Titre: Title:	Competence maps using agglomerative hierarchical clustering
Auteurs: Authors:	Ahmad Barirani, Bruno Agard, & Catherine Beaudry
Date:	2013
Type:	Article de revue / Article
Référence: Citation:	Barirani, A., Agard, B., & Beaudry, C. (2013). Competence maps using agglomerative hierarchical clustering. Journal of Intelligent Manufacturing, 24(2), 373-384. https://doi.org/10.1007/s10845-011-0600-y

URL de PolyPublie: PolyPublie URL:	https://publications.polymtl.ca/2312/
Version:	Version finale avant publication / Accepted version Révisé par les pairs / Refereed
Conditions d'utilisation: Terms of Use:	CC BY-NC-ND

Titre de la revue: Journal Title:	Journal of Intelligent Manufacturing (vol. 24, no. 2)
Maison d'édition: Publisher:	Springer
URL officiel: Official URL:	https://doi.org/10.1007/s10845-011-0600-y
Mention légale: Legal notice:	This is a post-peer-review, pre-copyedit version of an article published in Journal of Intelligent Manufacturing (vol. 24, no. 2) . The final authenticated version is available online at: https://doi.org/10.1007/s10845-011-0600-y

Competence Maps Using Agglomerative Hierarchical Clustering

Ahmad Barirani, Bruno Agard and Catherine Beaudry

Département de Mathématiques et de Génie Industriel

École Polytechnique de Montréal

C.P. 6079, succ. Centre-ville, Montréal (Québec), H3C 3A7, Canada

ahmad.barirani@polymtl.ca; bruno.agard@polymtl.ca; catherine.beaudry@polymtl.ca

Abstract: Knowledge management from a strategic planning point of view often requires having an accurate understanding of a firm's or a nation's competences in a given technological discipline. Knowledge maps have been used for the purpose of discovering the location, ownership and value of intellectual assets. The purpose of this article is to develop a new method for assessing national and firm-level competences in a given technological discipline. To achieve this goal, we draw a competence map by applying agglomerative hierarchical clustering (AHC) on a sample of patents. Considering the top levels of the resulting dendrogram, each cluster represents one of the technological branches of nanotechnology and its children branches are those that are most technologically proximate. We also assign a label to each branch by extracting the most relevant words found in each of them. From the information about patents inventors' cities, we are able to identify where the largest invention communities are located. Finally, we use information regarding patent assignees and identify the most productive firms. We apply our method to the case of the emerging and multidisciplinary Canadian nanotechnology industry.

***Keywords**: knowledge mapping, innovation, citation networks analysis, data mining, agglomerative hierarchical clustering, vector space model, nanotechnology.*

1 Introduction

Globalization is marked by a hyper-competitive economic landscape (Westphal *et al.*, 2010). Advances in industrial engineering and logistics have given the possibility for advanced countries to offshore their manufacturing activities to developing countries that offer cheaper labor wages. After a long period of rationalization, the same advanced countries are now facing the situation where those once developing countries are catching-up the technological gap (Albayrak and Erensal, 2009). In fact, emerging countries are suddenly leaders in certain high technology fields.

This new reality has an important impact on the industrial organization of advanced countries that are now forced to be more innovative if they want to benefit from economic growth. It has become vital for advanced countries to put in place institutions and policies that foster the development of their high technology industries. Innovation can be boosted when there are interactions among different technological fields (Taskin and Adali, 2004). Among multidisciplinary fields, one of the most promising high technology sectors is that of nanotechnology. Nanotechnology is often thought as a field that can have revolutionary applications in a wide range of industries. All advanced countries agree on the importance of this new field in the development of their economy. They have also put in place policies that would help develop their knowledge and competence levels in this promising area.

Innovative activities must however be performed in a context of resource scarcity. Even though advanced countries have greater access to resources compared to developing or emerging countries, it is impossible to explore and exploit all the technological paths that are available to them. Firms, organizations and countries must take their technological strengths and weaknesses into consideration when making strategic decisions about the directions they are willing to take. An important step in finding the strengths and weaknesses at national level consist in drawing a technological competence map of the country. In such contexts, the access and integration of information systems into the decision making process is crucial (Hsu *et al.*, 1994).

In this article, we propose a new method of assessing technological competences. Our method consists in developing a competence map of the Canadian nanotechnology industry by applying agglomerative hierarchical cluster analysis on a sample of patents obtained between 2005 and 2008. Nanotechnology has been selected because it is a recent, relatively well defined, active and still moving domain. We will be able to show the main branches of Canadian competences in nanotechnology and identify the most active regions and firms for each of these branches. The remainder of the article is organized as follows: the next section will provide some theoretical framework regarding strategic aspects of knowledge management and knowledge mapping as well as some elements regarding different methods used for knowledge mapping. We present two methods for measuring similarity between documents: citation network analysis and text mining. Then we provide a description of cluster analysis as a way to ordinate documents and techniques available for assigning labels to the groups of documents. The article then presents our methodology for mapping Canadian competences in nanotechnology. Finally, we will analyze the results of our study and make parallels with strategic management theory described.

2 State of the art

2.1 Knowledge management

The strategic managers' tasks often consist of performing an assessment of the organization's resources and *core competences* and of defining a strategic plan that will reinforce those competences (Barney, 1991; Prahalad and Hamel, 1990; Amin and Cohendet, 2004). In today's knowledge economy, the organization's stock of knowledge or intellectual capital is viewed as a strategic resource that constitutes its most valuable asset (Nahapiet and Ghoshal, 1998). This is the knowledge-based view of the firm in which organizations succeed because they have knowledge that is valuable, rare and inimitable (Grant, 1996). Another phenomenon which organizations are facing in the knowledge economy is constant change in their environment. In this regard, organizations need to have *dynamic capabilities* to reinvent themselves in the face of rapidly changing environment (Teece *et al*., 1997). They need to put in place processes that enable them to change their routines, products and markets over time.

This is part of the evolutionary economic perspective which studies the impact of initial technological decision on future directions (Nelson and Winter, 1982). In this regard, knowledge creation and diffusion is a *path dependent* process (David, 1985). Technologies that are developed and adopted at a certain point in time will shape the technological choices that are made at a later time. In other words, what organizations learn is always bound to what they have learned in the past (Cohen and Levinthal, 1990). It also follows from this line of thought that organizations can be trapped in *technological lockin* when they are unable to change their routines because they have invested too heavily in one technological branch (Arthur, 1989). Changing their technological paths becomes too cumbersome as these organizations are plagued with inertia. Taking into perspective the importance of intellectual capital and the path dependent nature of knowledge, it becomes vital for organizations to be self-aware of their core competences and of the opportunities that they have to absorb complementary knowledge (Feldman, 1994). It should be noted that knowledge is information in a specific context. In other words, it is useful only in that specific context. A firm's routines and best practices can change when the context changes (Chryssolouris *et al*., 2008; Wijnhoven, 2008).

One way to measure intellectual capital is through the analysis of patenting activity (Basberg, 1987). Patent databases have been used to derive the state of development in specific technologies (Duflou and Verhaegen, 2004). Patents are indications of research and development efforts endeavored by its inventors and assignees. They can therefore be counted as technological competence owned by the organization. Because patents must be novel and specific, they are also indicators of technological change. Organizations that are able to patent at a higher rate than others therefore show a capacity to bring technological changes to their industry. Certain organizations perform better than others when it comes to patenting. Larger firms that dispose of a greater quantity and diversity of resources are better equipped to patent than other. More important, they are able to patent in a much broader set of technological fields because their diverse knowledge-base allows them to innovate across many areas (Cantner & Graf, 2006; Boschma & ter Wal, 2005; Morrison 2008).

2.2 Knowledge Mapping

Börner *et al*. (2003) provide a thorough literature review regarding knowledge mapping. Knowledge mapping consists in gathering, analyzing and synthesizing bibliographical data in order to discover the location, ownership and value of intellectual assets. Knowledge maps can be used for the identification of scientific and technological know-how at firm, university or national level. Knowledge maps can be used for indicating current technological trends and can be helpful in forecasting future technological developments. Finally, knowledge maps can be used to find new opportunities to explore in emerging technological disciplines.

The first step in knowledge mapping usually consists in extracting a set of documents (articles or patents) from a bibliographical database (such as ISI-Thomson, Scopus or USPTO). Most studies use a *Boolean keyword-based document retrieval* method, i.e. documents that contain specific keywords are retrieved from the database for analysis. The process then consists in selecting *similarity attributes* for the documents. The two most popular attributes are citations and words, i.e. documents are similar if they cite the same sources or if they use the same words in their description. Based on the similarity attributes, documents are then grouped together, usually through *cluster analysis* or *dimension reduction*. Each of the resulting groups represents a knowledge branch to which a label is assigned by analyzing the content of the documents it contains. By analyzing other information associated with the documents, such as the authors, address or affiliations, it is possible to see who owns the intellectual capital and where the inventor communities reside. Interdependence between branches can be found by aggregating the citations made by the documents contained in each branch. For example, if many articles from branch A cite articles from branch B, then it can be said that branch A is technologically dependent upon branch B.

2.3 Measuring similarity through citation network analysis

In order to consider citation network analysis for similarity computing purposes, we will introduce some key concepts related to network theory. A network is defined by a pair of sets *G* = {*P*, *E*} where *P* is a set of *N* nodes P_1, P_2,…, P_n and *E* is a set of *m* edges that connect two nodes in *P* (Wasserman and Fraust, 1994). Each node has a *degree distribution* defined by the number of edges it shares with other nodes in the network. The number of edges that separate two nodes is called the *geodesic distance*. The shortest path is the smallest geodesic distance between two nodes. *Betweenness centrality*, for a node *i*, is therefore defined by

$$C_B(i) = {}_{j\quad k\quad i}\ \frac{{}_{jk}(i)}{{}_{jk}}$$

where σ_{jk} is the shortest path between nodes *j* and *k*, and $\sigma_{jk}(i)$ is the number of shortest paths between nodes *j* and *k* that pass through node *i*. Betweenness centrality is often an indication that a node is connecting two groups of nodes that would otherwise be disconnected (Grannoveter, 1973; Burt, 1992). These central nodes therefore are agents that imply a certain similarity between the groups of nodes that they help to move closer. For any given node *i*, the *clustering coefficient* C_i is defined by

$$C_i = \frac{2E_i}{K_i(K_i \quad 1)}$$

where E_i represents the number of edges between K_i nodes that are linked to node *i*. This metric shows the degree with which nodes connected to *i* are also connected to each other. A *clique* is a group of nodes that are all interconnected. A *community* is a network subgroup of nodes that are densely connected (Newman and Girvan, 2004). In both cliques and communities, average clustering coefficients are high since nodes tend to be interconnected. The presence of a clique or a community is therefore an indication of affinity and similarity between the nodes.

A network *component* is a subnetwork where at least one path exists between all nodes of the subnetwork. Disconnected components usually indicate that there is little similarity between nodes in each component.

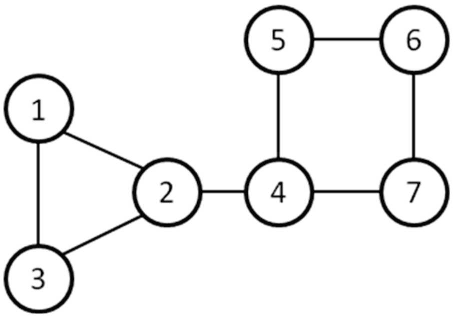

Figure 1: Network with 7 nodes and 8 edges.

Figure 1 is an example of a small network. Nodes 1, 2 and 3 are part of a clique and we can say that there are two communities in the network: one composed of nodes 1, 2 and 3 and the other composed of nodes 4, 5, 6 and 7. The network in figure 1 is composed of only one component since all nodes can be reached from any other node in the network. If node 2 and 4 were not connected, then the network would have contained two components: one composed of nodes 1, 2 and 3 and another composed of nodes 4, 5, 6 and 7. Nodes 1 and 3 have a clustering coefficient of 1.0 while node 2 has a clustering coefficient of 1/3. Betweenness centrality for nodes 2 and 4 are equal to 8 and 9, since nodes 1 and 3 must go through them to connect to nodes 5, 6 or 7. In this example, we can say that nodes 1, 2 and 3 are very similar. Also, node 4, 5, 6 and 7 are also similar to each other but at a lower degree.

Many kinds of networks have been observed in nature. Biological, social, electrical and hypertext networks are among some of the examples (Albert and Barabasi, 2002). Citation networks are networks where nodes are defined by documents and where edges are defined by the citations that connect the documents together. Citation networks are often categorized under directed networks, meaning that the relationship between the two nodes is unidirectional. In this regard, citation networks express interdependence and knowledge flows between documents (Small, 1999).

Small (1999) uses citation networks as a way to measure similarity in bibliographical data. Areas of high intercitation density then become indications of scientific activity around a certain subject. Bassecoulard *et al*. (2006) measure similarity and interdependence between nanoscience branches by using citation flows. From a seed of articles obtained by Boolean keyword-based retrieval, they build a larger sample by retrieving articles that often cite and get cited by the seed.

2.4 Cluster analysis

Cluster analysis is a data mining technique that consists in grouping a set of observations in a way such that similar elements are placed in the same group, called cluster (Berry and Linoff, 2004). These techniques are classified under *unsupervised learning techniques*. There are different types of clustering methods. All of the methods based on similarity require a measure of distance between two elements. The *Euclidean distance* between two documents q and p is a very popular metric that is computed by the following equation:

$$d_{q,p} = \sqrt{\sum\nolimits_{i} (q_i - p_i)^2}$$

where q_i and p_i are the attribute i's values for documents p and q respectively. Other metrics such as the cosine or *dice similarity* can be used for the same purpose. The goal of a clustering algorithm is to maximize intercluster distance while minimizing intracluster distance (Manning *et al*., 2008).

Clustering can be used to solve a variety of problems (Malakooti & Raman, 2000). Cluster analysis can be used in the customer support and relationship management industry (Berry and Linoff, 2004). Chen *et al*. (2007) use cluster analysis to perform customer segmentation aimed at improving customer retention in the telecommunication industry. Choudhary *et al*. (2009) provide a thorough review of clustering techniques used to solve manufacturing problems such as defect analysis, system rule generation, yield improvement and process optimization. Given the general purpose of unsupervised learning methods, cluster analysis has also been used for generating knowledge maps based on bibliographical data. The following two sub-sections provide a literature review of some of the most common techniques used in this area.

2.4.1 Partitional clustering

Partitional clustering techniques, such as k-means, group elements into a fixed (k) number of segments. The user can predefine or, after a few trials, deduct this number. The partitioning process starts by assigning one element to each cluster. This element will become the cluster's core. Remaining elements are then assigned to a cluster according to their distance with its core. At the next iteration, a new core is selected for each cluster from the elements that are assigned to it. Remaining elements are again assigned to the cluster having the less distant core (Berry and Linoff, 2004). The process stops after a maximum number of iterations or when a local optimum is found. Bassecoulard *et al*. (2006) use a variation of *k*-means clustering on citation networks to group articles into 7 broad scientific branches. By using affiliation data regarding articles, the authors were able to identify specialization levels of major countries in each branch of nanoscience. In addition, the authors show the interdependence between branches by analyzing citation flows at the cluster level. Kim *et al*. (2008) apply *k*-means clustering on a keyword vector space obtained from a sample of patents. Each formed cluster represents a technological branch. Branches are then linked together based on the co-occurrence of keywords in the clusters. By finding the patents that were filed earliest in each cluster and by linking clusters through citation analysis, the authors build a timeline showing when technological branches where introduced and to what technological branches they have led to.

2.4.2 Hierarchical clustering

Hierarchical clustering classifies observations under a tree structure after a number of iterations (Berry and Linoff, 2004). Clustering can be done by *agglomeration* (bottom-up: HAC, CURE) or by *division* (top-down: DIANE, BIRCH). Agglomerative methods initially assign each element to a segment. In e iteration, clusters that are similar are merged to form a larger cluster. The process stops when there is only one cluster left. Divisive methods in contrast start with one cluster that contains all the elements. In each iteration, clusters are split in a way that maximizes the distance between elements of one cluster and the other. The process stops when all segments constitute of only one element.

Newman and Girvan (2004) use hierarchical clustering for community detection in networks. They use network betweenness centrality as an indication of community boundaries. They place the most central nodes at the top of the dendrogram and the less central nodes at the bottom. Combined with citation networks analysis, hierarchical clustering also has the advantage of showing the relationship between scientific branches (Wallace and Gingras, 2009). Documents that cite sources common to lower-level clusters that do not cite common sources will more likely be positioned on higher levels of the dendrogram. They therefore connect those clusters and represent a broader branch. Tseng *et al*. (2007) have developed a hierarchical topic map by performing a multi-stage clustering method. They first cluster a large set of patents into small clusters based on their vector space similarity. At the next stage, these small clusters are then regrouped together based again on their vector space similarity.

2.4.3 Cluster Labeling

Weiss *et al*. (2005) list different methods for labeling clusters. *Feature selection* techniques are often applied in order to select a relevant set of words from a larger list. A simple approach in labeling clusters is to select the most frequent words in each cluster. Term ranking methods such as the *tf-idf* metric can also be used for the purpose of feature selection. The following procedure is usually applied in order to compute the *tf-idf* for terms appearing in a set of documents (Manning *et al*., 2008).

1. **Tokenising**: for every document in the sample, sentences are broken into single words. This leads to a vector of words representing each document.
2. **Stopwords removing**: common words (such as *the*, *and*, *or*, etc.) are removed for each vector representing a document.
3. **Weighting terms**: here the relative frequencies with which stemmed words appear in a single document with respect to the whole sample are computed. The *tf-idf* rank is the most common method used for this purpose. To compute the *tf-idf* rank of a term *i* in a document *j*, we first need to compute the *term's frequency* in the following way:

$$tf_{i,j} = \frac{n_{i,j}}{\sum_k n_{k,j}}$$

where $n_{i,j}$ is the number of occurrence of the term in document d_j and the denominator is the sum of the occurrences of all terms in document d_j. Then, we need to compute the *inverse document frequency* by using the following equation:

$$idf_i = log\frac{|D|}{|\{j:t_i \in d_j\}|}$$

where $|D|$ is the total number of documents in the sample and $|\{j: t_i \in d_j\}|$, called *document frequency*, is the number of documents in which the term appears. The *tf-idf* is then computed as follow:

$$tf\text{-}idf_{i,j} = tf_{i,j} \times idf_i.$$

Resulting from this definition, the *tf-idf* will be a) highest for terms occurring many times within a small number of documents, b) lower for terms occurring fewer times in a document or occurs in many documents, and c) lowest for terms appearing in virtually all documents (Manning *et al.*, 2008). Therefore, terms that have higher *tf-idf* scores can be selected as labels representing each document. This method can be extended to clusters where terms are taken from the documents that are assigned to each cluster (Weiss *et al.*, 2005).

Tseng *et al*. (2007) perform cluster labeling in the following manner. First, they find the most frequent words used by patents in each cluster from which they remove words that also frequently appear in other clusters. They then use an automatic Wordnet-lookup algorithm to classify those words into broad technological fields such as *material*, *chemistry* and *biomedecine*. Sometimes, labeling is performed manually. For example, if the most frequent word in a cluster is *biology*, then the user can assign that topic to the cluster.

3 Methodology

The method proposed in this article is based on five steps (figure 2). In order to simplify the reader's comprehension each step will be explained throw an example in building a map of Canadian competences in nanotechnology based on patent citation networks.

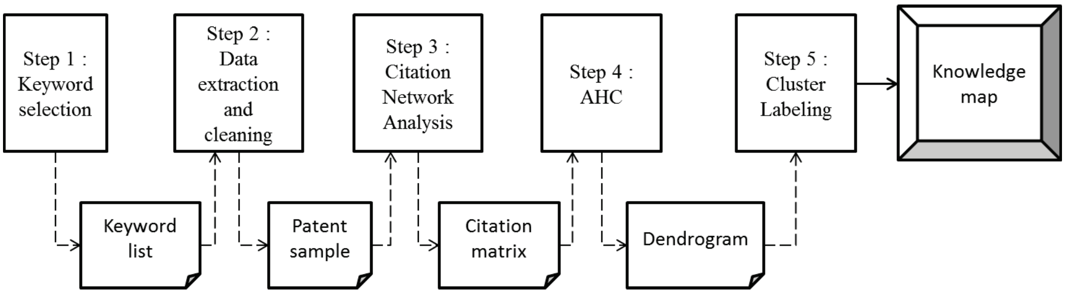
Step 1 : Keyword selection
Step 2 : Data extraction and cleaning
Step 3 : Citation Network Analysis
Step 4 : AHC
Step 5 : Cluster Labeling
Knowledge map
Keyword list
Patent sample
Citation matrix
Dendrogram

Figure 2: Methodology steps

Step 1: Keyword selection

We first need a set of nanotechnology related keywords. These keywords are obtained from bibliographic studies on nanotechnologies (Alencar *et al.*, 2007; Fitzgibbons and McNiven, 2006; Mogoutov and Kahane, 2007; Porter *et al.*, 2008; Schmoch *et al.*, 2003; Zitt and Bassecoulard, 2006). These studies, altogether, use more than 596 distinct keywords in their definition of nanotechnology. Yet, only 21 of them appear in more than one study. Therefore, we can see that there is great disparity in what these authors define as being nanotechnology-related keywords. In order to select significant keywords that represent the core of nanotechnology patents, we will select keywords that are used in more than one of the studies to form a query that is run on the United States Patent and Trademark Office database (USPTO, 2009). This method can be seen as an approximation to *tf-idf* weighting of keyword significance. Other weighting and indexing methods will be considered in future works.

Step 2: Data extraction and cleaning

All patents that contain one of the keywords and that have been granted to Canadian firms or for which one of the inventors resides in Canada are retrieved from the USPTO database. For the reminder of the article, these will be referred to as *Canadian patents*. For each patent, data about the title, abstract, application and granted date, number of claims, references, citations, as well as the name, city and country of inventors and firms are extracted. We will refer to the patents that are cited by our Canadian patents by *cited patents*. The resulting sample is then cleaned of incomplete entries and different representation of the same assignee names (ex: Nortel and Nortel Networks are the same assignee). Finally, suburban areas are associated to their metropolitan areas (for instance, Laval is associated to Montreal's metropolitan area).

Step 3: Citation Network Analysis

The third stage of our study consists in building the citation network from our sample of Canadian nanotechnology patents. In our citation network, the nodes are the Canadian patents in our sample and the patents that are cited by them, and the edges are defined by the citation relationship between Canadian patents and those that they cite. We use the *open source* software application NodeXL (CodePlex, 2011) for this step of our study. From the resulting network, we select the largest component for the next step in our analysis. This is a necessary measure given the fact that we use AHC. Since we use the co-citation as a way to measure similarity, it is unavoidable that AHC groups two disconnected network components at a certain point in the process. In such cases, the AHC will perform an arbitrary merger of the two components, which will lead to incorrect representations of technological fields' hierarchies. By selecting the largest network component, we are certain that cluster mergers always involve a certain level of similarity in patent co-citations. Another advantage of working with the largest citation network component resides in that it acts as a natural cleaning process on the patents obtained by Boolean keyword-based retrieval. In fact, this retrieval method is bound to precision and recall issues, i.e. that the retrieval process will always miss some of the relevant documents and will add some undesirable documents to the retrieved sample. Removing patents that are not part of the largest citation network will rid us of some irrelevant patents that figure in our sample. However, this method has the disadvantage of discarding, from the competence map, relevant nanotechnology patents that are not connected to the main network component. This is a limitation imposed by the choice of AHC as a method for competence mapping. **Step 4: Hierarchical clustering**

In the fourth step of our method, we first build the citation matrix used for cluster analysis. This matrix will have rows representing Canadian patents from the largest component and column representing all the cited patents. In order to reduce the size of the attribute set (i.e. cited patents), we will only consider patents that have been cited by at least two Canadian patents. This is natural since patents that have been cited by only one patent do not contribute to the similarity of that patent with other patents. The citation matrix will be filled with 1s when a Canadian patent in the rows cites one of the cited patents in the columns and with 0s otherwise.

We then perform the actual AHC on the citation matrix. We will use the open source software application RapidMiner (Rapid-I, 2011) for this purpose. We will use cosine similarity as a way to measure patent similarity and the *average linkage* method of merging clusters together. Cosine similarity between Canadian patents *A* and *B* represents whether patent *A* and *B* cite the same patents. Average linkage means that clusters are merged together based on the average similarity of the patents they contain. Proceeding in this way has the advantage of merging clusters based on their overall citation patterns and will be helpful in measuring interrelatedness between different branches of the Canadian nanotechnology competences. From the dendrogram resulting from the AHC process, we select the clusters at the top levels to build our competence map.

Step 5: Cluster labeling

Our final step consists in finding labels for the clusters that are at the lower level of the competence map. By merging patent titles for each cluster, we build a vector space representing the *tf-idf* rank of the words appearing in each cluster. We then sort the words based on their *tf-idf* rank and select the top five words as labels for each cluster. As a result, clusters are represented by the words that they most frequently contain relative to other clusters.

4 Results and Analysis

This section will show detailed results of the methodology and final analysis of the knowledge map.

Step 1: Keyword selection

The first column in Table 1 shows the keywords selected for our study and the number of patents our extraction process has provided in December 2009. As described in section 3, these keywords have been used at least twice in a collection of bibliographic studies regarding nanotechnologies.

Step 2: Data extraction and cleaning

Data extraction was performed using PatentBot, a software application developed internally by our team. The second column in Table 1 shows the number of patents our extraction process has provided in December 2009. From these 8,076 patents, 5,811 have been selected after cleaning was performed on incomplete patent documents. From these patents, we have selected those that were obtained during the years 2005 to 2008. This gives us a more accurate map of current Canadian competences in nanotechnology. Our sample contains 1,697 Canadian nanotechnology patents granted between 2005 and 2008.

Table 1: Nanotechnology keywords

	Number of patents extracted
nano*	4 568
atom* force microscop*	88
biosensor	231
mesoporous material*	31
molecular beam pitaxy	95
molecular switch	25
nems	9
polymer composite*	379
polymer dna	10
polymer rna	3
quantum	1287
scanning probe microscop*	16
self assem*	219
supramolecular chemistry	18
tunnel* microscop*	2
photonic*	969
scanning prob*	41
single electron*	85

Step 3: citation network analysis

By analyzing the sample of patents obtained in the previous step, we find that the 1,697 Canadian patents obtained between 2005 and 2008 cite 22,017 distinct patents and the citation network is composed of a total of 36,961 citations. From the 22,017 distinct patents, only 6,712 (~30%) are cited more than once by the Canadian patents. The citation network has (1,697 Canadian patents + 22,017 cited patents =) 23,714 nodes and 36,961 edges, implying that it is expected to be relatively fragmented. In fact, when building the citation network (figure 3) with the help of NodeXL, we observe that the main network component is formed by 10,853 out of 23,714 nodes (~46%). Furthermore, only 691 (~41%) patents from our initial list of 1,697 Canadian patents are part of the main network component. The network is composed of 622 disconnected components, 484 of which contain only one Canadian patent. These are patents that a) are not cited by any of the Canadian patents and b) do not cite any of the other patents that have been cited by the Canadian patents. Although we cannot conclude that these 484 patents are false positives (that they have been extracted because containing ambiguous nanotechnology keywords), we cannot use them for the purpose of knowledge mapping with regards to our methodology. In fact, not having any citation in common with other Canadian patents, they will be at *infinite* distance of other patents or clusters. This will wrongfully place them at the top of the dendrogram which will result in a loss of precision in our technological hierarchy. The 3 largest components after the main component contain respectively 38, 26 and 22 Canadian patents. While these components are large enough to be treated as clusters, they suffer from the same issue than those 484 patents. Although we could apply AHC on each of those components, we cannot *situate* them with regards to the clusters found for the main component because no similarity in terms of co-citations exists between them.

Figure 3 shows the Canadian nanotechnology network's main component. Big-colored nodes represent Canadian patents and small-black nodes represent patents cited by the Canadian patents. Each color represents one of the clusters found during our AHC (4th) step. As we can see, the clustering process regroups patents that are situated in the same region in the network graph.

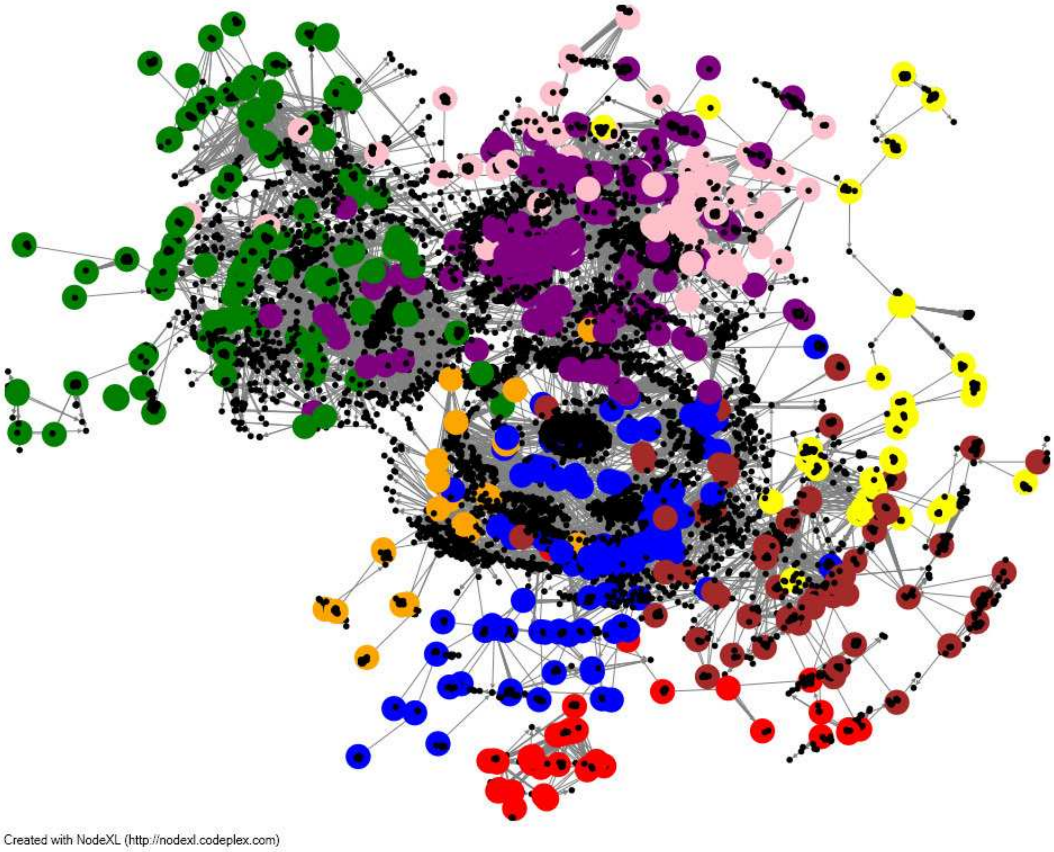

Figure 3: Canadian nanotechnology patents citation network's main component between years 2005 to 2008. Big-colored nodes represent Canadian patents. Small-black nodes represent patents cited by Canadian patents. Each color represents one cluster found by our AHC method. Since the two-dimensional representation of the network will place nodes that cite the same sources in the same region, nodes from the same cluster are also located in the same regions.

Step 4: Hierarchical clustering

From the citation network of the main component, we build a citation matrix of size 691 by 3,765 (this is the number of patents that are cited more than once by the 691 Canadian patents). By running an AHC on this matrix, we obtain the dendrogram shown on the right side of figure 4. As expected, the average linkage method offers a better hierarchical representation of the technological branches than the single linkage method (left side of figure 4) which has a stairway-like shape. This is due to the fact that single linkage, by merging clusters based on the most similar elements, will delay the merger of *outsider* patents to later steps in the linkage process. The competence map resulting from the selection of top-level clusters will show distinct technological branches separately but will embed them one into another instead of having a balanced tree of branches.

At the highest level of the dendrogram resulting from the average linkage method, the two top level clusters are at a distance of 1.57078. We then select all clusters that have a distance above 1.57 for our competence map, which gives us around 20 clusters, with the smallest clusters having more than 20 patents. This seems reasonable, given the fact that we need to have clusters large enough to be able to have meaningful labels for each of them. As shown in figure 5, each cluster is represented by a circle that is sized according to the number of patents it contains. Child clusters are drawn inside the parent cluster to represent the hierarchical dimension of clusters. Each cluster is also identified by the cluster ID provided by RapidMiner. This ID represents the iteration number in which the cluster was created. As we can see in figure 5, higher-level clusters have higher IDs because they are formed later in the clustering process.

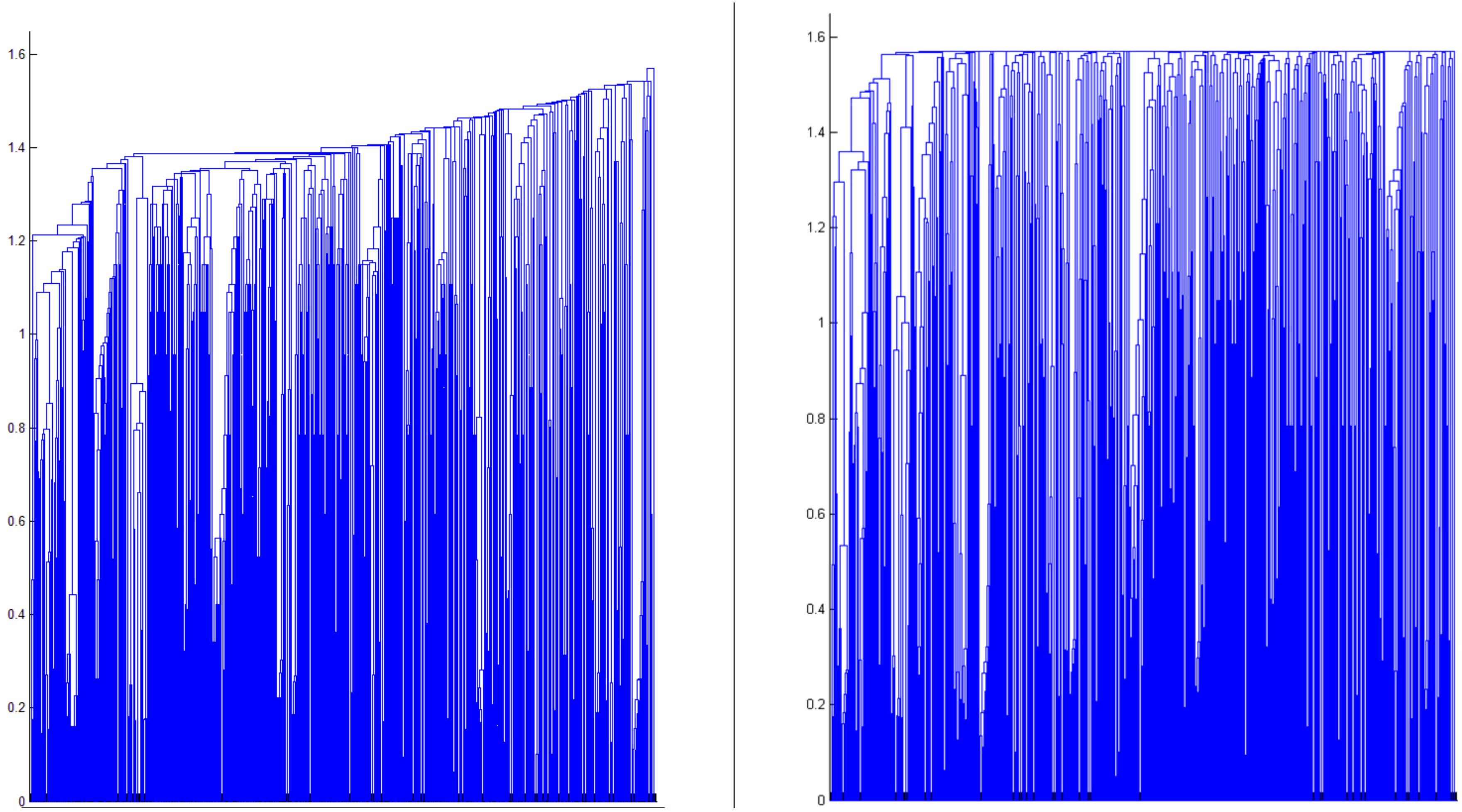

Figure 4: Dendrogram resulting from AHC using single linkage (left) and average linkage (right). (Plot using Matlab 2009b).

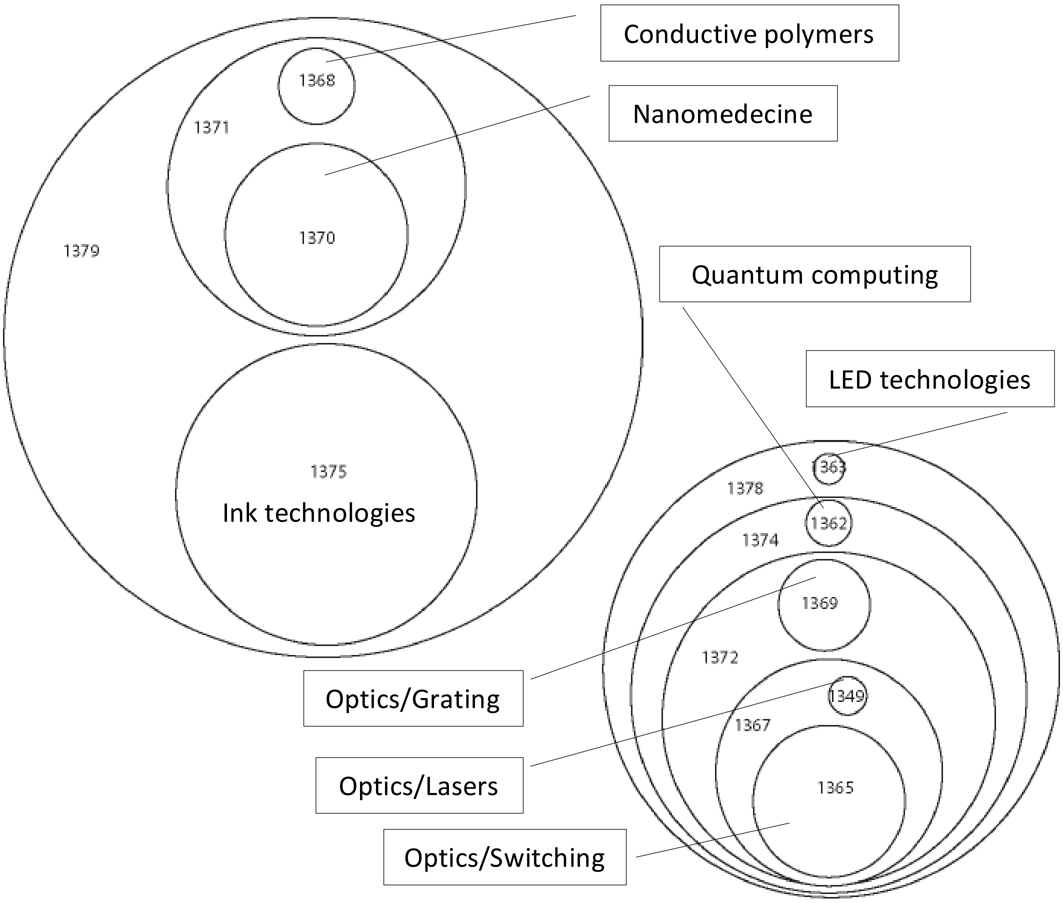

Figure 5: Competence map based on the main components of Canadian nanotechnology citation network.

Step 5: Cluster labeling

To label clusters, we merge the titles for the 8 clusters that are at the lowest levels of our competence map (clusters 1349, 1362, 1363, 1365, 1368, 1369, 1370 and 1375) and select the highest *tf-idf* ranked terms appearing in the merged titles of each cluster. We also search for the top three patent holders and active cities in each cluster. The results are shown in table 2. As we can see, *Xerox Corporation*, *Nortel Networks* and *D-Wave* are globally the most active firms. Xerox is particularly dominant in electrophoretic technologies for printer toner solutions (cluster 1375) and polithiophenes technologies (cluster 1368). Nortel Networks, as expected, is very active in optical solutions for networking and communications (clusters 1349 and 1365). D-Wave is the leading firm in quantum computing technology (cluster 1362). On the other hand, some branches, such as nanomedecine (cluster 1370), are not dominated by one big player. For instance, the biopharmaceutical company *Geron Corporation* is the number one patent holder in nanomedecine but owns less than 8% of all patents in this branch of nanotechnology. The same observation applies to LED and lighting technologies (cluster 1363) where the main player (*Brasscorp Ltd.*) holds less than 15% of all patents.

Analysis

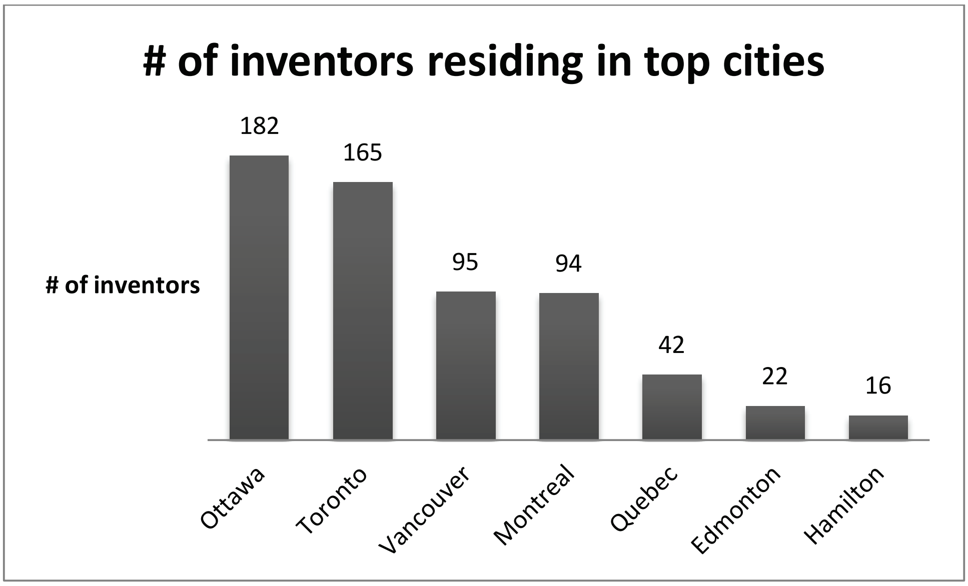

Figure 6: Top cities per number of inventors living in metropolitan area.

If we examine Canadian cities and the number of inventors residing there, we obtain the graph shown in figure 6. As we can see, the Ottawa metropolitan area, dubbed the Silicon Valley North, hosts the largest community of nanotechnology inventors. Toronto, Vancouver and Montreal follow with the second, third and fourth positions with somehow smaller communities given their population size compared to Ottawa. Table 2 shows the concentration of nanotechnology inventors in top Canadian cities. As we can see, Ottawa has an incredibly higher ratio of nanotechnology inventors by population. Quebec City has the second largest ratio of inventors per thousand inhabitants. Yet, Ottawa's ratio is 2.7 times larger than Quebec City's. Other cities have ratios of the same magnitude although small differences exist between cities. Toronto, Montreal and Vancouver, the three largest Canadian metropolitan areas, have relatively the same ratio of inventors by population.

Table 2: Ratio of nanotechnology inventors by metro area population.

City	Population	Number of inventors	Ratio (per thousand inhabitants)
Ottawa	1,130,761	182	0.16
Toronto	5,113,149	165	0.03
Vancouver	2,116,581	95	0.04
Montreal	3,635,571	94	0.03
Quebec	715,515	42	0.06
Edmonton	1,034,945	22	0.02
Hamilton	692,911	16	0.02

Although figure 6 indicates the domination of the technological scene by two cities (Toronto and Ottawa), the last column in table 3 shows that Montreal and Vancouver are not in such bad positions. For instance, Vancouver is the national leader in two technological areas (quantum computing and LED technologies) and has second position in nanomedecine. The latter technological branch is led by Montreal. Interestingly, these technological areas are either smaller (quantum computing and LED) or not dominated by one firm (LED and nanomedicine).. Given the importance of nanomedicine and the fact that it is not dominated by a big player, Montreal and Vancouver must take proper measures to strengthen their competitive position in this area. A complementary strategy for these cities can be to develop competences in neighboring branches. For instance, nanomedicine (cluster 1370) is very close to conductive polymers technologies (cluster 1368) as our knowledge map shows that they rely on the same technological base. Incidentally, Vancouver and Montreal (the leaders in nanomedicine) have the second and third most important communities in conductive polymers technologies even if they are far behind Toronto.

Table 3: Top words and firms per cluster

Cluster	**Top Words**	**Top Firms** **(# of patents obtained)**	**Top Cities** **(# of inventors)**
1349	optical ray x communications compensation	Nortel Networks (16) Applied Micro Circuits Corporation (3) FSONA Communications Corporation (2)	Ottawa (42) Montreal (5) Toronto (3) Quebec (3)
1362	qubit Quantum Resonant Superconducting fiber	D-Wave (25) University of Toronto (3) Luxtera, Inc. (2) MagiQ Technologies, Inc (2)	Vancouver (12) Toronto (7) Montreal (6)
1363	LED lamp Light inspection	Brasscorp Ltd. (4) EXFO Photonics (3) UView Ultraviolet Systems, Inc. (2) Mattson Technology	Vancouver (12) Toronto (11)

	systems	Canada, Inc. (2)	
1365	switch network switching optical wavelength	Nortel Networks (56) PTS Corporation (5) Enablence Inc. (4) JDS Uniphase Corporation (4) Raytheon Company (4)	Ottawa (87) Vancouver (8) Edmonton (4)
1368	Polythiophenes Organic film devices gelable	Xerox Corporation (36) LG Display Co., Ltd. (6) Chemokine Therapeutics Corp. (3)	Toronto (18) Vancouver (18) Montreal (11)
1369	optical grating chromatic wave wavelength	Lxsix Photonics (7) Teraxion Inc. (6) Photintech Inc. (5)	Ottawa (36) Quebec (29) Montreal (10)
1370	expression protein cells compositions acid	Geron Corporation (10) Arius Research Inc. (6) QLT Inc. (6)	Montreal (52) Vancouver (34) Toronto (16) Quebec (15) Edmonton (11)
1375	members Toner processes display Electrophoretic	Xerox Corporation (136) iFire Technology, Inc. (13) Nucryst Pharmaceuticals (12)	Toronto (103) Montreal (11) Hamilton (7) Vancouver (7) Ottawa (5)

5 Conclusion

This paper proposes a method to build a citation network from a sample of patents. It explains how to select the main network component and to build a citation matrix that is used to perform an AHC. With the hierarchical structure of the dendrogram generated by the AHC, we are able to deduce the technological relationship that exists between the clusters. Furthermore, an analysis of the patent titles for each cluster shows the most relevant words in each cluster. We use these words as labels describing the different branches of competences. By examining major patent holders in each branch we are able to identify the most active firms and institutions in each branch. Furthermore, by aggregating data about inventor cities, we are able to see where the largest community of practitioners resides.

We validated the method with the analysis of Canadian nanotechnology patents. From this application, many conclusions could be observed with a large practical impact for politics, deciders and researchers. The results show that Toronto and Ottawa are the most important Canadian centers for nanotechnology development with Nortel Networks and D-Wave being the most important Canadian firms holding patents in the USPTO. This shows that Canadian firms are in a stronger position in optical networking and communication solutions (with Nortel Networks) as well as in quantum computing (D-Wave). Since patenting is an indication of past investment in research and development, these firms have proven that they own a greater proportion of the stock of knowledge than any other Canadian firm when it comes to nanotechnology. The vast amount of knowledge these firms hold should give them the power to act as central players in the development of Canadian competences in nanotechnology. It is regrettable for Canada that Nortel has filed for bankruptcy and that Google has bid for its patent portfolio (GoogleBlog, 2011). If Nortel's bankruptcy leads to the dismantling of activities that were previously performed its nanotechnology R&D units, then a national-level intervention that would keep these activities running at more or less the same pace than before is highly recommended. In fact, high technology inventors have the privilege to be mobile, which could lead to their relocation to nanotechnology poles outside the country if local firms do not fill the void left by Nortel. Given the size of Nortel's nanotechnology patent portfolio compared to other Canadian firms, it wouldn't be sound to expect that all of its R&D activities can be taken over by one or even a group of local firms.

Finally, our study shows that our competence maps can be used as a decision tool when it comes to questions regarding the exploitation of a technological position or the exploration of new technological areas. We have seen that cities with limited overall capabilities can concentrate in developing one or a few areas of expertise and then expand their competences to other areas that rely on the same technological know-how. This is especially important in the case of cities like Montreal and Vancouver that are two main Canadian cities that are shadowed by a smaller but more technologically savvy city that is Ottawa. The former can take advantage of their leading position in the area of nanomedecine and expand their sphere of influence to conductive polymers technologies.

Next studies in this area may consider improving the visualization approach of the results. Also an interactive approach that will precise a step by step analysis, adding keywords search facilities at any time, will help decision makers for a more accurate competence map. One of the limitations of our methodology consists in the discarding of secondary network components from the competence map. As discussed in the article, this is a limitation due to the choice of AHC technique for organizing technological branches hierarchically. In future work, we hope to tackle this limitation by developing methods for the interaction of technological branches from disconnected network components. .

Acknowledgements

We would like to thank the SSHRC, the CIHR and the NSERC for their financial support. We thank the OST and particularly Vincent Larivière for access and support to the USPTO citation data. We thank Ricard-Olivier Moreau and Terence-Shawn Ramine Kananian for their rigorous contribution in cleaning and administrating our patent database. Finally, we thank Ilse-Carolina Monila Gutiérrez for preparing the nanotechnology keyword list.

References

Albayrak, Y. E. and Erensal, Y. C. (2009) 'Leveraging technological knowledge transfer by using fuzzy linear programming technique for multiattribute group decision making with fuzzy decision variables', *Journal of Intelligent Manufacturing*, 20, pp. 223-231.

Albert, R. and Barabási, A.-L. (2002) 'Statistical Mechanics of Complex Networks', *Review of Modern Physics*, 74, 47.

Alencar, M. S. M., Porter, A. L. and Antunes A. M. S. (2007) 'Nanopatenting patterns in relation to product life cycle', *Technological Forecasting and Social Change*, 74 (2007), pp. 1661–1680.

Ankerst, M., Breunig, M. M., Kriegel, H.-P. and Sander, J. (1999) 'OPTICS: Ordering Points To Identify the Clustering Structure', *ACM SIGMOD international conference on Management of data*, pp. 49–60.

Bassecoulard, E., Lelu, A. and Zitta, M. (2006) 'Mapping nanosciences by citation flows: A preliminary analysis', *Scientometrics*, 70 (3), 2007, pp. 859-880.Berry, M. J., and Linoff, G. S. (2004) *Data Mining Techniques for Marketing Sales and Customer Relationship Management*, Wiley.

Borner, K., Chen, C. and Boyack, K. W. (2003) 'Visualizing Knowledge Domains', *Annual Review of Information Science and Technology*, 37, pp. 179-255.

Chen, Y., Zhang, G., Hu, D. and Fu, C. (2007) 'Customer segmentation based on survival character', *Journal of Intelligent Manufacturing*, 18 (4), pp. 513-517.

Choudhary, A. K., Harding, J. A. and Tiwari, M. K. (2009) 'Data mining in manufacturing: A review based on the kind of knowledge', *Journal of Intelligent Manufacturing*, 20 (5), pp. 501-521.

CodePlex (2001) http://nodexl.codeplex.com.

Chryssolouris, F., Mavrikios, D., Xeromerites, S. and Georgoulias, K. (2008) 'Manufacturing Knowledge Work: The European Perspective' in Bernard, A. and Tichkiewitch, S., *Methods and Tools for Effective Knowledge Life-Cycle-Management*, pp. 213-225, Springer.

Duflou, J.R. and Verhaegen, P.-A. (2011) 'Systematic innovation through patent based product aspect analysis', *CIRP Annals – Manufacturing Technology*, 60 (1), pp. 203-206.

Ester, M., Kriegel, H.-P., Sander, J. and Xu, X. (1996) 'A density-based algorithm for discovering clusters in large spatial databases with noise' in Simoudis, E., Han, J. and Fayyad, U. M. *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (KDD-96)*, pp. 226–231.

Fattori, M., Pedrazzi, G. and Turra, R. (2003) 'Text mining applied to patent mapping: a practical business case', *World Patent Information*, 25, pp. 335-342.

Fitzgibbons, K. and McNiven, C. (2006) 'Towards a Nanotechnology Statistical Framework', *Blue Sky Indicators Conference II*.

GoogleBlog (2011) http://googleblog.blogspot.com/2011/04/patents-and-innovation.html.

Hsu, C., Babin, G., Bouziane, M., Cheung, W., Rattner, L., Rubenstein, A. and Yee L. (1994) 'The metadatabase approach to integrating and managing manufacturing information systems', *Journal of Intelligent Manufacturing,* 5 (5), pp. 333-349.

Kim, Y. G., Suh, J. H. and Park, S. C. (2008) 'Visualization of patent analysis for emergin technology', *Expert Systems with Applications*, 34, pp. 1804-1812.

Li, X., Hu, D., Dang, Y., Chen, H., Roco, M. C., Larson, C. A., Chan, J. (2009) 'Nano Mapper: an Internet knowledge mapping system for nanotechnology development', *Journal of Nanotechnology Research*, 11 (3), pp. 529-552.

Malakooti, B. and Raman, V. (2000) 'Clustering and selection of multiple criteria alternatives using unsupervised and supervised neural networks', *Journal of Intelligent Manufacturing*, 11, pp. 435-451.

Manning, C. D., Raghavan, P. and Schutze, H. (2008) *An Introduction to Information Retrieval*, Cambridge University Press.

Mogoutov, A. and Kahane, B. (2007) 'Data search strategy for science and technology emergence: A scalable and evolutionary query for nanotechnology tracking', *Research Policy*, 36, pp. 893–903.

Newman, M. E. J. and Girvan, M. (2004) 'Finding and evaluating community structure in networks', *Physical Review E*, 69, 026113.

Porter, M.F. (1980) 'An algorithm for suffix stripping', Program, 14 (3), pp 130-137.

Porter, A. L., Youtie, J., Shapira, P. and Schoeneck, D. J. (2008) 'Refining search terms for nanotechnology', *Journal of Nanoparticle Research*, 10, pp. 715–728.

Rapid-I (2011) http://rapid-i.com.

Schmoch, U., Heinze, T., Hinze, S. and Rangnow, R. (2003) *Mapping Excellence in Science and Technology across Europe: Nanoscience and Nanotechnology*, Centre for Science and Technology Studies.

Small, H. (1999) 'Visualizing Science by Citation Mapping', *Journal of the American Society for Information Science*, 50 (9), pp. 799-813.

Taskin, H. and Adali, M. R. (2004) ‘Technological intelligence and competitive strategies: An application study with fuzzy logic’, *Journal of Intelligent Manufacturing*, 15, pp. 417-419.

Tseng, Y.-H., Lin, C.-J. and Lin, Y.-I. (2007) ‘Text mining techniques for patent analysis’, *Information Processing and Management*, 433, pp. 1216-1247.

USPTO (2009) http://uspto.gov.

Wallace, M. L., Gingras Y., Duhon R. (2009) ‘A new approach for detecting scientific specialties from raw cocitation networks’, *Journal of the American Society for Information Science and Technology*, 60 (2), pp. 240-246.

Wasserman, F. and Fraust, K. (1994) *Social Network Analysis: Methods and Applications*, Cambridge University Press.

Weiss, S. M., Indurkhya, N., Zhang, T. and Damerau, F. J. (2005) *Text Mining: Predctive Methods for Analyzing Unstructured Information*, Springer.

Westphal, I., Thoben, K.-D. and Seifert, M. (2010) ‘Managing collaboration performance to govern virtual organizations’, *Journal of Intelligent Manufacturing*, 21 (3), pp. 311-320.

Wijnhoven, F. (2008) ‘Manufacturing Knowledge Work: The European Perspective’ in Bernard, A. and Tichkiewitch, S., *Methods and Tools for Effective Knowledge Life-Cycle-Management*, pp. 23-44, Springer.

Xu, R., and Wunsch II, D. (2005) ‘Survey of clustering algorithms’, *IEEE Transactions on Neural Networks*, 16(3), pp. 645-678.

Zitt, M., Bassecoulard, E. (2006) ‘Delineating complex scientific fields by an hybrid lexical-citation method: An application to nanosciences’, *Information Processing and Management*, 42, pp. 1513–1531.