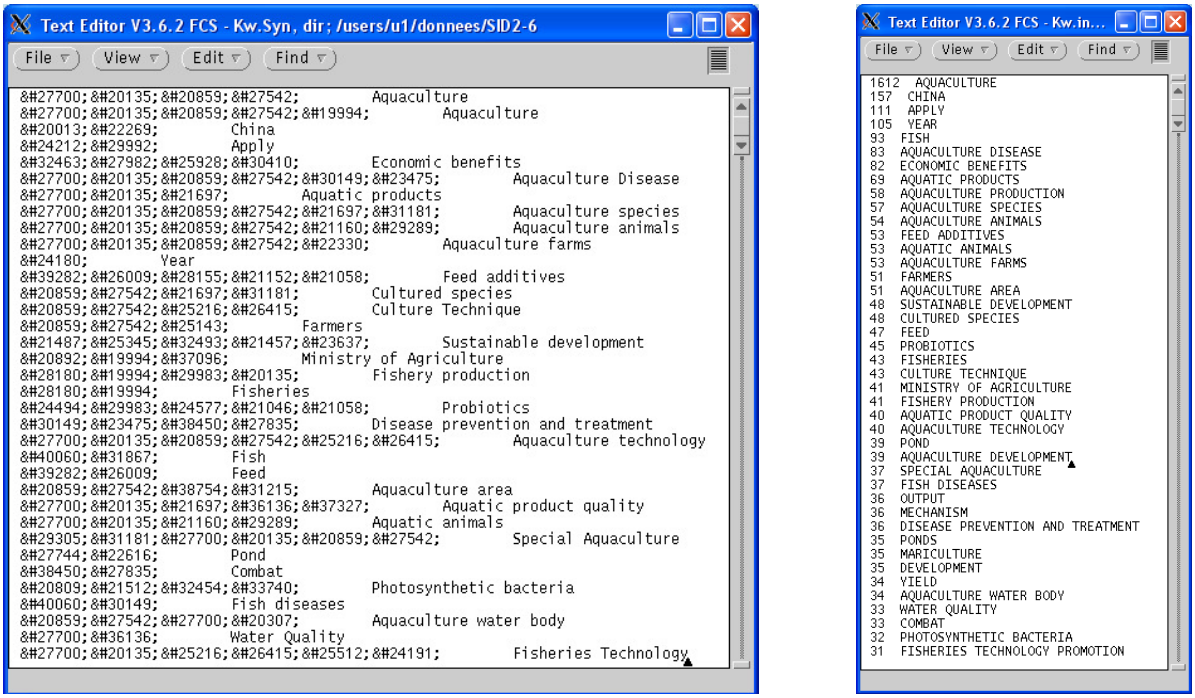


Fig. 6. UNICODE and corresponding phrase translation and synonyms (left side), phrase occurrences (right side), extracted from keywords.

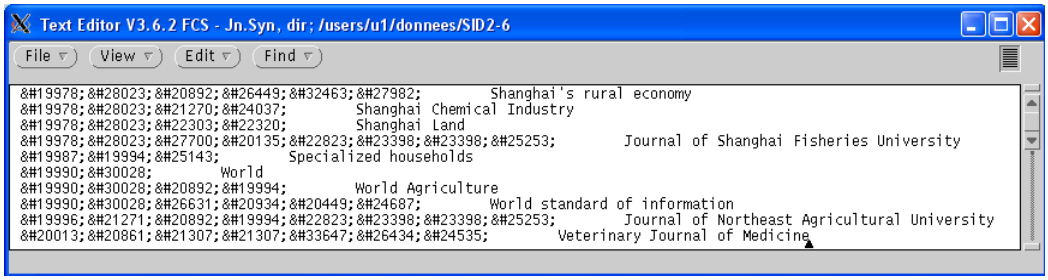


Fig. 7. Extract from the journal dictionary.

IV. ANALYSING AQUACULTURE IN CHINA

A. *Social networks*

As explained in the previous section, to begin with, authors' names are translated into English; then we resolve the problem of English homonyms where Chinese names have been translated.

Next we create a cross referencing table that cross references the authors' names; in this cross referencing table we consider authors that have written at least two publications. Indeed those who have published only one publication are not of any help when trying to extract relationships between authors.

Figure 8 presents the topology of the main teams. We can immediately see that there is very little co-authoring in the Chinese scientific publications we analysed. A second observation is that the teams are generally directed by a main author who has *control* of 2, 3 or 4 distinct sub-teams. Notice that in the figure, some names are not translated, whereas others are translated word by word and mean something in English. This has no impact on the results of the analysis.

- 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
- …

B. *Semantic networks*

In the same way it is meaningful to cross reference the keywords suggested in the documents and thus to extract a map of the terminology chosen by the editors or authors of the publication via the keyword field. Of course, using the keyword field does not help much to extract weak signals or novel signals because usually the keywords are more common terms. Conversely, strong signals and domain diversity are elements that we can extract.
Figure 9 displays the terms, which are circled in figure 10, belonging to one of the extracted term clusters. This figure displays the entire semantic network extracted from the analysed data.

C. *Analysing evolution*

Evolution can be analysed and visualised in many ways. In the next sub-sections we first analyse this evolution by taking into account the correlation that exists between journal names and dates. Then we consider the evolution of social networks or relationships between authors over time.

Correlation between time and journal names

In this section we analyse the profile of how the journals in which authors published during the four years of the study, namely 2004 to 2007, evolve. Correspondence analysis (Mardia et al, 1979), (Jolliffe, 2002) applied to the cross referencing table in which the two dimensions are Journals and Dates (Jn x Dp) allows us to visualise the various profiles on a regular tetrahedron (one dimension for each year) presented three dimensionally in figure 11.

In figure 11, top left corner, the sub-figure shows the years only and their corresponding direction with regard to the factorial axes. The same projection is applied to the journals in the rest of the figure, for example, in the top right corner the journals are those associated with 2007, meaning that they are associated with 2007 only, i.e. they are probably new journals or journals that have been recently integrated into CQVIP. On the edge of the tetrahedron the journals appear in the data collection over a 2 year period (for example 2006 and 2007 are on the edge of the right hand side of figure 11). Journals that appear over a 3 year period lie on one face of the tetrahedron. Finally, those appearing over a 4 year period are displayed inside the tetrahedron and converge towards the year in which they appear most.

Evolution of author relationships

A second method consists in using a three dimensional cross referencing table where two dimensions represent the authors (thus co-authoring is represented) and the third dimension corresponds to time. We can then visualise the evolution of the author network on a graph. This graph was developed in (Loubier, 2009). Time is distributed chronologically on a circle like the hours on a clock. The nodes corresponding to authors are attracted by these artificial nodes and are positioned towards the centre of the graph if they occur within the four time periods. On the contrary, the author nodes tend to be positioned in the direction of the corresponding reference when the author appears only once. They tend to be in a central position if the author appears in several consecutive periods. Figure 12 displays this network. At the bottom left corner, for example, the authors associated with 2006 are the only ones to appear.

This space-time analogy is similar to the correspondence analysis presented in figure 11, to which graph drawing techniques can be added. We then obtain a graph which shows the main teams (as in figure 8) with their respective evolutions.
The colour histogram attached to each node indicates its quantitative evolution; the end time period is represented in green whereas the beginning one is represented in red. The position with respect to its collaborative nodes indicates the time of the author's involvement with the team. The node bonds specify with whom and how long the collaboration lasted.

Figure 12 brings together the evolution of the main Chinese teams in the field of aquaculture. Some specific collaboration continues whereas others can be seen as emergent, moreover there are collaborations that either finish for a period of time or stop altogether. It is easy to locate the leaders of the author groups; indeed the size of each histogram is proportional to the appearances of the author in the collection. It is also easy to extract the authors that appear in the end year only (green) or in the beginning year (red). Finally figure 12 also shows the main authors who are responsible for the connections between teams, for example, when considering the team represented at the top of figure 12, the only leader who still publishes in the last period is Chen Changfu. He used to collaborate frequently with Meng Chang-Ming until 2006. He headed two separate teams of collaborative authors in 2004, worked with Shen Ke-Ray in 2005 and with one team consisting of 2 authors in 2006. In contrast, the three teams on the left side of figure 12 have many emergent authors and long-standing leaders. Other teams disappear; the four on the right hand side disappeared in 2006.

This analysis can be completed using a correspondence analysis based on the same three dimensional cross referencing tables. This analysis shows the trajectories of the authors when they collaborate with other authors. In the data we analysed, no such mobility could be extracted.

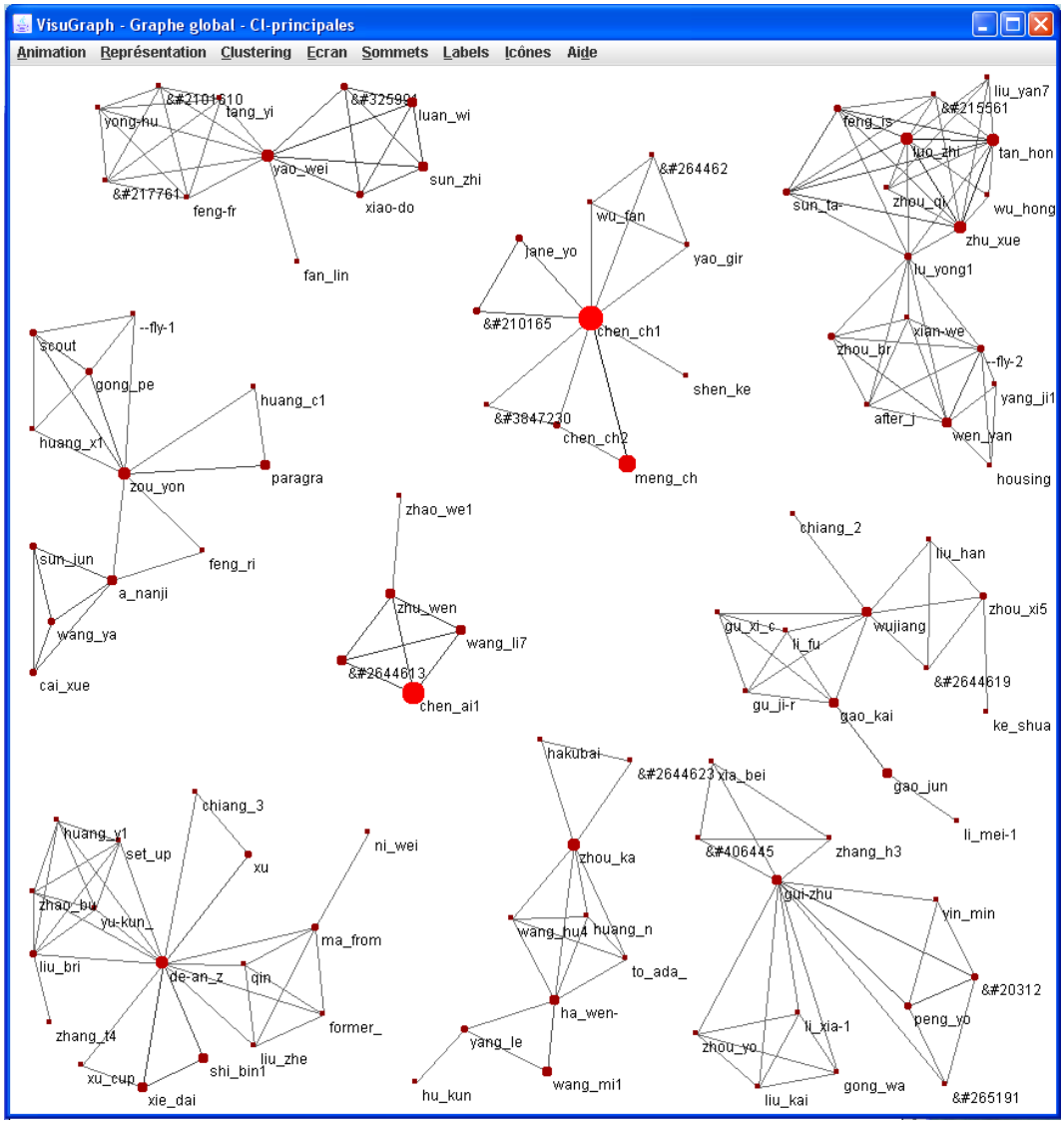


Figure 8 : Social network analysis - extraction of the main teams by authorship.

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, Kind, Water quality, etc.

Figure 9 : Terms belonging to one of the extracted term clusters.

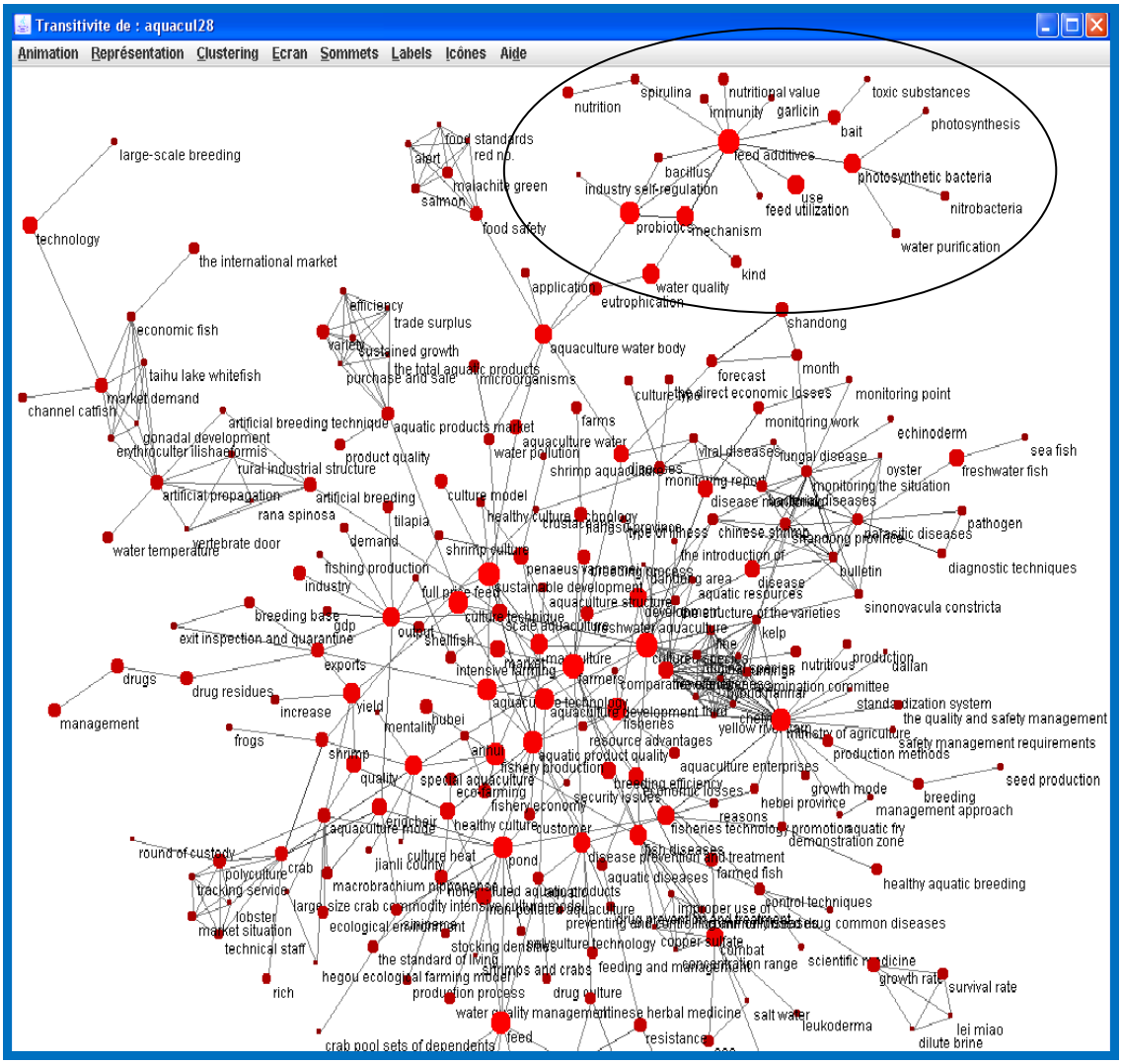


Figure 10: Semantic network based on the keywords from CQVIP.

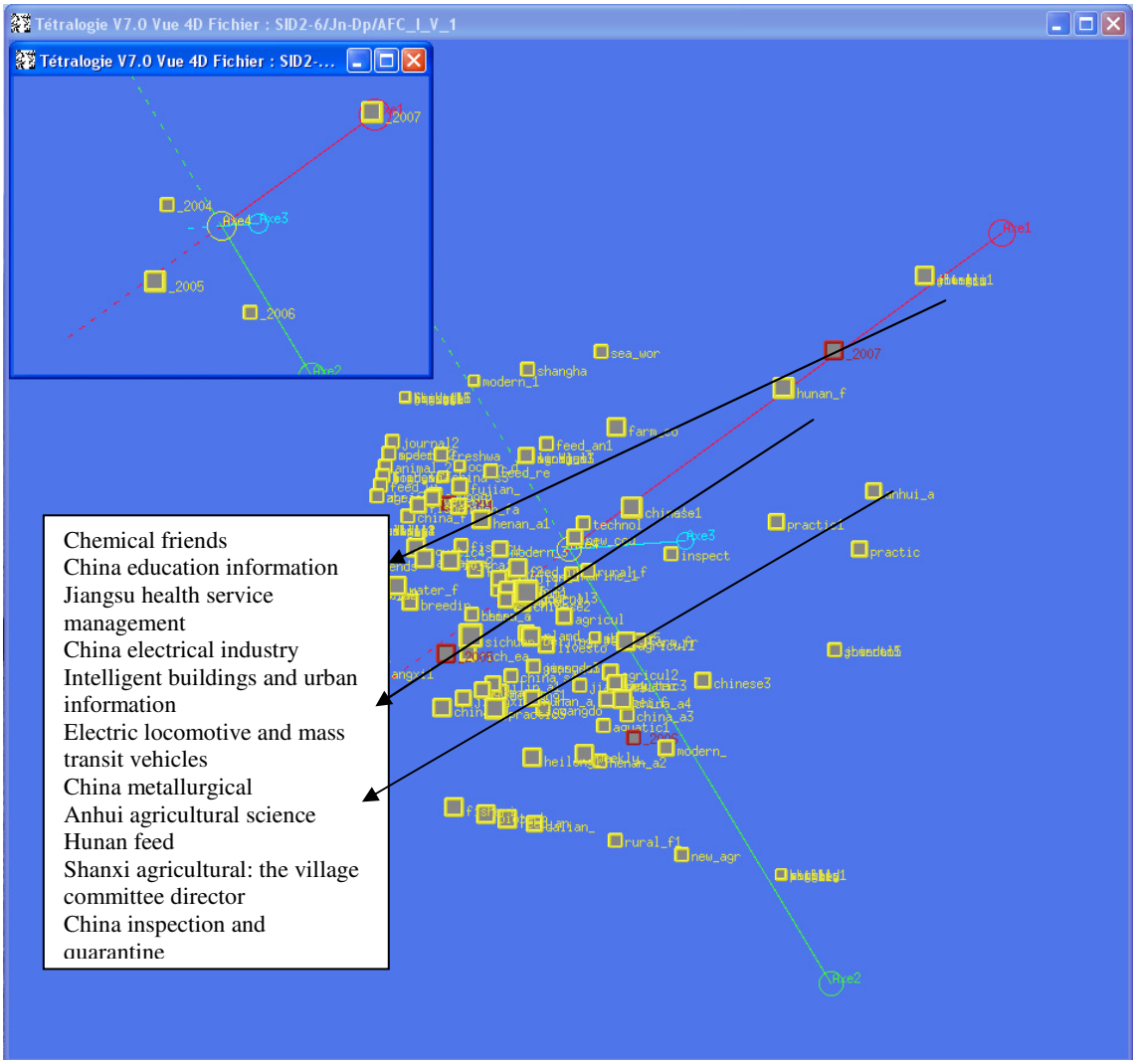


Figure 11: Visualising the results of a correspondence analysis on the first axes – journals x dates cross reference table.

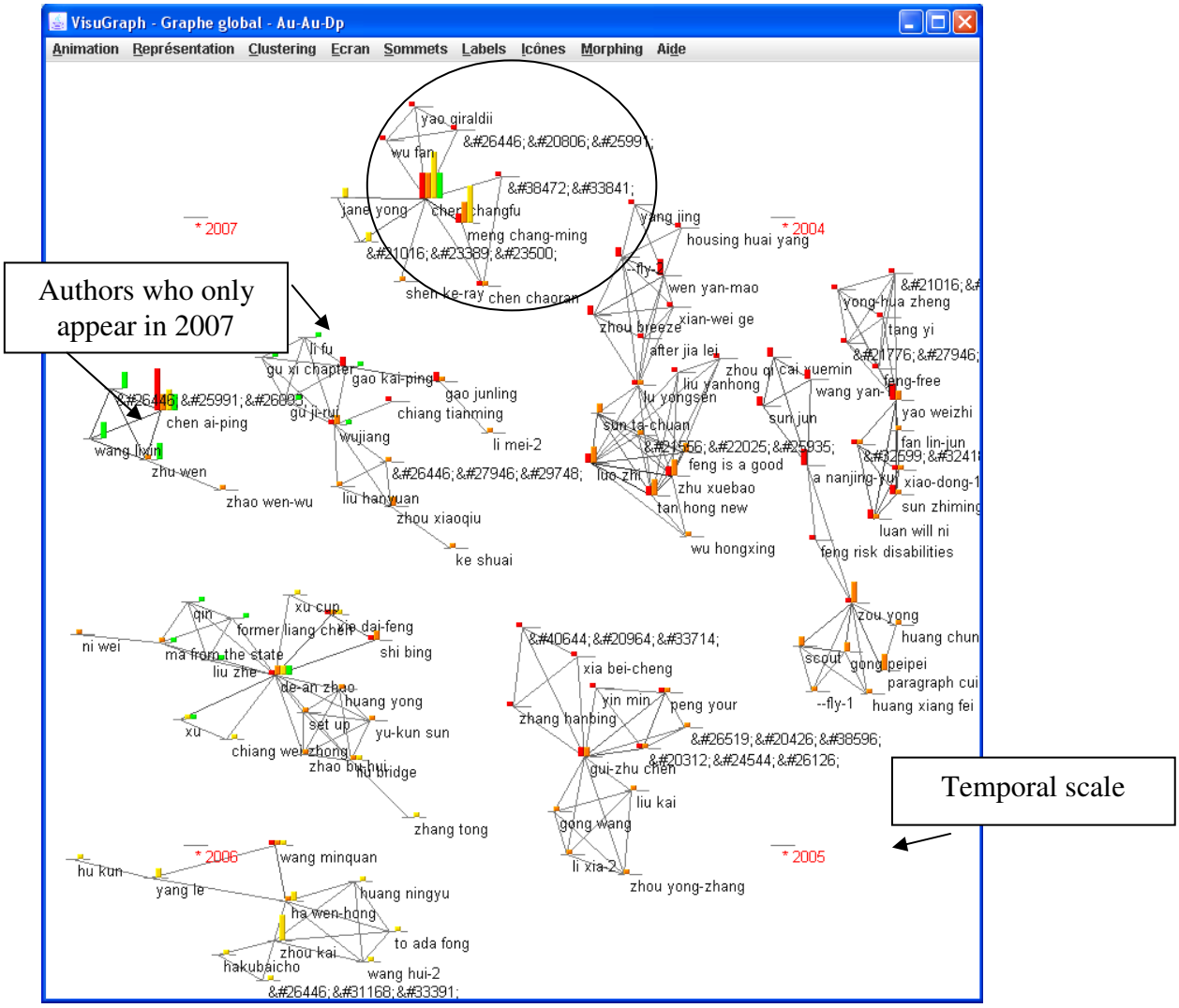


Figure 12: networking and evolution of the main teams (co-authoring).

D. *Semantic analysis of free text*

We use the dictionary of keywords we built and of which we present an extract in figure 6, including a stop-word list and a dictionary of synonyms (terms that are known to have similar meaning), to analyse the free text. Free text from the title and the abstract field of the documents is first reduced to chunks of text using punctuation. The n-grams of ideograms corresponding to the known keywords (from the keyword field) are then extracted from the text and completed by new n-grams of ideograms extracted automatically according to their frequency.

These new phrases of ideograms, which can include existing keywords, are translated into English in order to try to understand their meaning. If the translation we obtain using an automatic translator is meaningful with regard to the context but corresponds to a new term, then it is vital to have access to an expert in order to understand the context for this term and to confirm that it is an important term for the domain. These terms can correspond to important terms that are missing in the keyword field. Alternatively, we can analyse whether these new n-grams form clusters or not (Russet and Red, 2009). This can be carried out by analysing their co-occurrences in the document set. In this way it is possible to detect a weak signal (coherence, simultaneity, and consensus). Another way to validate the findings is to cross reference the new term with the other extracted elements (authors, organisations, keywords, journals and dates) and consider those that are related. This will be explained in the next section.

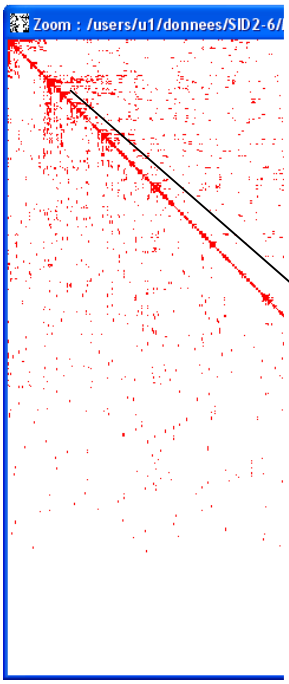
Using this approach and without knowledge of a language it is thus possible to detect implicit information that occurs in the corpus and which is inaccessible from a simple reading. The detection of the weak signals is in fact much in demand by decision makers because it corresponds to the need to detect innovation in order to make the right decisions (new avenues to explore, new products to use, etc.). Figure 13 presents a list of detected terms (new n-grams of ideograms) and an emergent semantic cluster.

E. Detecting weak signals

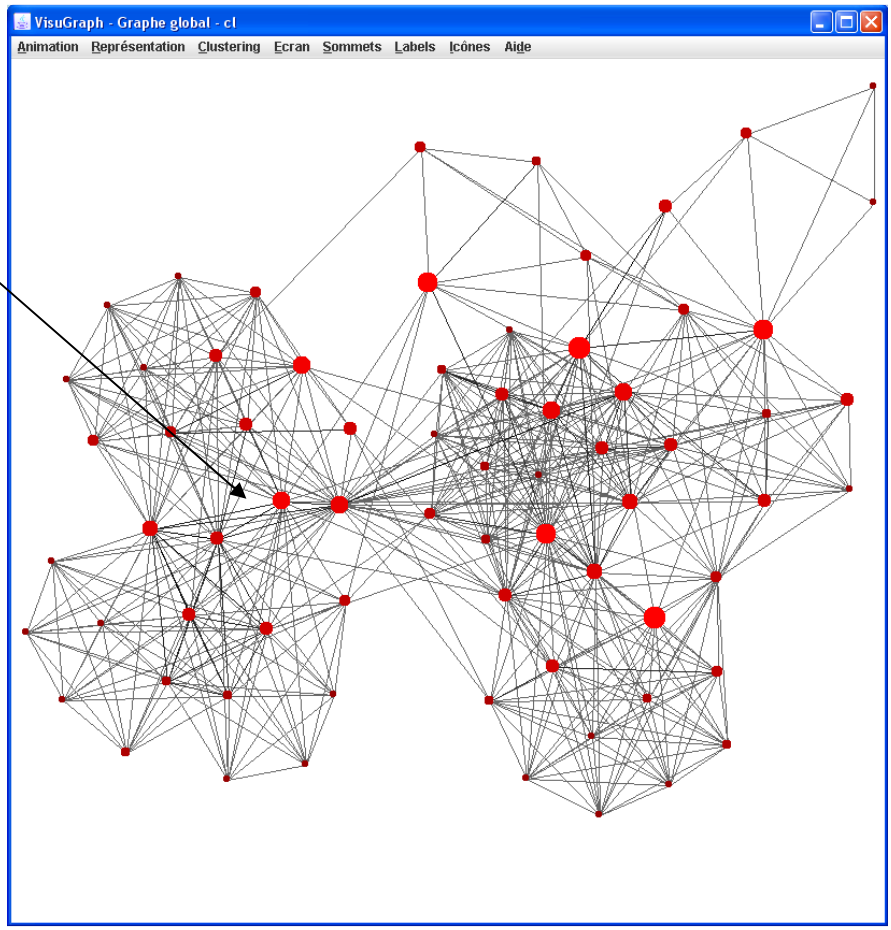
To detect weak signals, we first extract the keywords and the known terms from the title and abstract. Then we detect the new sequences that exceed a number of occurrences. Afterwards we cross reference these new n-grams with time and we keep only those which occur frequently during the end time period (here 2007). Finally these terms are cross referenced (co-occurrence) and we sort the subsequent matrix to obtain diagonal blocks. Each block represents an emergent concept identified by a new terminology which does not exist in the keyword field and which only occurs in some documents. Weak signals can then be validated by cross referencing them with all the other fields and in particular the keywords. In figure 14, part a) we represent the cross referencing matrix; each plot indicates a non-nil value for the cross referencing. Along the diagonal of the matrix, a certain number of clusters consist of new terms and correspond to a semantic group. Each cluster is extracted in a square sub-matrix and can be visualised in the form of a semantic graph (figure 14 b.). This information should then be submitted to an expert in the field for validation.

养殖塘	Breeding pond	养殖塘
养殖可持续&#	Sustainable development of	养殖可持续
养殖持续健&#	Sustained and healthy	养殖持续健
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&#	Work culture forum	养殖工作座
养殖息	Farming income	养殖息
养殖高产高&#	Breeding high yield and high	养殖高产高
养殖经济效&#	Economic benefits of	养殖经济效
养殖罗非鱼	Tilapia culture	养殖罗非鱼
养殖螃蟹	Breeding crabs	养殖蝻蟹
大水产养殖&#	Large aquaculture households	大水产养殖
水产品消费	Consumption of aquatic products	水产品消费
水产品出口	The export of aquatic products	水产品出口

Figure 10 : New terms extracted from free text that do not occur in the keyword field.



a) Cross referencing matrix in the form of plots.



b) Semantic network based on n-grams

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

c) n-grams and their translation.

Figure 11 : Analysis of newly detected terms and their clusters

V. FURTHER ANALYSIS: ARABIC

In this section we briefly present two other examples of resources on which an analysis can be carried out using the method we presented in the previous sections for Chinese. UNICODE UTF-8 can be extracted from the HTML source code.

With regard to the first example, Al Jazeera, the originality is being able to analyse the reactions of the blog users (see figure 15) and with regard to the Korean library we chose to analyse, we can see that the scale of the characters devoted to this language is different, but that the principle of analysis remains the same (see figure 16).

No matter what the collection and the data are, the challenge is to detect tagging that enables us to extract elements of information and hence build the cross referencing tables (actors, semantics, dates, etc.). Dictionaries of keywords and expressions are also very useful in the treatment of free text and in the detection of innovation therein.

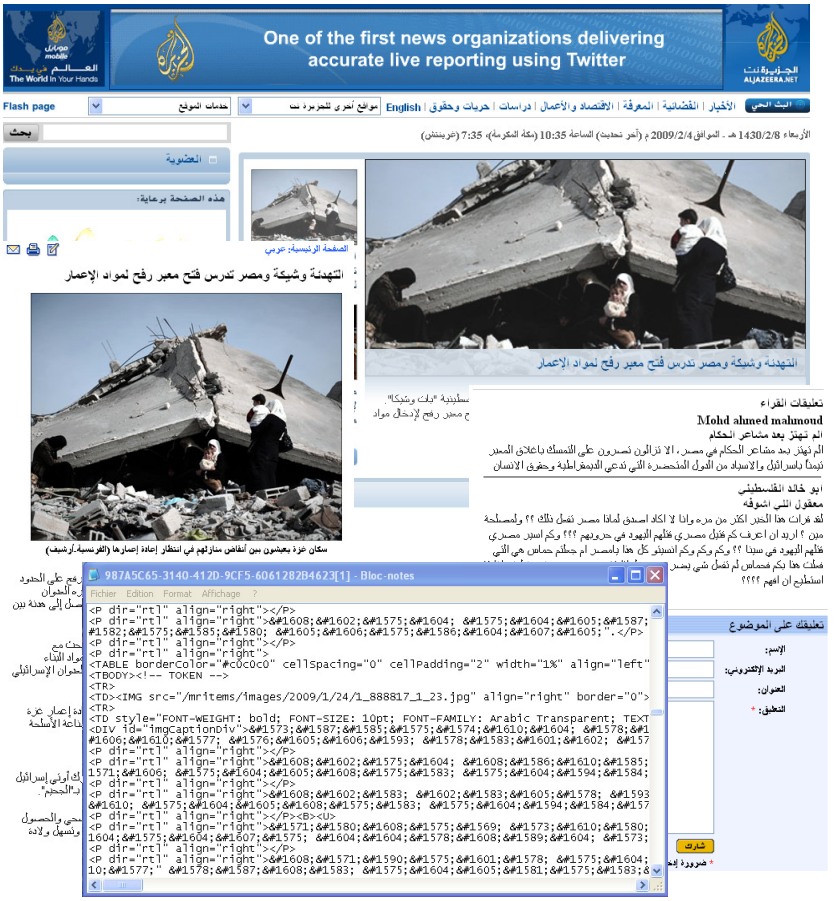


Figure 12: Aljazeera.net (document brief and associated blog)

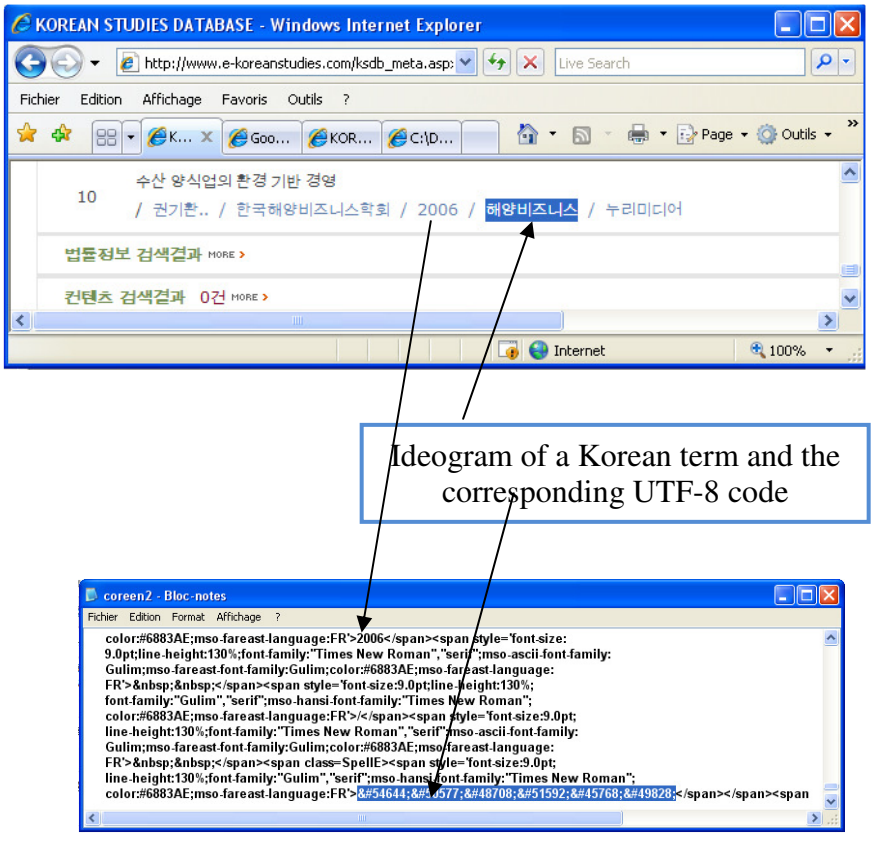


Figure 13: Korean from www.e-koreanstudies.com

VI. CONCLUSION

In this paper, we presented a method that answers the challenge of analyzing documents one cannot read because it is written in a foreigner language. This is crucial for enterprises in competitive intelligence and science watch activities.

The CQVIP library on which we carried out this analysis represents an example of the multiple sources that can be analysed using the method we present throughout this paper. Any language can be treated the same way. However, some issues have to be resolved in order to make this process fully usable and some additional work has to be undertaken:

-Building dictionaries (terms, etc.) and translating them into English (and/or into another language).

-Treating the named entities (for authors, organisations or journals): an automatic translation is sufficient, but there remain many ambiguities that have to be dealt with (importance of accents, pronunciation, context).

-The translated terms obtained by translating new detected terms or phrases in a statistical way will not be part of traditional dictionaries, either because they are too new or because other forms will be referenced. Checking the validity is then an issue if no expert is available to validate manually.

In future work it will thus be necessary to contemplate collaboration between different domain experts in:

-Text and data mining,

-Natural language processing (semantics, morphosyntaxic, ontologies, etc.),

-Languages (Chinese, Korean, Japanese, Arab, etc.),

-The fields to be analysed (scientific, technological, economic, geopolitical, etc.).

This collaboration between different experts could be useful as part of a two staged approach:

-Pre-processing data: homogenisation of the vocabulary, choice of the information granularity, translation, clarification, etc.).

-Interpreting results: very often it is useful to go back to those document sources consisting of free text, in which case it is important to understand both the language and the domain.

VII. REFERENCES

Chen C., Visualisation of Knowledge Structures, Handbook of Software Engineering and Knowledge Engineering, 2002.

Dousset B., (2009). Extraction de l'information implicite par analyse textuelle de sites Web en UNICODE, Veille Stratégique Scientifique et Technologique, CD-ROM.

Dousset B., (2008). Extraction of strategic information through analysis of major components. Datametrics Journal, 2(1).

Gaolin F., Hao Y., and Fumihito N., (2006). Chinese-English term translation mining based on semantic prediction, Proceedings of the COLING/ACL on Main conference poster sessions, 199–206.

Geroimenko V., and Chen C., (2002). Visualizing the Semantic Web XML-based Internet and Information Visualisation, Springer, ISBN 1-85233-576-9.

Ghalamallah I., Grimeh A., and Dousset B., (2007). Processing data stream by relational analysis, MODULAD, INRIA, n°36.

Ghalamallah I., Loubier E., and Dousset B., (2008). 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.

Guénec N., Loubier E., Ghalamallah I., and Dousset B., (2008). Management and analysis of Chinese database extracted knowledge. BCS IRSG Symposium: Future Directions in Information Access, British Computer Society.

Jolliffe, I.T. (2002). Principal Component Analysis, second edition, Springer.

He D., Wang J., Oard D. W., and Nossal M., (2003). User-assisted query translation for interactive CLIR, annual international ACM SIGIR conference on Research and development in information retrieval, 461-461 (demo)

Leydesdorff, Loet. (1995) The Challenge of Scientometrics: The development, measurement, and self-organization of scientific communications. DSWO Press/Leiden University, Leiden, 1995; at http://www.upublish.com/books/leydesdorff-sci.htm)

Li H., Cao Y, and Li C., (2003). Using Bilingual Web Data to Mine and Rank Translations, IEEE INTELLIGENT SYSTEMS, 54-59.

Loubier E., and Dousset B., (2007). Visualisation and analysis of relational data by considering temporal dimension. International Conference on Enterprise Information Systems, Vol. ISAS, INSTICC Press, 550-553.

Lu C., Xu Y., and G., Shlomo (2008). Web-Based Query Translation for English-Chinese CLIR. Computational Linguistics and Chinese Language Processing (CLCLP), 13(1). pp. 61-90.

Mardia, K.V., Kent, J.T., and Bibby, J.M. (1979). Multivariate Analysis, Academic Press.

Mothe J., Chrisment C., Dkaki T., Dousset B., and Karouach S., (2006). Combining mining and visualization tools to discover the geographic structure of a domain, Computers, Environment and Urban Systems, Geographic Information Retrieval (GIR), 30(4): 460-484

Mothe J., and Dkaki T., (1998). Interactive multidimensional document visualization, International ACM SIGIR conference on research and development in information retrieval, 363-364.

Peters C., (2009). What happened in CLEF 2009 - Introduction to the Working Notes, Cross Lingual Evaluation Forum.

Roux C., (2009). Methods to extract weak signals. International Journal of Competitive Intelligence, Strategic, Scientific and Technology Watch, SciWatch Journal, Hexalog, Barcelona - Spain, 2(1):23-29.

White H.D. and McCain K.W. (1998). Visualizing a discipline: an author co-citation analysis of information science, 1972-1995, In JASIS, 49(4), 327-355.

White H.D., (2003). Pathfinder networks and author co-citation analysis: A remapping of paradigmatic information scientists, In JASIST, 54(5), 423-434.

Zitt M., and Bassecoulard E., (1994). Development of a method for detection and trend analysis of research fronts built by lexical or co-citation analysis, In Scientometrics, 30, 333-351.