# Discerning Industrial Networks, Clusters and Competences - An Alternative View Using Web Mining Techniques

John R. Williams, Dimitris Assimakopoulos

# Discerning Industrial Networks, Clusters and Competences–an Alternative View using Web Mining Techniques

John R Williams[1], Dimitris Assimakopoulos[2]

[1]Director, Fabriam Ltd,, United Kingdom.
jrw@fabriam.com; J.Williams@ncl.ac.uk
[2]Professor and Director of Doctoral Programs, Grenoble Ecole de Management,
Europole, 12 rue Pierre Semard, BP127, 38003 Grenoble, France.
dimitris.assimakopoulos@grenoble-em.com

**Abstract.** This short extract is part of a wider study into the use of the web for research into the presence and structure of industrial clusters and is concerned here with the discernment of networks of firms and the presence of commonalities of competence amongst firms within the identified networks.
The research shows that the information that can be extracted using web based methods is sufficiently informative to gain a wide and detailed picture of industrial activity in the locale under study. In addition to finding evidence of industry clusters and networking activity hidden by company list based investigations the methodology developed has shown that for the region studied, we are looking not so much at the clustering of artefacts *per se* but the clustering of 'competencies' in a wide range of sectors that share both common antecedents and current practice in engineering skills for the design and manufacture of large structures that operate in difficult or even hostile environments.

## 1 Introduction

In recent years the notion of industrial clustering has been the subject of much research and investigation and the number of available publications on the subject is significant. Much of this research activity stems from the part that industrial clusters are thought to play in the competitiveness of defined areas, such concepts having being promoted on a world stage by prominent authors of whom Porter (1998a)(1998b)(2000) is perhaps the best known.
A good definition of clustering in an industrial context is from (Van den Berg, Braun and van Winden 2001):

> *"The popular term cluster is most closely related to this local or regional dimension of networks. Most definitions share the notion of clusters as localised networks of specialised organisations, whose production processes are closely linked through the exchange of goods, services and/or knowledge".*

However it has been observed that many of the tools and data available to those engaged in cluster research have known shortcomings particularly related to descriptive data on firm activity and a number of commentators have given the opinion that such deficiencies can lead ultimately to significant errors and omissions related to the discernment of clustering activity (Feldman, Francis and Bercovitz 2005) and particularly by (Porter 1998c) when he complained, in relation to cluster boundaries, that they:

*"rarely conform to standard industry classification systems [SIC] which fail to capture many important actors in competition as well as linkages across industries ….Because parts of a cluster often fall within different traditional industrial or service categories, significant clusters may be obscured or even go unrecognised".*

## 2 Problem Statement

In the introduction we briefly outlined the difficulties and now postulate that the Web as a vast and powerful information source could provide some additional insights into the basic description of individual firms that go to make up sectors, industries, networks and clusters. In this context the research question becomes:
*'Can the use of the internet and the world wide web as an information resource add anything useful to more conventional methods of researching industrial clusters and networks?'*

## 3 Towards a Web Based Methodology

The basic principle was to submit a database of regional company URLs to a spider[1] to extract relevant text present on each company website. This sounds a simple task but the presence of large amounts of irrelevant noise makes the task less straightforward. Even a simple excursion into the web using URLs for say a group of industrial companies shows that the above approach is fraught with problems due to the mass of 'messy' text and other data including hidden programming instructions that characterise most web pages. The result of this is that some way has to be found to deal with the type of information that obscures 'relevant' data or text in the search to find information that helps the objective of describing the firm and its activities.

In the event and after some considerable development and test a proprietary program (http://www.phantomsearch.com) written initially to spider large single sites proved effective and this was used in all subsequent tests.

In designing the test the following boundary conditions were adopted:

1. The catchment area would be the North East of England. This is the smallest of the English RDA[2] regions having in 2007 about 42000 VAT[3] registered companies and a long tail of micro firms falling below this VAT threshold.
2. The number of companies scanned should be sufficient to give confidence in the robustness of any comparisons

The next stage therefore involved acquiring the URLs of as many regional firms and other relevant organisations as was reasonably practical. No single database existed of all companies and in the records obtainable from many of the most prominent data providers the URL was often absent as a field in the company record. The basic method of approach was to acquire as many URLs from regional organisations as

[1] A spider can be thought of a web crawler with some inherent capability that allows it to act in the manner of an intelligent agent operating to a specific set of instructions.
[2] Regional Development Agency
[3] Value Added Tax. In the UK at the time this was £57000 of company turnover

possible from a credit rating agency followed by further company data obtained from trade associations, the regional Chamber of Commerce, other membership organisations and trade directories. The net result of the whole exercise was that the list of what appeared to be valid URLs numbered 14000.

We were now in a position to submit these URLs to the spider program with the aim of acquiring text from all text, html, word documents and translated pdf files to be placed into a suitable database which was then fully indexed.
The metrics for this database when complete were :

- File size 60.81Mb
- Number of unique keywords 177432 with 231 noise words (not indexed)

What this means is that we now have, fully indexed the collective text from almost 14000 websites in the region. It is fully acknowledged that at this stage, any biases within the collected seed URLs will be similarly reflected in the keyword database.

The final stage in this part of the research was to undertake a comparison between the information coming out of the keyword database and that which could be elicited regarding company activity from the same set of firms but using SIC derived activity. The keyword database program has the ability to choose one or more or indeed all of the datasets noted above. The most robust of the datasets in terms of quality of the data (when checked independently by sample) was the DNB company credit rating derived dataset numbering some 8518 records. It was thus possible to search for certain types of industry keywords known to be present in regional industry. Examples such as ‘engineering’ could be refined by Boolean operators to find such specialism’s as ‘precision engineering’.

A short test was set up using the 8518 records to find companies by (a) SIC and (b) SIC accompanying text description and then by submitting activity descriptors to the keyword database. It is known that some industrial activities are hard to find in an SIC based system, for example ‘motorsport’ (Pinch and Henry 1999) and a number of these were chosen for the test. The results of the comparison between the Keyword System and the SIC based system yielded Table 1 - Test of keyword frequency by number of firms.

**Table 1**. Test of keyword frequency by number of firms

| Word | Count by DNB 'Main Activity' | Count by any DNB 'SIC text' | Count by Keyword |
|---|---|---|---|
| Motorsport | 0 | 0 | 11 |
| Motor racing | 0 | 0 | 9 |
| Motorcycle | 3 | 24 | 38 |
| Automotive | 10 | 8 | 104 |
| Defence | 0 | 0 | 48 |
| Military | 2 | 0 | 25 |
| Offshore | 5 | 0 | 105 |
| Subsea | 1 | 0 | 15 |
| Yacht | 2 | 0 | 13 |
| Sail | 1 | 0 | 9 |
| boat | 6 | 8 | 36 |
| TOTALS | 30 | 40 | 413 |

Whilst this exercise is relatively simplistic compared with the more elegant and focussed route taken by for example (Hajlaoui and Boucher 2009) the conclusion here is that not surprisingly, a search system based on whole text searching in a closed set of websites is far superior for finding activity than by searching on SIC based activity alone. This can be seen in the right hand column where the number of firms found against each activity word is sometimes as much as an order of magnitude greater than the SIC based columns.

This result indicates that the basic methodology of finding firms engaged in some nominated activity using only the Web does work within limitations. The methodology does however require the user to input various supposed activities or groups of activities that may be prominent in the region under study. It does not of itself find clusters in the same way that a study based on location quotients does but on the other hand it does find activities that are often hidden to more conventional methods of discerning firm activity.

Various other searches of the Keyword database were tried exploring the industrial make up of the region and in particular that associated with a sub-sea 'cluster'. It had been known from other studies (Siedlock and Andriani 2006) that this type of activity was prominent amongst the region's offshore and engineering firms but an SIC based search yielded little evidence of this. This is because the firms involved do not put 'sub-sea' down as their primary or even secondary activity, they regard themselves as being in various kinds of engineering. However many of these same firms have the word 'sub-sea' somewhere on their website and this is picked up in the keyword database. Additionally such searches do not limit the scope to companies as often educational establishments, quality assurance organisations, specialist law firms and a whole raft of supporting organisations also appear and this sort of thing helps to gain a wider understanding of the potential cluster effect.

## 4 Finding Industrial Networks

Here with regard to finding linkages between organisations either as a buyer, or as a supplier or for knowledge exchange we have made the assumptions that a link exists if, on a company website:

(i) a firm puts a reference to another organisation on its own website as a clickable link (found by a short search program)
(ii) there are visual clues to other organisations on a company website e.g a client list (carried out by inspection)
(iii) an external firm references the URL being looked at, so called in-links (using an in-link finder such as www.linkpopularity.com)

For comparison purposes we used the list of firms noted above in the 'subsea' cluster. This corpus was derived from the first part of the research and comprised firms and other organisations with the words 'sub-sea' somewhere on the firm's website and in addition the firms identified by (Siedlock and Andriani 2006) in their snowball sampling based study were also added.

The only limitation applied was that as we were seeking evidence of activity on a closed geographical basis then any connection found should be to another regionally based entity.

When all these three types of link were merged, for each of the organisations in the subsea cohort it became possible to discern some interesting patterns of connectivity within the cohort. For the combined dataset of 282 nodes (organisations), website searching found 394 intraregional links.

## 5 Analysis of Firm Linkages

The linkages have been drawn and analysed using UCINET (Borgatti, Everett and Freeman 2005) and also VISONE ( www.visone.info ). These are simple links and say nothing about strength although direction can be inferred. The overall network of the subsea cohort is as shown in Figure 1 and it is of course possible to carry out all manner of analyses regarding the metrics of the network.

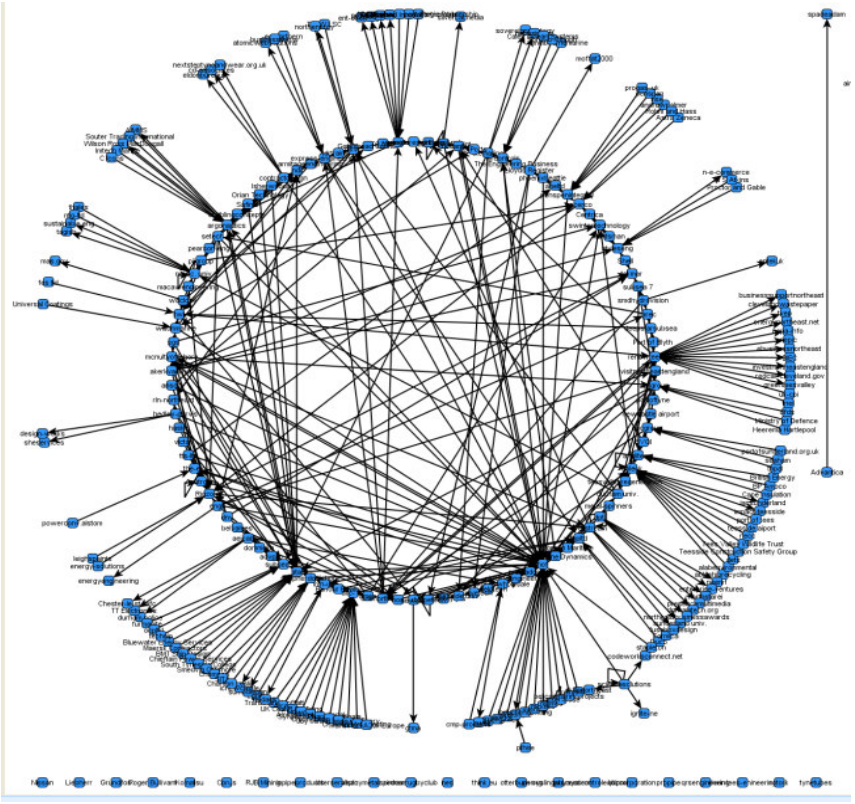

**Fig. 1.** Subsea Network in the N.E. of England (circular representation)

## 6 Overlaps Between Clusters

In the work looking at the activity of clustered groups of firms it was observed that many firms had 'membership' of more than a single grouping i.e. they used their expertise to address a variety of market segments. To test the extent to which such overlaps took place a test was set up to track the activities of a nominated group with respect to its activities in one or more other clusters. The groups chosen were Marine

Engineering, Environmental technologies, Offshore Oil and Gas, Subsea and Defence as these groups had shown up strongly in the wider study of the region as distinct groupings although whether they could each be regarded as a Porter type cluster is open to discussion.
As a result of this exercise it was possible to draw Table 2 which shows the extent to which firms in main or primary cluster are also engaged in other clusters.

**Table 2.** Degree of Involvement by Firms in other Sectors

| | Main Sector | | | | | |
|---|---|---|---|---|---|---|
| | Marine | Envir | O'shore | Subsea | Defence | Totals |
| Populations | 199 | 386 | 344 | 283 | 290 | 1502 |
| Firms in other sectors | 49 | 45 | 40 | 11 | 123 | 268 |
| % 'hits' | 25% | 12% | 12% | 4% | 42% | 18% |
| Sector by sector detail: | | | | | | |
| Marine | | 22 | 4 | 1 | 30 | |
| Env | 1 | | 1 | 3 | 29 | |
| Offshore | 29 | 15 | | 6 | 38 | |
| Subsea | 18 | 8 | 31 | | 26 | |
| Defence | 1 | 0 | 4 | 1 | | |
| TOTALS | 49 | 45 | 40 | 11 | 123 | 268 |

Of the 1502 firms in Table 2, it was noted that 140 had a presence in another grouping in addition to their 'main' activity, 47 had a presence in 2, 10 in 3 groups and 1 firm had a presence in 4 in addition to their own. When we looked in detail at the capability of these organisations it was possible to discern from their websites that they have a combined capability in supporting engineering manufacture of medium and large sized artefacts such as for example ships, armaments, rigs and subsea ploughs.

## 7 Discussion and Concluding Remarks

At this stage of the research we are firstly seeking to establish a methodology rather than a comprehensive picture of industry in a specified locale. The URLs used in this research are but a sample of commerce and industry in the region under scrutiny and a complete or near complete set would be required to have complete confidence that we would have a comprehensive database of all the activities being undertaken. Secondly, initial attempts to use the keyword database for discerning clusters free from the known constraints of an SIC based system clearly have some limitations. What the keyword database does is find activity by firm and in considerable detail. It does not of itself find clusters although in conjunction with other methods as noted earlier it may help to find the presence of firms and supporting organisations to feed into a database of candidate firms for more conventional cluster analyses.

In the research into linkages it is fully acknowledged that to discern networks we are using the presence of links on (or into) a site as evidence of some form of collaboration or of economic dependency or some measure of esteem. Whilst this may be true for some firms, for many others with all manner of networks of their own, no

such relationships at all are inferred on their own websites. It was observed that the system works better for smaller firms as their websites tend to have more regional linkages. For example the corporate website of a very large offshore operator who had commissioned a multi-million dollar rig project in the region would be unlikely to have links back to the yard and main contractors who built the rig. Conversely the many smaller companies who had contributed to that project, even in a small way, put a link on their sites citing the corporate as a prestigious client. Many links that might have been gleaned by snowball sampling or other practical means are thus missed by the web based methods described here. However on-the-ground sampling methods are far from perfect and it is perhaps salutary to note that the internet derived links were greater in number than those from the practical observations cited. The links as shown in Figure 1 are also useful for determining the presence of 'key nodes' i.e. those organisations that are 'important' in network terms as they are gatekeepers of knowledge or of trade within the network as shown. Further, the phenomena known as 'the strength of weak links' becomes evident (Granovetter 1973) and that is the ability of one dense network to be joined to another remote network by a small number of individual links but the important part played by these 'connectors' was out of proportion to the actual number of occurrences.

A key question when looking at web derived networks is do such gatekeepers appear because they put lots of links on their sites and they are then similarly reinforced by external sites through in-links and thus are they really economically important in the cluster of activity? Comparison with networks derived by sampling would indicate about a 50% match in terms of indentifying 'important' nodes. However as with the first part of the research in finding firms, the web based system found networks and sub-nets not found by manual sampling and vice versa. The two methods in combination therefore helped to discern a much richer network of industrial interactions than either method in isolation.

Although the example of the subsea cluster given here has been used partly because of the opportunity for comparison it should be noted that in many sectors it is almost impossible to carry out any form of a snowball sampling exercise because of suspicion over commercial confidentiality from individual firms. In such cases the 'hands off' methodology shown here may be the only option for finding the presence of interconnectedness.

Spidering whole text from organisational websites frees the researcher from the constraints imposed by an SIC based system and one of the advantages is that it is possible to look at firms' capabilities or competences rather than looking simply by aggregated sector or product. SIC based systems cannot easily cater for this type of search as they are based only on a main and a small number of secondary activities. In the work on overlaps therefore it was possible to find firms that could sell into different sectors and markets because of the range of their competences as noted on their own websites. In the region studied and with the sample database used some 198 organisations were identified that when examined had competencies in engineering design and manufacture for medium complex engineering artefacts often designed for harsh environments. This outcome probably has much to do with the history of the North East of England and the region's traditions in shipbuilding, mining machinery and offshore rig building. The same skills and competencies however are now being deployed in advanced subsea operations and offshore wind technologies. This plays

well with the notion of 'related variety' (Boschma 2008) in that competences established either as part of a cluster or trading network in one sector can be transposed to new markets and technologies and the ability to see this amongst the admittedly engineering related groups used was one of the unexpected results of this research

This has been an attempt at using the web to try and discern the presence of agglomerations of connected industrial activity within a regional boundary. At the present stage of research it is considered that the methodology developed shows promise but that it clearly needs more work to improve the confidence in the results found. As web tools develop and as commerce and industry generally further embrace web based promotion and trading such progress will help to strengthen the possibilities for developing a robust web based methodology as outlined above.

## References

Barabasi A. B.: Linked - How everything is connected to Everything Else and what it means for Business, Science and Everyday Life. Penguin Books (2003).

Borgatti S.P., Everett M.G., Freeman L.C.: Ucinet for Windows: Software for Social Network Analysis. Harvard, MA: Analytic Technologies. V.6.181 (2005)

Boschma R.: Constructing Regional Advantage: related variety and regional innovation policy. Report for the Dutch Scientific Council for Government Policy. University of Utrecht, (2008)

Feldman M. P., Francis J., Bercovitz J.: Creating a Cluster While Building a Firm: Entrepreneurs and the Formation of Industrial Clusters. Regional Studies, Vol.39.1 pp. 131 and 139. (2005)

Granovetter M.S.: The Strength of Weak Ties. American Journal of Sociology 78, pp. 1360-1380 (1973).

Hajlaoui K., Boucher X.: Neural Network Based Text Mining to Discover Enterprise Networks. Proceedings of th 13th IFAC Symposium on Information Control Problems in Manufacturing. Moscow, (2009)

Pinch S., Henry N.: Paul Krugman's Geographical Economics, Industrial Clustering and the British Motor Sport Industry, Regional Studies, 33.9, pp. 815-827 (1999).

Porter M. E.: On Competition. Harvard Business School Press. (1998a)

Porter,M.E.: Location, Clusters and the 'New' Microeconomics of Competition, Business Economics, 33, 1, pp. 7-17. (1998b)

Porter M.E.: Clusters and the New Economics of Competitiveness, Harvard Business Review, December, pp. 77-90. (1998c).

Porter M. E.: Locations, Clusters and Company Strategy, in Clark, G.L., Feldman, M. and Gertler, M. (Eds) Handbook of Economic Geography, Oxford: Oxford University Press, pp. 253-274. (2000).

Siedlock F. and Andriani P.: The emergence of the sub-sea technology clusters in the North-East of England: Some evolutionary considerations and implications for cluster policy. Conference on 'The Organising Society'. EGOS Bergen. www.egosnet.org (2006).

Van den Berg, L., Braun, E. and van Winden W.: Growth Clusters in European Cities: An Integral Approach, Urban Studies, 38, 1, pp. 187 (2001)