

People and semantic networks analysis based on texts in languages we do not understand

Bernard Dousset, Anass Elhaddadi, Josiane Mothe

Institut de Recherche en Informatique de Toulouse, IRIT UMR 5505
 Université de Toulouse, Université Paul Sabatier
 118, Route de Narbonne, F-31062 Toulouse cedex 9 (France)

dousset@irit.fr, haddadi@irit.fr, mothe@irit.fr

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. 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.

Index Terms— text mining, graph, semantic network, social network.

I. INTRODUCTION

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 [1]. Many approaches are employed to resolve this problem and query translation is one of them [2]-[3]. Reading documents that are not written in a language the

user is familiar with is an issue. [4] present an *English reading-assistance system* which suggests translations of words and phrases based on mining techniques. [5] 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 [6]. 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 [7]-[8]. 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 [9]-[8]. Groupings other than authors can be used for the purposes of correlation when mining scientific publications such as keywords, journals, etc. as presented in [10]. 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 [11]-[12]. 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 [13] scientific publications are mined in order to highlight groups of authors and their geographic relationships.

This paper extends an earlier work from [14]. This new version aims at spreading the results to an international audience.

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.

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

用户名: 密码: [登录](#) [IP登录](#) [用户注册/忘记密码](#) [电子期刊阅览室](#) [充值中心](#) [Google 学术](#) | [旧版入口](#)

VIP 维普资讯 题名/关键词 搜索 8000 余种期刊文章..... [文章搜索](#) [高级检索](#) [常见问题](#)

中文期刊·专业文章 [首页](#) [期刊导航](#) [知识社区](#) [学者空间](#) [学术机构](#) [专题导读](#) [充值中心](#)

2007 非常盘点: 民生关注热门榜

学科分类

医药卫生
 临床医学 基础医学
 预防医学 中医中药
 药学

工程技术
 石油工业 石油工程
 冶金工程 材料科学
 机械仪表 能源动力
 电气工程 电子电信
 化学工程 建筑科学
 水利工程 交通运输
 航空航天 环境安全
 自动化计算机
 纺织 食品

自然科学
 生物 化学 力学
 数学 物理 天文地球

农林牧渔
 畜牧兽医 水产渔业
 农学 林学

人文社科
 管理 财经 政治
 法律 哲学 宗教
 军事 教育 文化
 体育 艺术 语言
 文学 历史 地理

订阅服务
[电子期刊订阅](#)
[学科分类订阅](#)
[纸质期刊订购](#)
[在线充值中心](#)
[阅读充值资费](#)

论文发表/论文写作
 专业提供论文发表和写作服务
 服务保发保过 010-63397249
 QQ: 278123888

LPG——后石油时代“枯木逢春”
 作为车用清洁替代燃料中技术最成熟、最容易推广使用的LPG(液化石油气),近十年的推广运用却经历了一个曲折的发展过程,没能如人愿地顺利发展。在我们大力建设节约型社会的今天,关注和推广成熟替代燃料无疑是调整能源结构,改善大气环境,实现可持续发展的有效措施。

- 警惕不典型的干燥综合征 干燥综合征
- 工程轮胎生产现状与发展前景 工程轮胎
- 垂直管理模式下权力的配置与制约 垂直管理
- 中央与地方关系法治化基本问题研讨 中央与地方
- 我国私募基金“收益保底”探析
- 个人投资理财渠道大盘点 理财渠道
- 对有效市场假说的再认识 资本市场
- 伟大的资本博弈 资本博弈
- 社会舆论监督的法律保障问题研究 舆论监督
- “藏独”没有希望——国际舆论纷纷质疑达赖喇嘛

21世纪清洁能源——二甲醚
 · 人人参与,保护肝脏——爱肝日特别专题
 · 2007非常盘点:民生关注热门榜
 · 抗生素——治病还是致病?
 更多专题>>

搜索风向标 生物燃料 网络电视 节能减排 下一代网络 多功能干细胞 月球

期刊推荐 [期刊导航](#) [期刊评价](#)

医药卫生

- 警惕不典型的干燥综合征 《中国临床医生杂志》
- 噪声性聋发病机制 《国际耳鼻咽喉头颈外科杂志》
- 视网膜血管炎的研究进展 《眼科研究》
- 近视眼巩膜胶原的研究现状 《眼视光学杂志》
- 全子宫切除术手术途径研究进展 《中国妇产科临床杂志》
- 创伤性下颈椎不稳的诊断与稳定性重建 《实用诊断与治疗杂志》
- 糖代谢异常在冠心病发病机制中作用... 《中国现代临床医学》

工程技术

- 我国稀土工业发展现状及进展 《稀有金属》
- 轧钢技术的现状和新发展 《钢铁》
- 浅析国际矿产资源经济贸易发展趋势及转型 《黄金》
- 冷轧取向电工钢生产技术的进展 《特殊钢》
- 中国十大钢铁企业能耗分析及节能工作建议 《冶金能源》
- 彩电面板面临的机遇与挑战 《安徽冶金》
- 影响钨铜复合材料烧结致密度的两个因素 《中国铝业》

学者推荐 [更多学者>>](#)

张靖周
 所在的机构:南京航空航天大学
 主要研究方向:1. 传热传质学
 2. 燃烧学 3. 红外隐身

- 直升机红外抑制器两种进气方式下的实验研究
- 压电驱动液电液耦合射流流动特征研究
- 爆震管内波与火焰相互作用机理的试验
- 压电驱动自耦合射流流动特性实验
- 薄片射流冲击冷却换热特性的实验

水利·工程 倪晋仁 杨志峰 黄润秋 陈守煜
化学 俞善信 高廉 寇元 张克武
财经 王缙慈 张维迎 杨开忠 魏守华
教育研究 张尧学 谈松华 王翼生 何克抗
天文·地球 王绍武 郑荣才 付广 施雅凤

学术机构 [更多>>](#)

[浙江] 浙江大学附属第二医院
 [云南] 昆明冶金研究院
 [广西] 广西师范大学
 [湖北] 中国长江三峡工程开发总公司
 [江西] 东华理工学院

快乐学习,下一站?
灌水专区 -->

知识社区

推荐有礼,推荐有礼,推荐有礼

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

全选题录 下载题录 打印题录 Email题录
 按时间筛选: 显示方式:

1 **【标题】 日本对虾精养高产技术初探**
 【作者】 周晨光 陈佳颖 黄秀荣
 【机构】 浙江省舟山市定海区海洋与渔业局, 316000
 【文摘】 舟山市绿源水产养殖公司2005年采用围塘塑料大棚进行日本对虾精养高产技术研究, 每667m²养殖产量达到366.87kg, 经济效益十分显著。本文主要介绍该公司围塘日本对虾精养高产技术, 供参考。
 【刊名】 >>>齐鲁渔业-2008:25(2)-27-27 相关文章

2 **【标题】 微生态制剂及其在水产健康养殖中的应用**
 【作者】 潘小红 [1] 陈国贤 [1] 徐学峰 [2]
 【机构】 [1]河南省平顶山市农业技术推广站, 467000 [2]平顶山市水产技术推广站
 【文摘】 近年来, 在水产养殖业中, 微生态制剂作为绿色饲料添加剂、水质改良剂以及对鱼类健康、预防疾病、促进生长和品质改善所起的显著作用, 越来越被人们所重视。并以其无毒副作用, 无耐药性, 无残留污染, 效果显著等特点逐渐得到广大水产养殖从业者的认可, 不少地方把目光投注在微生物技术在水产养殖的应用上来, 利用微生物制剂的辅助作用建立水产健康养殖模式, 实现无公害化养殖。
 【刊名】 >>>齐鲁渔业-2008:25(2)-50-52 相关文章

3 **【标题】 高青县渔业生产驶入“高速路”**
 【作者】 李恒永 史春来
 【文摘】 高青县渔业生产进入快速发展阶段, 2007年全县水产养殖面积达到3400hm² (5.1万亩), 水产品总产量2.6万t, 产值达2.2亿元。发挥了水产业在全县农业中的四大支柱的作用, 被省海洋与渔业厅列入全省水产养殖重点县。
 【刊名】 >>>齐鲁渔业-2008:25(2)-61-61 相关文章

4 **【标题】 禹城渔业污染源清查摸底全面完成**
 【作者】 尹国存
 【文摘】 最近, 禹城市水产局根据《第一次全国污染源普查清查工作细则》、《山东省水产养殖业污染源普查实施方案》的要求, 认真落实省海洋与渔业厅《关于开展水产养殖业污染源清查工作的通知》精神, 成立专门班子, 强化工作措施, 制定清查方案, 确保人力、物力、车辆到位, 向各乡镇下发通知, 召开会议层层部署, 按要求深入农户,
 【刊名】 >>>齐鲁渔业-2008:25(2)-63-63 相关文章

5 **【标题】 无棣县狠抓水产品品质**
 【作者】 邵红梅 温孟泉
 【文摘】 经省水产品质检检测中心对无棣县水产养殖示范区对虾、缢蛏、梭子蟹等近10个水产品抽样监测化验表明, 各项化验指标均高于国家标准, 产品合格率100%。
 【刊名】 >>>齐鲁渔业-2008:25(2)-64-64 相关文章

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

The figure shows two windows. The top window is a text editor displaying a Perl script that fetches a webpage from cqvip.com and parses its content. The script uses the LWP::Simple module and the HTML::TreeBuilder module to parse the HTML. The bottom window shows the Google Translate interface with the same URL entered and the resulting HTML content displayed.

```

mso-fareast-language:FR'</span><span style="font-size:7.5pt;mso-bidi-font-size:11.0pt;font-family:'PMingLiU',serif;mso-bidi-font-family:PMingLiU;mso-fareast-language:FR">
<span style="font-size:7.5pt;mso-bidi-font-size:11.0pt;font-family:'Times New Roman',serif;mso-fareast-font-family:'Times New Roman';mso-fareast-language:FR'>
<span style="font-size:7.5pt;mso-bidi-font-size:11.0pt;font-family:'MS Mincho',mso-bidi-font-family:'MS Mincho';mso-fareast-language:FR">
<span style="font-size:7.5pt;mso-bidi-font-size:11.0pt;font-family:'PMingLiU',serif;mso-bidi-font-family:PMingLiU;mso-fareast-language:FR">
<span style="font-size:7.5pt;font-family:'Times New Roman',serif;mso-fareast-font-family:'Times New Roman';mso-fareast-language:FR'>
</span></p>
<p class=MsoNormal align=center style="margin-bottom:0cm;margin-bottom:.0001pt;text-align:center;line-height:normal">
<span style="font-size:7.5pt;font-family:'Times New Roman',serif;mso-fareast-font-family:'Times New Roman';mso-fareast-language:FR;mso-no-proof:yes">
<!--[if gte vml 1]>
<v:shape id="Image_x0000_3" o:spid="x0000_i1352" type="#x0000_t75" alt="http://www.cqvip.com/search/images/pdf.gif" style="width:12pt; height:12pt; visibility:visible;mso-wrap-style:square">

```

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

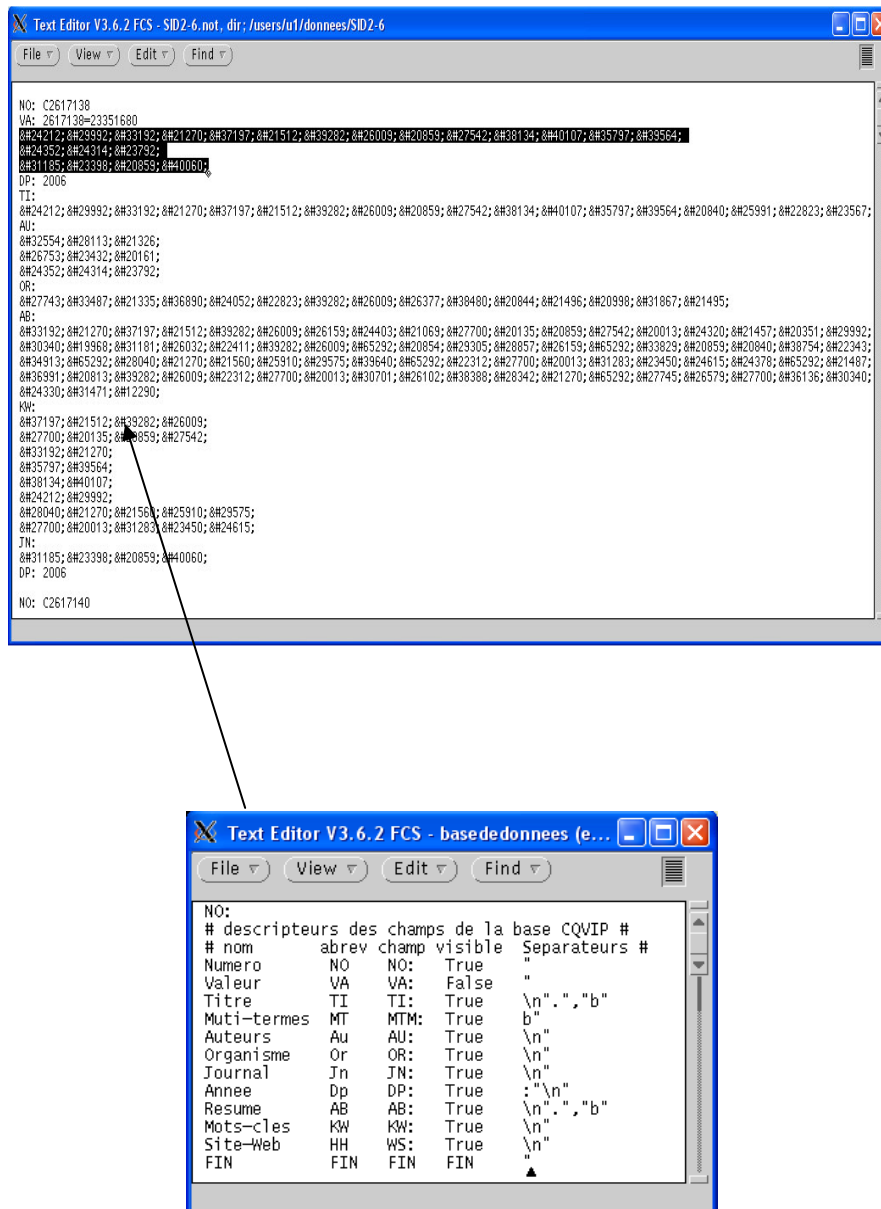


Fig. 4. A bibliographical reference that has been re-coded (tags in ASCII and content in Chinese UNICOD) and the corresponding metadata.

Figure 4 illustrates the results. Tags are written in ASCII whereas text (content) is in UNICOD. 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 UNICOD (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.

The figure shows two screenshots of a text editor window titled 'Text Editor V3.6.2 FCS - Au.Syn, dir; /users/u...'. The left screenshot displays a list of authors with their UNICODE codes and corresponding Pinyin names. The right screenshot displays a list of authors with their UNICODE codes and corresponding Pinyin names, including occurrence counts.

UNICODE	Pinyin	Occurrences
李 婷	Li Ting	
李 同 庆	Li Tongqing	
汪 莉	Li Wang	
李 励 年	Li was	
李 利 卫	Li Wei-1	
李 巍	Li Wei-2	
李 霞	Li Xia-1	
黎 夏	LI Xia-2	
李 向 群	Li Xiangqun	
李 先 明	Li Xianming	
李 欣	Li Xin	
李 建 星	Li Xing-1	
李 星	Li Xing-2	
李 学 良	Li Xueliang	
李 学 付	LI Xue-pay	
李 学 迅	Li Xun	
李 影	Li Ying	
李 勇	Li Yong	
李 勇 帆	Li Yong-fan	
李 永 民	Li Yongmin	
李 玉	Li Yu	
李 原	Li Yuan	
李 治 兵	Li Zhi-bing	
李 博 邃	Li Zongsui	
连 常 平	Lian Chang Ping	
梁 全 文	Liang full text	
梁 剑 平	LIANG Jian-ping	
梁 嘉 一	LIANG Jia-one	
梁 萌 青	LIANG Meng-qing	
梁 天 红	Liang Tian-hong	
梁 学 民	Liang Xue-min	
伍 嘉 亮	Liang-1	
岳 耀 亮	Liang-2	
翟 旭 亮	Liang-3	
廖 海	Liao	
廖 昌 容	LIAO Chang-rong	
廖 梅 杰	LIAO Mei-jie	
廖 功	Liao Zhigong	
解 放	Liberation	
王 烈 华	Lie-hua	
黎 光 程	Liguang Cheng	
范 立 民	Li-Min Fan	
黎 明	Liming	
林 春 友	Lin Chun-you	
林 启 存	Lin Cun	
24	CHEN CHANGFU	
18	CHEN AI-PING	
13	MENG CHANG-MING	
12	WANG WENBIN	
10	XIONG YAN CHENG	
9	PAN JIN-MING	
9	潘 秀 莲	
8	SONG CHANG-TAI	
8	LI-3	
8	CHANG KIM	
7	XIAO PEI PHILIP	
7	SUN KE-YEAR	
7	LIU JIN	
7	DE-AN ZHAO	
6	ZHOU NEW	
6	ZHANG WEN-GE	
6	XIE GANG-1	
6	XIANGYANG	
6	SHANGHAI RED	
6	LI TONGQING	
5	ZOU YONG	
5	ZHU XUEBAO	
5	ZHOU KAI	
5	YAO WEIZHI	
5	XU ZIRONG	
5	WANG KAI-YU	
5	TAN HONG NEW	
5	SHAO QING ARE	
5	LUO ZHI	
5	LIU WENBIN	
5	王 彦 波	
5	李 正 飞	
4	ZHU WEN	
4	ZHANG ZHONGYUAN	
4	YI	
4	YANG LING	
4	XIAO YUAN-JIN	
4	XI LIU WEI	
4	WUJIANG	
4	WU-SHENG	
4	WANG GUANG-JUN	
4	SHU-LIN WANG	
4	PAN LIN TAK	
4	MIAO XIANGWEN	
4	LIU XUN	

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

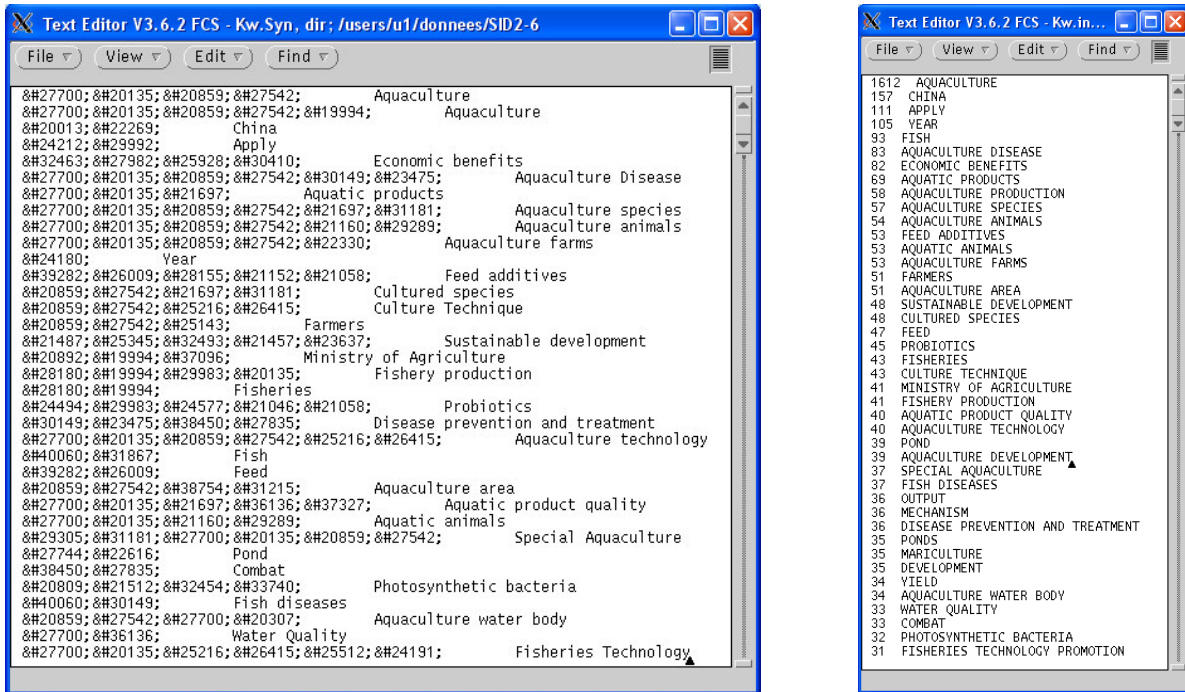


Fig. 6. UNICODE and corresponding phrase translation and synonyms (left side), phrase occurrences (right side), extracted from keywords.

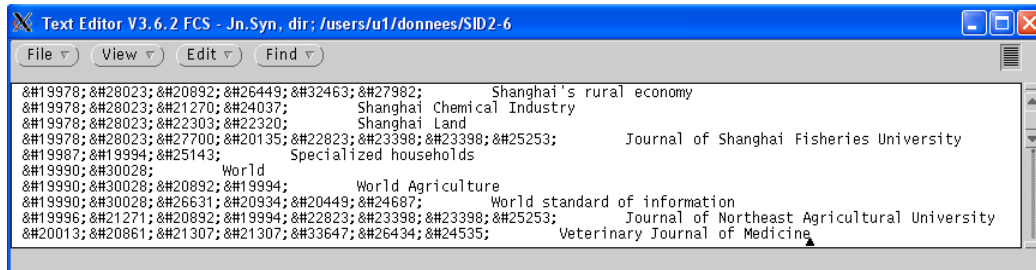


Fig. 7. Extract from the journal dictionary.

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 [15]-[16] 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 [17]. 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.

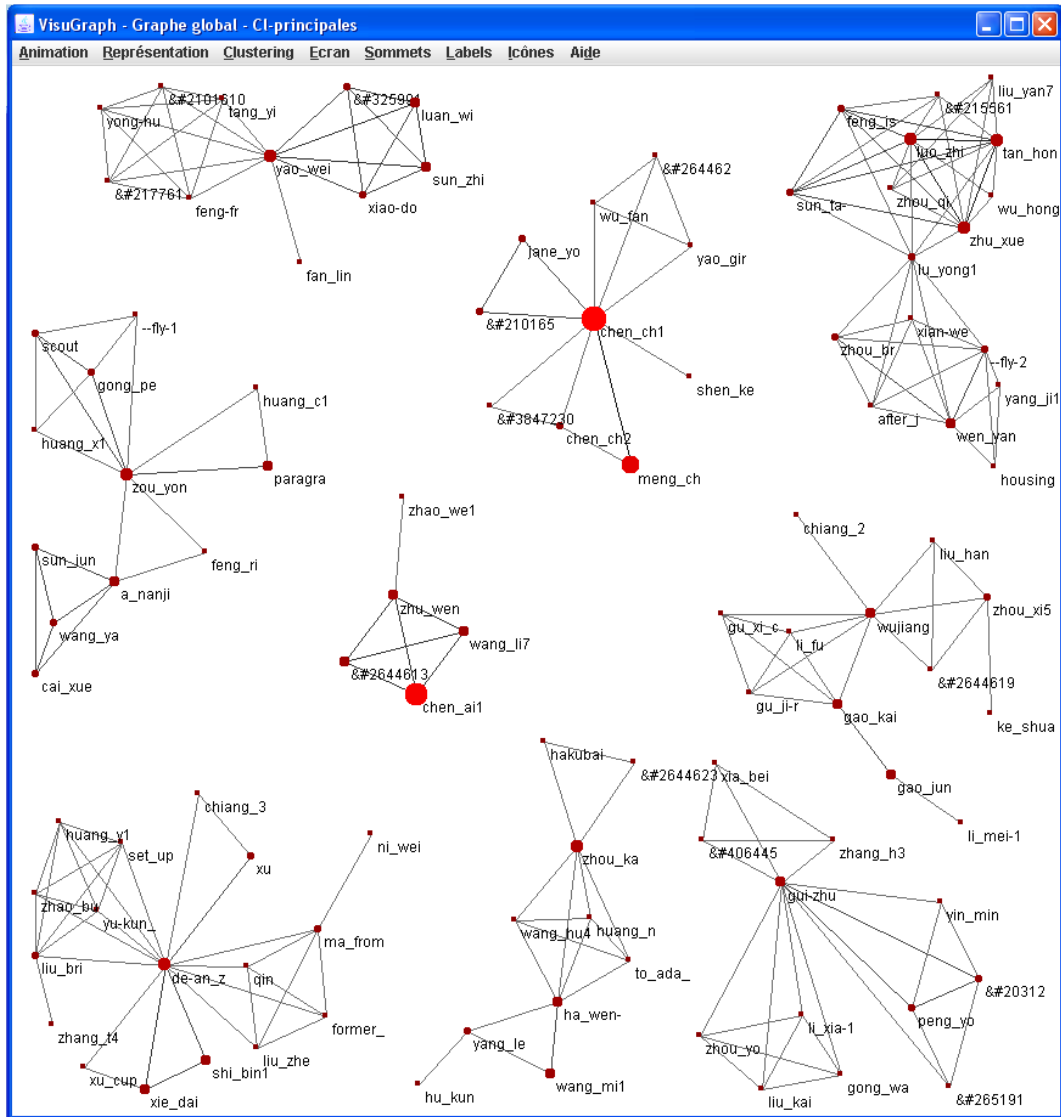


Fig. 8. Social network analysis - extraction of the main teams by authorship.

Feed additives, Nutrition, Spirulina, Nutritional value, Immunity, Garlicin, Bait, Toxic substances, Photosynthetic bacteria, Photosynthesis, Nitrobacteria, Water purification, Feed utilization, Bacillus, Probiotic, Industry self-regulation, Mechanism, Kind, Water quality, etc.

Fig. 9. Terms belonging to one of the extracted term clusters.

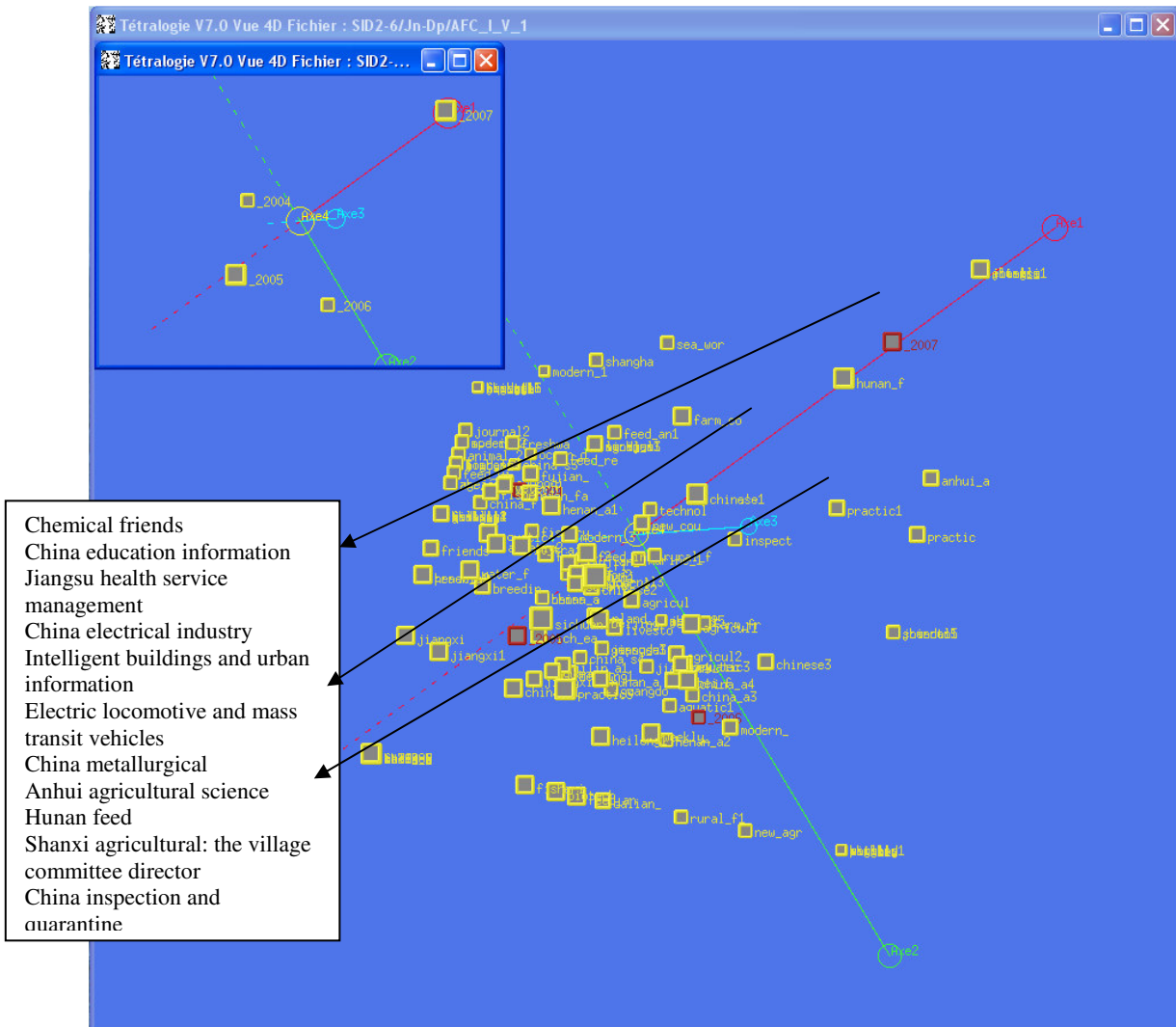


Fig. 11. Visualising the results of a correspondence analysis on the first axes – journals x dates cross reference table.

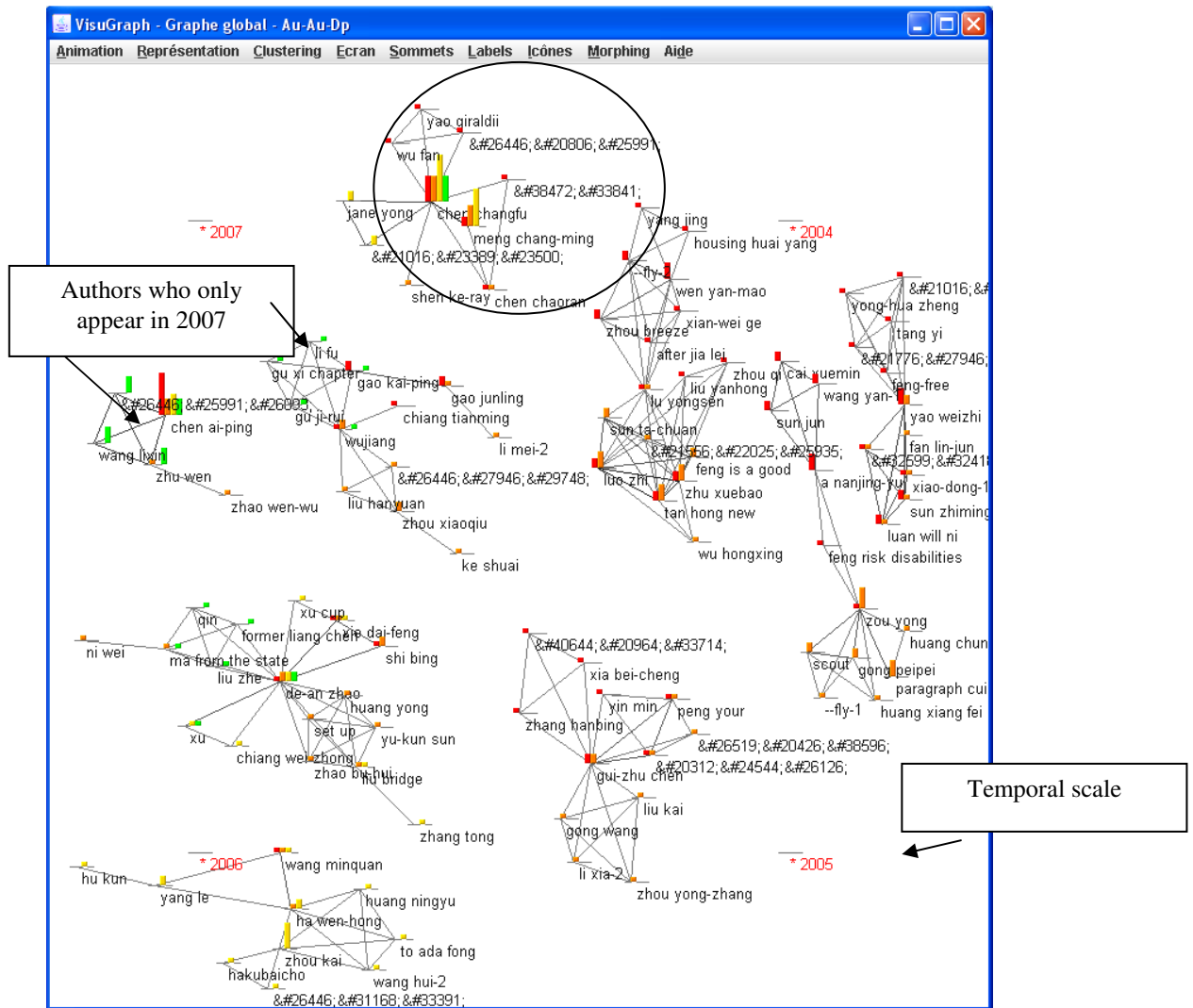


Fig. 12. Networking and evolution of the main teams (co-authoring).

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 [18]. 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

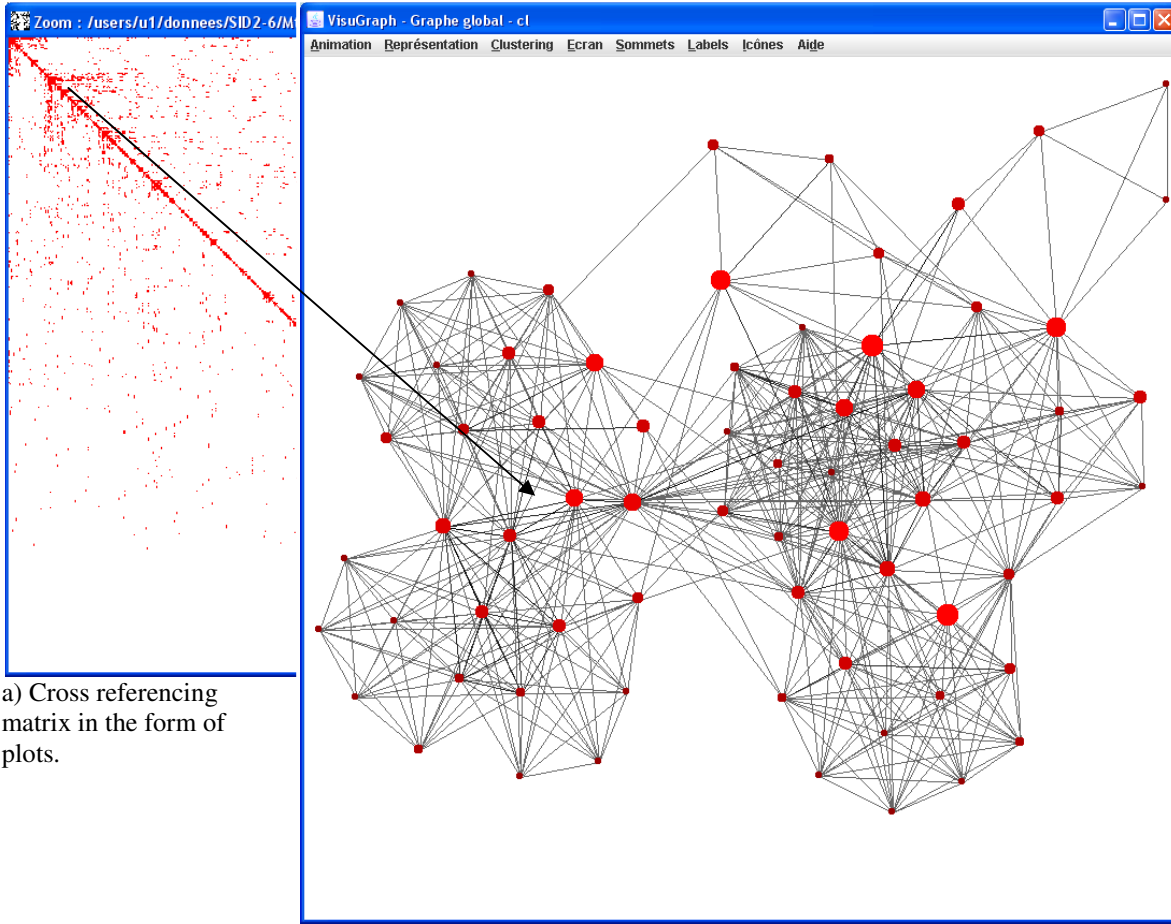
the new sequences that exceed a number of occurrences. Afterwards we cross reference these new n-grams with time and we keep only those which occur frequently during the end time period (here 2007). Finally these terms are cross referenced (co-occurrence) and we sort the subsequent matrix to obtain diagonal 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

养殖塘
 养殖可持续发
 养殖持续健康
 养殖河蟹
 养殖船
 养殖良种
 养殖大菱鲆
 养殖农户
 养殖病原体
 养殖工作座谈
 养殖息
 养殖高产高效
 养殖经济效益
 养殖罗非鱼
 养殖螃蟹
 大水产养殖户
 水产品消费
 水产品出口

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

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

Fig. 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.

The screenshot displays the Aljazeera.net website interface. At the top, it states "One of the first news organizations delivering accurate live reporting using Twitter". The main content area features a large photograph of a severely damaged building after an earthquake, with people seen amidst the rubble. A video player is positioned below the photo. To the right, there is a text section with Arabic text and a list of names, including "Mohd ahmed mahmoud". A small window in the foreground shows the HTML code for the text, including various attributes like font size and family, and a list of names encoded in UTF-8.

Fig. 12. Aljazeera.net (document brief and associated blog)

The screenshot shows a Windows Internet Explorer browser window displaying search results on the website www.e-koreanstudies.com. The search results include a link to "해양비즈니스 / 누리미디어". A text box highlights the Korean term "해양비즈니스" and its corresponding UTF-8 code. Arrows point from the text box to the HTML code in a separate window below, which shows the code for the highlighted text, including attributes like font size and family, and the UTF-8 code.

Fig. 13. Korean from www.e-koreanstudies.com

VI. CONCLUSION

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, morphosyntactic, 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

- [1] Peters C., (2009). What happened in CLEF 2009 - Introduction to the Working Notes, Cross Lingual Evaluation Forum.
- [2] 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)
- [3] 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.
- [4] Li H., Cao Y., and Li C., (2003). Using Bilingual Web Data to Mine and Rank Translations, IEEE INTELLIGENT SYSTEMS, 54-59.
- [5] 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.
- [6] 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>
- [7] 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.

- [8] White H.D., (2003). Pathfinder networks and author co-citation analysis: A remapping of paradigmatic information scientists, In JASIST, 54(5), 423-434.
- [9] Zitt M., and Bassecouard 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.
- [10] Mothe J., and Dkaki T., (1998). Interactive multidimensional document visualization, International ACM SIGIR conference on research and development in information retrieval, 363-364.
- [11] Chen C., Visualisation of Knowledge Structures, Handbook of Software Engineering and Knowledge Engineering, 2002.
- [12] Geroimenko V., and Chen C., (2002). Visualizing the Semantic Web XML-based Internet and Information Visualisation, Springer, ISBN 1-85233-576-9.
- [13] 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
- [14] Dousset B., (2009). Extraction de l'information implicite par analyse textuelle de sites Web en UNICODE, Veille Stratégique Scientifique et Technologique, CD-ROM.
- [15] Mardia, K.V., Kent, J.T., and Bibby, J.M. (1979). Multivariate Analysis, Academic Press.
- [16] 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.
- [17] 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.