Scientometrics (2012) 90:983–999
DOI 10.1007/s11192-011-0562-z

Mapping the research on aquaculture. A bibliometric analysis of aquaculture literature

Fabrizio Natale · Gianluca Fiore · Johann Hofherr

Received: 12 July 2011 / Published online: 26 November 2011

Abstract Research on aquaculture is expanding along with the exceptional growth of the sector and has an important role in supporting even further the future developments of this relatively young food production industry. In this paper we examined the aquaculture literature using bibliometrics and computational semantics methods (latent semantic analysis, topic model and co-citation analysis) to identify the main themes and trends in research. We analysed bibliographic information and abstracts of 14,308 scientific articles on aquaculture recorded in Scopus. Both the latent semantic analysis and the topic model indicate that the broad themes of research on aquaculture are related to genetics and reproduction, growth and physiology, farming systems and environment, nutrition, water quality, and health. The topic model gives an estimate of the relevance of these research themes by single articles, authors, research institutions, species and time. With the co-citation analysis it was possible to identify more specific research fronts, which are attracting high number of co-citations by the scientific community. The largest research fronts are related to probiotics, benthic sediments, genomics, integrated aquaculture and water treatment. In terms of temporal evolution, some research fronts such as probiotics, genomics, sea-lice, and environmental impacts from cage aquaculture, are still expanding while others, such as mangroves and shrimp farming, benthic sediments, are gradually losing weight. While bibliometric methods do not necessarily provide a measure of output or impact of research activities, they proved useful for mapping a research area, identifying the relevance of themes in the scientific literature and understanding how research fronts evolve and interact. By using different methodological approaches the study is taking advantage of the strengths of each method in mapping the research on aquaculture and showing in the meantime possible limitations and some directions for further improvements.

Keywords Aquaculture · Bibliometrics · Computational semantic · Topic model · Latent semantic analysis · Co-citation analysis

F. Natale (✉) · G. Fiore · J. Hofherr
European Commission. Joint Research Centre, Institute for the Protection and Security of the Citizen,
Via E.Fermi, 2749 I - 21027 Ispra (VA), Italy
e-mail: fabrizio.natale@jrc.ec.europa.eu

Introduction

Aquaculture is one of the world's fastest growing food production industries. In recent years (1970–2008) it has recorded an average annual growth of 8.3%, three times the rate of the world meat production (FAO 2010). An important factor determining this growth has been attributed to improved control over the production process, which has been made possible, among others, by dynamic research and technological developments. Since aquaculture is a relatively young industry in comparison to other forms of animal husbandry, there is still potential for productivity growth through further research and technological development (Asche 2008). Especially in this case it is therefore important for scientific research to adapt constantly in order to respond to policy and economic drivers.

A way for identifying key challenges for the future is to analyse how scientific literature is evolving, which are the "hot topics" emerging and which disciplines and technologies are receiving most attention. From this understanding policy makers may find targets for scientific funding and research institutions may identify areas of collaborations and decide about their positioning and priorities for future research activities.

In this study we analysed with quantitative methods the corpus of scientific literature on aquaculture since 2000 to understand what are the main research areas and research fronts, how these are interacting and how they have developed over time.

We used Latent Semantic Analysis (LSA), Topic Model (TM) and Co-citation Analysis (CCA), which are popular methods applied in computational semantics and in bibliometrics for analysing scientific papers extracted from bibliographic databases.

LSA and TM are methods enabling an unsupervised analysis of text in large collections of documents to identify topics and documents similarities from the contextual usage of words in the documents.

With the LSA method a document-term matrix is constructed on the basis of the frequencies of words appearing in a corpus of documents (Landauer et al. 1998). This matrix is resolved by singular value decomposition (SVD) in a term-vector and a document-vector matrix. Through SVD a dimensionality reduction is applied on the original document-term matrix which eliminates some of the noise given by different usage of words across documents and gives a deeper representation of the semantic concepts in the documents in respect of a simple computation of word frequencies and word co-occurrences (see Wild et al. 2005 for a graphical representation of the method). In the reduced latent semantic space, terms are considered similar on the basis of associative closeness. The reference to words and documents similarities in the reduced semantic space gives the possibility of solving problems of polysemy (different words for the same meaning) and synonymy (words with similar meaning). This idea is extensively applied in information retrieval and web search engines to find similar documents and establish latent semantic indexing of documents. In this study we used a spatial representation of the similarity values between words to give an indication of the subjects represented in the corpus.

TM is a more recent method for finding scientific topics in a literature corpus which uses a probabilistic rather than a spatial approach as in LSA (Blei et al. 2003; Griffiths and Steyvers 2004; Steyvers 2007). TM consists of a generative model based on the idea that each document is expressing a mixture of topics. Each topic is expressed using a set of characteristic words and can be represented by a multinomial distribution over words (bags of words). According to this model different documents can be generated by choosing a distribution over topics and sampling randomly from the corresponding bag of words.

In TM, the inference of the probability distributions over words and over topics that define the generative model is carried out by introducing a prior distribution over topics and fixing a parameter for the number of expected topics.

TM is useful for finding interpretable topics with semantic meaning and for assigning these topics to the documents. Since the approach is probabilistic, the word composition of each topic and the assignment of topics to the documents are expressed as probabilities and allow for a word to be present in more topics and for a document expressing multiple topics. From these assignments it is possible to measure the relevance of topics in single documents and, by taking the average of the values across the entire corpus, to have a global evaluation by affiliation, author, time and other variables available in the bibliographic data.

Both LSA and TM are unsupervised methods that try to identify the cognitive structure, respectively, from a spatial and a statistical analysis of the text in the corpus.

An alternative approach followed in bibliometrics is to identify research fronts by examining the social interactions between scientists as they emerge from the co-citation mechanism (Small 2006). The choices of authors in building their list of references determine an association of citing and co-cited documents that gives a delineation of common research themes.

A co-citation network can be constructed by establishing a link between two documents that are cited together by a third paper. The fact that they are frequently cited together by many other documents indicates that the relation is particularly strong and this can be considered a sign of the emergence of a defined research front (Small and Griffith 1974). The research front is represented by a set of key and frequently cited seminal documents forming its knowledge base and by the documents citing them which constitute the more recent evolution of the knowledge base. Further applications stemming from the original co-citation method include visual and network analysis methodologies to study interactions between topics and their dynamics over time (Chen 2006).

In this study we have applied the three methods of LSA, TM and CCA to analyse the titles, keywords, abstracts and list of references in the literature on aquaculture extracted from the bibliographic database Scopus. Although the main objective was not to compare or improve the methods, we applied to our knowledge for the first time these different approaches to the same corpus of bibliographic data. The three described methods offer the possibility to analyse efficiently the large amount of data which is available in bibliographic databases. It would be reductive to use this data only as a way to quantify scientific output by number of publications and number of citations. The associations between terms in the texts, abstract and titles and the interactions between authors through the citation mechanism represent instead an opportunity to map the structure of knowledge in a scientific area and to understand how it is evolving. Although using different perspectives the three methods have in common the idea of using the great amount of bibliographic data to let emerge in an unsupervised way the underlying knowledge base. Instead of defining a preventive hypothesis to model the structure of themes, the analysis of the correlations between words and citations links speak by themselves about the content and the evolution of ideas in the scientific area. By applying a mixed-method approach to the concrete case of the literature on aquaculture we exploit the main advantages of each method in getting a synthetic representation of the research trends on aquaculture and give some indications about limitations and possible future developments of these methods.

Materials and methods

The bibliographic data used in the study was extracted from the Scopus database by searching the word "aquaculture" in the title, abstract and keywords list. The bibliographic data related to aquaculture included the title, authors, affiliations, abstract, references and keywords.

The LSA method was performed using the package "lsa" for R (Wild 2005) on a corpus of documents each represented by the title and keywords of the documents.

The title and keywords were used to create a word-document matrix, after filtering out function words and common terms (stop-words), words with global frequency lower than 100 and length of less than three characters. A similarity score between words was calculated as cosine on the document-term matrix in the latent semantic space and these scores were converted into dissimilarities, ranging from zero (exactly the same) to one (exactly the opposite). Finally, the matrix of dissimilarity values was subjected to classical multidimensional scaling in order to allow visualisation in a bi-dimensional space. The threshold of 100 for the global frequency of words was chosen to produce a visualisation limited to the most relevant words and avoid visual cluttering from a great number of more specialised terms.

In the case of TM the entire text of abstract, keywords and titles for each document was considered in the corpus. The text was subject to stemming, exclusion of stop-words and of words with global frequency below 10. The model generation was run using the package "topicmodels" in R (Gruen and Hornik 2011). Following the Latent Dirichlet Allocation method described by Blei et al. (2003), the distribution of topics over documents was defined as symmetric Dirichlet distribution with a single hyper parameter of 0.1. This parameter controls the smoothing of the topic distribution with low values favouring the assignments of few topics to a document. Another parameter in the model was about the number of topics to be generated. Although statistical methods and metrics of model fit have been proposed to determine the best number of topics, Chang et al. (2009) showed that these methods do not always address the explanatory goal of the model; models performing better in terms of held-out likelihood may in fact generate less semantically meaningful topics. In alternative, they propose a method based on human judgement to find intrusion words in the list of most probable words of the identified topics. Following the same idea we tested models with an increasing number of topics and assessed the coherence of the generated topics looking at the number of intruding words among the top 20 most probable words.

As a variant to the classical TM method we introduced some control in the generation of topics by excluding the words identifying main animal species from the corpus. In this way the topics generated by the model were forced to represent exclusively disciplines and methodologies. The relevance in respect of animal species was assessed independently by finding word occurrences in the abstracts, title and keywords and than combined with the results of the topic model.

Since Scopus does not provide unique identifiers allowing direct connections to the cited references, in order to perform the CCA, the text of references was searched to identify a string of the title of the cited papers. Citations links were only established between papers in the selected aquaculture corpus which implies that external citations to documents from different subject areas were not considered in the study. The list of links between citing and cited documents was represented as an un-weighted directed network where each source node corresponded to the citing paper and each destination node to a cited paper. This network was further transformed in a co-citation undirected and weighted network using the software Sci2 (Sci2 Team 2009). In this network each node corresponded to a cited paper and the link between nodes to the fact that two documents were co-cited by at least one document in the corpus. For each edge a weight attribute was calculated on the basis of the number of documents co-citing the node pairs and a normalised co-citation measure was calculated as frequency of co-citations between the two documents divided by the square root of the product of their citation frequencies (Small 2006).

The co-citation network was simplified by excluding links with weight below two and normalized co-citation below 0.3 (Thomson 2008), with the two thresholds punishing respectively links based on few co-citations, and co-citations representing only a small share of the citations received by the two linked papers. The relatively high value of the co-citation threshold contributed to domain specificity, by penalising the common tendency to cite generic and highly cited papers in important journals, like Nature and Science, rather than more specialist papers with more limited dissemination.

In order to identify clusters on the co-citation network we used a network-agglomerating algorithm named FAG-EC, which identifies modules as group of nodes having a number of connections within the group higher than the number of connections out of the group (Li et al. 2008). FAG-EC is governed by two parameters indicating: the number of connections within the group in respect of the number of connections out of the group, and the minimum number of nodes in each cluster. For the first parameter a value of one was chosen to get the maximum number of clusters and therefore more specific research fronts, while the second parameter was set to three to avoid an excessive fragmentation in research fronts composed by only two co-cited papers. Finally, to represent the interactions between research fronts at a higher level of aggregation, a new co-citation network was created by grouping all the documents in each cluster in a single node and considering the number of co-citations between these new nodes.

Results

The dataset of bibliographic data extracted from Scopus included a total of 14,308 documents published between 2000 and May 2011. These documents were mostly articles (86.5%) and conference papers (8.4%). The evolution over time of the number of published documents showed a constant expansion in line with similar trends in other research areas such as the research on poultry (Fig. 1).

The main source journals on the basis of the number of publications were: Aquaculture (12.3%), Aquaculture Research (5.1%) and the Journal of World Aquaculture Society (3.9%). The main countries on the basis of the affiliations of authors were United States (16.5%), China (6.3%) and United Kingdom (5.6%) (Table 1).

For the LSA method a total of 337 out of 24,349 unique words in the titles and keywords lists were retained in the document-term matrix after filtering. In addition to aquaculture, the seven most used words were fish (in 3,306 documents), growth (in 2,106 documents), water (in 1,522 documents), sea (in 1,382 documents), shrimp (in 1,289 documents) and salmon (in 1,141 documents). Figure 2 shows for the top 100 words, with global frequency above 286, the positioning in a bi-dimensional space on the basis their dissimilarities calculated from the cosine values. A short distance between words indicates a strong association which is emerging by a common usage in the same context; from these associations it is possible to identify broad areas of research which characterise the scientific literature on aquaculture. The disciplines tend to arrange into three main branches.

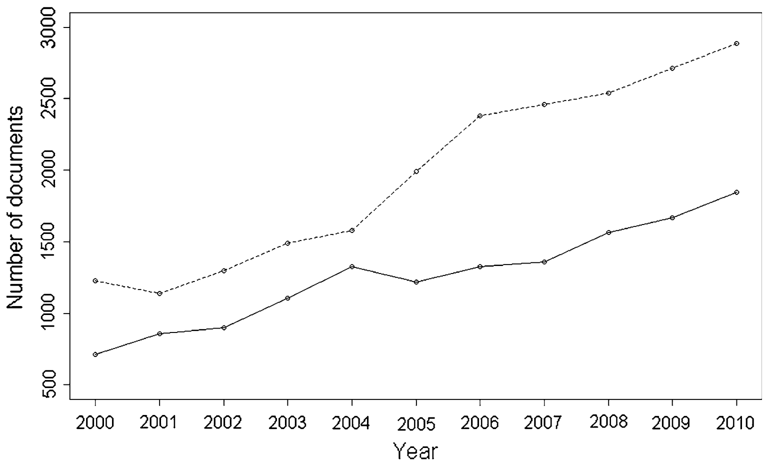

Fig. 1 Number of documents on aquaculture (*solid line*) and on poultry (*dotted line*) recorded in Scopus between 2000 and 2010

Table 1 Top ten journals and top ten countries on the basis of the number of documents on aquaculture between 2000 and May 2011

Journal	Nr (%)	Country	Nr (%)
Aquaculture	1823 (12.3)	United States	2950 (16.5)
Aquaculture Research	760 (5.1)	China	1124 (6.3)
Journal of the World Aquaculture Society	583 (3.9)	United Kingdom	995 (5.6)
Diseases of Aquatic Organisms	398 (2.7)	Canada	988 (5.5)
Aquacultural Engineering	366 (2.5)	Australia	940 (5.3)
Aquaculture International	279 (1.9)	Spain	894 (5.0)
Journal of Fish Diseases	201 (1.4)	Norway	697 (3.9)
Hydrobiologia	185 (1.2)	France	690 (3.9)
Journal of Shellfish Research	170 (1.1)	India	688 (3.8)
Fish and Shellfish Immunology	163 (1.1)	Japan	678 (3.8)

The ranking of countries is calculated considering the affiliations of all the authors of each article

A central branch is composed by associations of words which span from left to right, from economic aspects to environmental assessment, production systems, water quality and control of the production environment. From this central branch two lateral branches depart, on one side, towards reproduction, growth performance and nutrition and, on the other side, towards disease control and genetic methodologies. Words identifying the farmed species tend to be positioned in a central area equidistant from the disciplines, with some specificities: catfish, tilapia, sea bass and sea bream appear to be more associated with research on reproductive and nutritional aspects; mussel and oyster are closer to water quality and shrimp and salmon are closer to genetic and disease studies.

The TM method provided an additional insight in the structuring of the aquaculture research into different topics. By increasing the number of topics the model started to produce more specific aggregations although this affected the semantic coherence of the topics by progressive intrusion of non-relevant terms among the most probable words. We

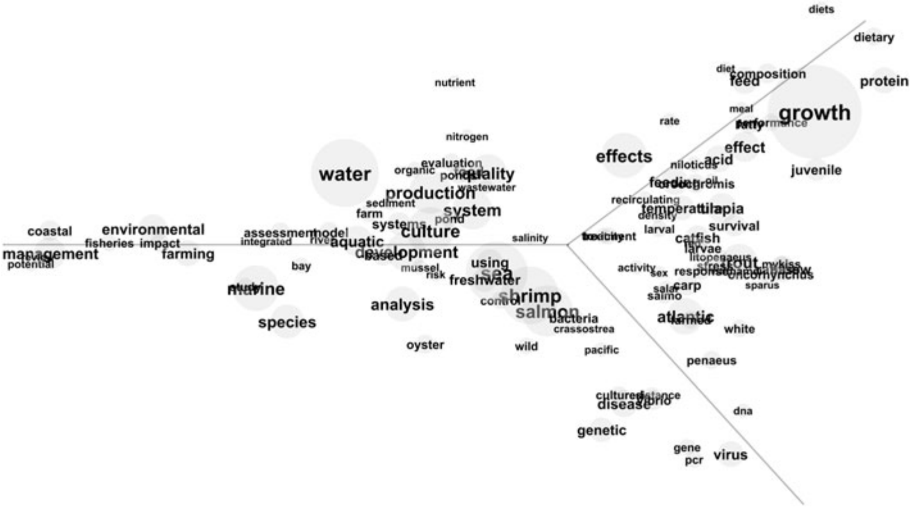

Fig. 2 Map of the aquaculture literature, based on the Latent Semantic Analysis of keywords and titles. The position and distance reflects the semantic association of words. The *font* and *circle* size reflect word global frequencies. The map shows top words (with a global frequency above 286) with the exclusion of fish and aquaculture. The words are spatially distributed along three main branches identifying main themes of aquaculture research

considered that a model with six topics offered an acceptable compromise between level of aggregation and coherence. Each topic generated by the model was defined by its own probability distribution over a total of 8,861 words selected from the abstracts, with the most probable words explaining its meaning. Figure 3 represents the top 20 words for each of the six topics on the basis of their probability. Since the topics are defined by the probabilities of words rather than by hard clustering a word may appear in more than one topic. The six identified topics can be synthetically described as dealing with: genetics and reproduction, growth and physiology, farming systems and environment, nutrition, water quality and health. Using the model with six topics we generated a posterior distribution of documents over topics. This distribution gives a measure of relevance of the topics in each document of the corpus. The examples in Table 2 show the assignments of topics to seven randomly selected abstracts from the corpus. The classification of documents provided by the TM was assessed by comparing the content of the abstract with the probabilities assignments across the six topics of 200 randomly selected documents. Overall the classification appeared to be reliable and only 10% of the articles were judged as misclassified. The last document listed in Table 2 shows an example of misclassification. The article was part of the corpus since it included the term aquaculture in the list of keywords; however, since the main content is about geothermal energy and only incidentally relevant for aquaculture, also the scores attributed by the model are not correctly reflecting the relevance in respect of the main topics which characterise the aquaculture literature. By averaging the topic probability in each document a measure of relevance for each topic is given for the entire corpus of aquaculture literature and for the different segmentations on the basis of the variables provided in the bibliographic data. According to the model results the relevance of the identified topics in the aquaculture literature was as follows: farming systems and environment 26.5%, water quality control 17.5%, health 16.4%, genetics and reproduction 13.9%, growth and physiology 13.8% and nutrition 11.9%. Figure 4 shows

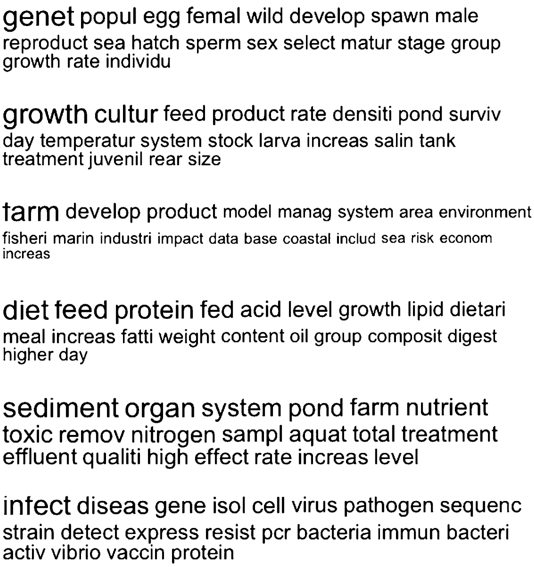
genet popul egg femal wild develop spawn male reproduct sea hatch sperm sex select matur stage group growth rate individu
growth cultur feed product rate densiti pond surviv day temperatur system stock larva increas salin tank treatment juvenil rear size
farm develop product model manag system area environment fisheri marin industri impact data base coastal includ sea risk econom increas
diet feed protein fed acid level growth lipid dietari meal increas fatti weight content oil group composit digest higher day
sediment organ system pond farm nutrient toxic remov nitrogen sampl aquat total treatment effluent qualiti high effect rate increas level
infect diseas gene isol cell virus pathogen sequenc strain detect express resist pcr bacteria immun bacteri activ vibrio vaccin protein

Fig. 3 Topics in the aquaculture literature identified by the Topic Model. Each topic is represented by the 20 most probable words in the distribution over words. The font size reflects the word probability. The topics from top to bottom can be synthetically described as: genetics and reproduction, growth and physiology, farming systems and environment, nutrition, water quality, and health

the relevance of the topics for different years. The relevance of the topic on farming systems and environment peaked in 2007–2008, and on water quality in 2009. The evolution over time shows also the emergence of the health topic, which is almost reaching the relevance of the water quality topic.

The content of the topics represents exclusively disciplines and methodologies since the terms related to species were filtered out before the inference of the model. The relevance in respect of common farmed species for each document was quantified separately on the basis of the occurrence of generic or scientific names identifying the species in the abstract, title or keywords. In case of multiple species occurring in an article the score of one was distributed in equal proportion to each species. This provided an additional dimension to evaluate the relevance of documents and of the global literature on aquaculture for the different combinations between species and disciplines or methodologies. The score for the topic-species combinations was obtained as product between the topic probability and the occurrences of species terms for each article and the results were averaged across the entire corpus to have a total measure of relevance for each topic-species pair. Figure 5 shows that a relevant part of research on shrimp and salmon is related to health while in the case of sea bass and sea bream the focus is more on nutritional aspects and reproduction. Reproduction has the highest relevance in the case of sturgeon, while the topic on farming systems and environment has the highest relevance in the case of tuna.

Through the analysis of the literature with the CCA method a network of 999 highly co-cited documents was generated. This network was produced by analysing 25,603 citations by 7,463 citing documents. The remaining 6,485 documents in the aquaculture corpus had either no references or external references to documents not included in the corpus. By examining the connectivity of this co-citation network with the FAG-EC algorithm a total of 103 clusters were identified. Each cluster was made by highly cited documents which were often cited together and which could therefore be considered as representing the knowledge base of a specific research front.

From the analysis of the links in the co-citation network it was possible to identify interesting relations between research fronts, which underline the multi-disciplinary nature

Table 2 Classification of documents on the basis of the six topics generated by the Topic Model

Extract from the abstract	Genetics and reproduction	Growth and physiology	Farming systems and environment	Nutrition	Water quality	Health
Molecular tools to assist breeding programs in the gilthead sea bream (*Sparus aurata* L.) are scarce. A new multiplex PCR technique (OVIDORPLEX), which amplifies nine known microsatellite markers, was developed in this work	0.9	0.0	0.0	0.0	0.0	0.1
The sapphire devil, *Chrysiptera cyanea*, were reared for 45 days during the non-reproductive season (September) under LD14:10 at four different wavelengths produced by light emitting diodes (LEDs): red (peak at 627 nm), green (530 nm), blue (455 nm) and white (5,000 K). Ovarian maturation occurre	0.6	0.4	0.0	0.0	0.0	0.0
Limited information is available on vaccine performance in parasitized fish. The objective of this study was to determine if parasitism of fish affected vaccine efficacy. Antibody level, hematology and survival of *Nile tilapia* vaccinated	0.0	0.0	0.0	0.1	0.0	0.9
The concentrations of 16 PAHs in surface sediments collected from four Italian lagoons, exploited for aquaculture and fishing activities, during the period 2004–2007, were analysed	0.0	0.0	0.2	0.0	0.8	0.0
Removing solids is an essential task when recirculating water an aquaculture system. Dissolved solids production directly from particulate solids	0.0	0.0	0.1	0.0	0.9	0.0
The proliferation of bacteria in intensive aquaculture systems may be responsible for poor growth and mass mortality of marine fish larvae. Essential fatty acids provided in the diet could protect larvae by modulation of the immune response via arachidonic acid (AA) and eicosapentaenoic acid (EPA)	0.0	0.2	0.0	0.5	0.0	0.2

Table 2 continued

Extract from the abstract	Genetics and reproduction	Growth and physiology	Farming systems and environment	Nutrition	Water quality	Health
The potential for integrating aquaculture with agriculture has been widely recognized as a means of improving the use of inputs, diversifying output and economic opportunity, and enabling smallholder producers to maintain and strengthen livelihoods. This paper describes the outcomes of this approach	0.0	0.1	0.9	0.0	0.0	0.0
The worldwide application of geothermal energy for direct utilization is reviewed. This paper attempts to update the previous survey carried out in 1995 (Freeston 1995) and presented at the World Geothermal Congress 1995 in Florence, Italy. For each of these updates since 1975, the recording of data has been similar, but not exactly the same. As in 1995, an effort was made to quantify geothermal heat pump data and the investment in geothermal energy development	0.0	0.0	0.4	0.0	0.6	0.0

The examples include seven randomly selected abstracts and in the last row an example of misclassification

of aquaculture research and the crossbreeding between different disciplines and research subjects. Figure 6 shows some examples of relations between four different clusters representing a portion of the entire co-citation network. A first relation is between two clusters dealing respectively with sex control and genomics. This relation reflects the potential use of induced triploidy, on one side, to avoid problems of lower growth rates associated with sexual maturation and, on the other side, for the containment of genetic contamination from escaped farmed fish affecting wild populations.

The genomics cluster is also related to the cluster dealing with sea-lice through a paper about the assessment of salmon aquaculture impacts on wild salmonids. A third relation is between the sea-lice cluster and the cluster on the epidemiology of salmon anaemia virus which is based on a paper dealing with risk-analysis methods for emerging diseases in salmon. These examples outline possible areas of collaboration between research fronts making use of both, epidemiological modelling and genomics to asses the risk of spreading of major diseases from farmed fish into wild populations.

To represent more synthetically the relations between clusters they were manually labelled on the basis of the title of their members and represented themselves as nodes in a cluster co-citation network at a higher level of aggregation (Fig. 7). In this higher level co-citation network the importance of each research front was quantified on the basis of the

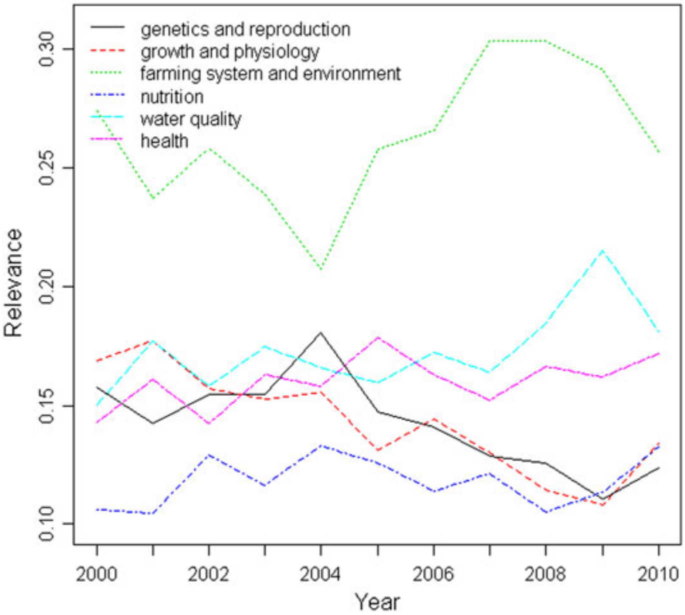

Fig. 4 Relevance of topics in the aquaculture literature by year. The relevance is calculated using the Topic Model, as mean of the probability of each topic being represented in the documents of the aquaculture literature

number of documents citing the members of each cluster and the strength of the relations between research fronts was given by the number of documents co-citing members of two connected clusters. Based on the number of citations the top ten research fronts were: probiotics (648 citations), benthic sediments (537), genomics (436), integrated aquaculture (416) and water treatment (347). Strong relations on the basis of the number of documents co-citing members of the clusters are evident in particular in the following cases:

benthic sediments with shrimp farming effluent (52), with environmental impacts of cage aquaculture (53) and with phosphorous sediment (40);

water treatment with bio-filters (55), with de-nitrification in re-circulating systems (33) and with integrated aquaculture (30);

genomics with sex and reproduction controls (28);

off-shore aquaculture with carrying capacity in shellfish (38);

carrying capacity in shellfish with integrated aquaculture (25).

As documented by other bibliometric studies, research fronts tend to exhibit a life-cycle characterized by a phase of expansion, as important documents introducing key and innovative ideas start to be highly cited, and a subsequent phase of decline, when these documents representing the knowledge base are superseded by more recent and innovative ideas.

The evolution of the size of the research fronts, measured as number of citations over time, gives useful indications about the trends in a research area and on the main topics where research is currently focused. Figure 8 shows the frequency of citation for the major research fronts in the period considered in the study.

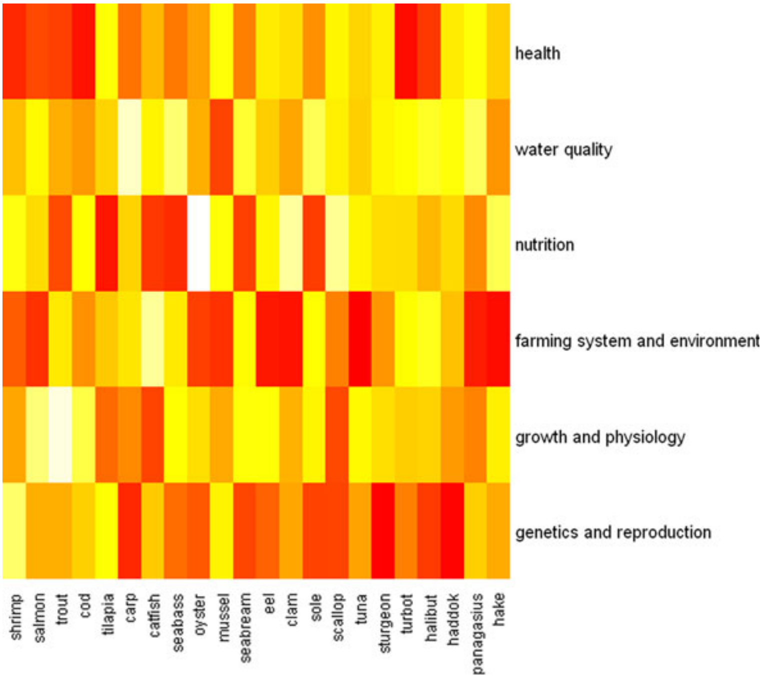

Fig. 5 Relevance of topics by main farmed species. The heat for each topic-species combination is derived from the product between probabilities from the Topic Model and occurrences of terms identifying the species in titles, abstracts and keywords

Some research fronts like probiotics, genomics, sea-lice, environmental impacts from cage aquaculture are more recent and still expanding while other, such as mangroves and shrimp farming, and benthic sediments, after an expansion in past years, are gradually loosing importance. This indicates either that the issue has lost interest by the scientific community or that a new knowledge base for the research front is needed to bring it to further developments. The appearance of multiple peaks is generally a sign of the appearance of new key documents which revamp the interest in a research front keeping it large enough to emerge in the CCA.

Discussion

In this study we applied popular methods used in cognitive science and bibliometrics to map the aquaculture literature between 2000 and May 2011. The LSA method gives a spatial representation of word similarities according to their usage in the title and keyword list of each of the documents forming the corpus of aquaculture research. From the arrangement of words in a bi-dimensional space it was possible to recognise visually broad areas of research and their associations. For the purpose of producing a map of the words association, the LSA using only titles and keywords was preferred instead of using also the abstracts. Title and keywords already give a pre-selection directed by the authors of the important words that are likely to represent the research area, while the inclusion of abstracts did not improve the clarity of the map since it added many common and less relevant terms.

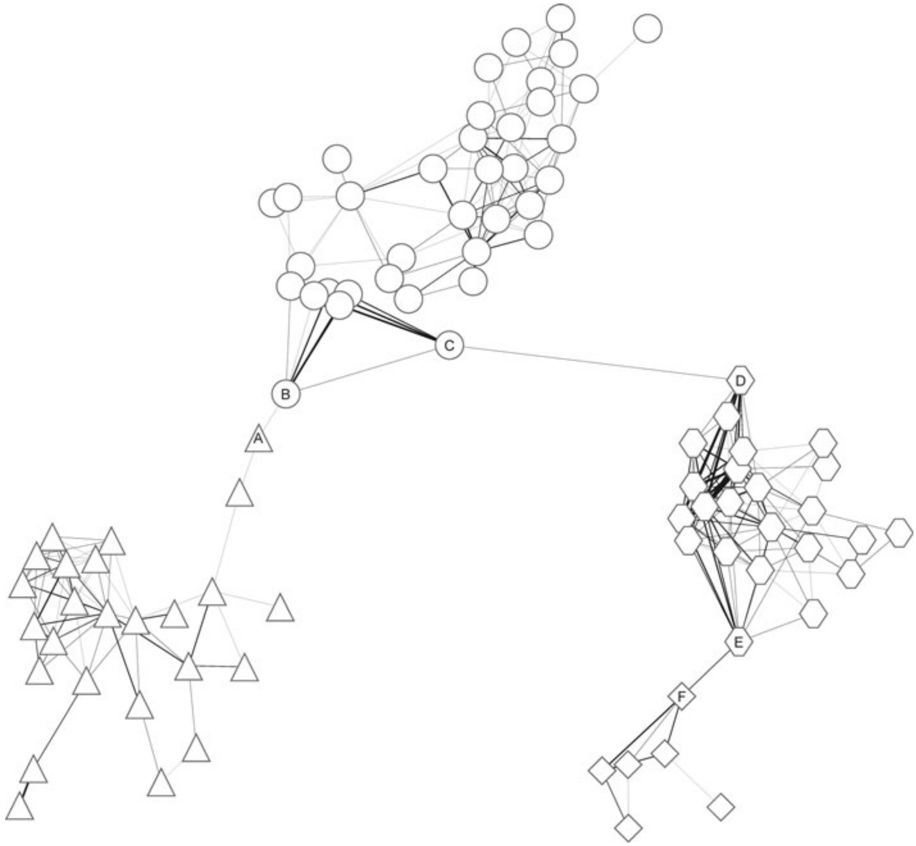

Fig. 6 Relations between documents in a portion of the co-citation network. Each node represents a highly cited document and edges are built on the basis of co-citations. These documents form four clusters related respectively to: sex control (*triangles*), genomics (*circles*), sea-lice (*hexagons*) and the epidemiology of salmon anaemia virus (*diamonds*). The titles of the documents building connections between clusters are: *A* Interactions between Aquaculture and Wild Stocks of Atlantic Salmon and other Diadromous Fish Species: Science and Management, Challenges and Solutions. An introduction by the Conveners; *B* Dominance relationships and behavioural correlates of individual spawning success in farmed and wild male Atlantic salmon, *Salmo salar*; *C* Genetic and ecological effects of salmon farming on wild salmon: modelling from experimental results; *D* A global assessment of salmon aquaculture impacts on wild salmonids; *E* A model of salmon louse production in Norway: Effects of increasing salmon production and public management measures; *F* A framework for understanding the potential for emerging diseases in aquaculture

The TM method offered the possibility of identifying topics and evaluating their relevance in the corpus. In the case of TM the text of the abstract included in the corpus provided a larger basis of natural language for the inference of the distributions of the probabilistic model. The TM model with six topics showed an acceptable level of coherence in terms of word composition and provided a method for the classification of documents in probabilistic terms which catered for the possibility of documents to deal with multiple topics.

The identified topics correspond to a classification scheme of the sector of aquaculture, which is recurring in the subdivision of subject areas in scientific journals, in the

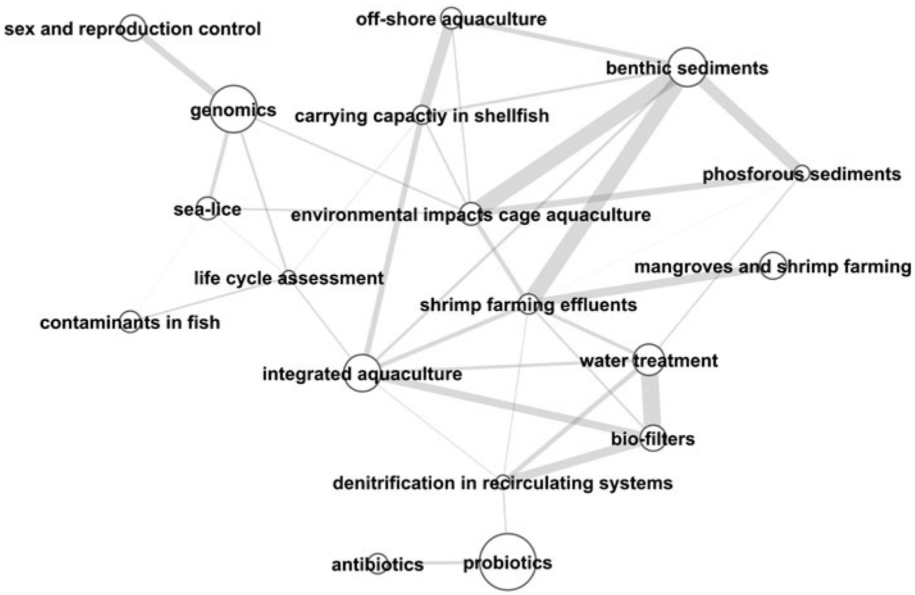

Fig. 7 Relations between research fronts identified by the Co-citation Analysis. Each node represents a cluster of highly cited papers. The relations between two research fronts (edges) are established on the basis of co-citations between papers in each research front. The size of the node represents the total citations of the documents belonging to each research front

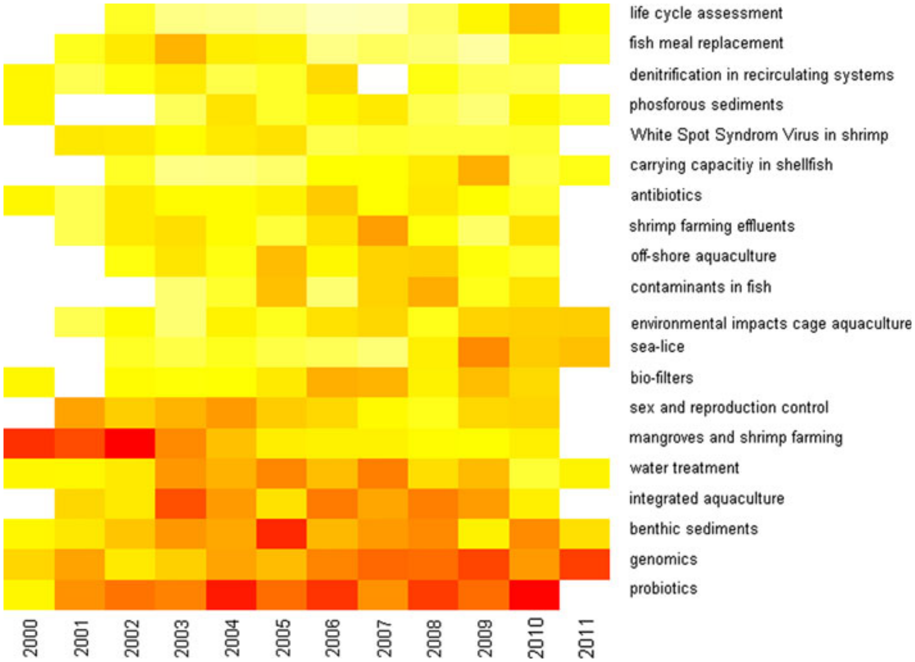

Fig. 8 Relevance of the research fronts by year. The heat is calculated from the number of citations received by the documents belonging to each research front by year

organisation of research institutions and in the structuring of themes in knowledge platforms (EATIP 2011). Although the TM method does not offer new insights in representing the main areas of the research it gives, to our knowledge for the first time, the possibility of assessing their relevance with a quantitative method.

On the basis of the assignments to documents we could quantify the topic relevance for single authors, journals, institutions and time. In respect of the classical approach used in TM, we excluded from the identification of topics any reference to species and looked independently at relevance in respect of species on the basis of simple word occurrences. This innovative application of the method improved the coherence of the topics and added an additional dimension for the classification of the corpus on the basis of combinations of species and methods or disciplines. Whilst providing a method for quantifying the relevance for a given set of topics, TM remains an automated classification system and a decrease in semantic coherence was observed as the number of topics increased. The need to evaluate the explanatory validity of the generated topics through human judgment indicates that there are still limitations of such an automated approach in identifying coherent topics especially if, by increasing their number, they become too narrow and specific. The filtering of words in the corpus operated in our study in reference to the main aquaculture species may represent a way forward in guiding the model generation process towards a more coherent topics definition. This idea could be further extended by applying standard lists of stop-words related to methodologies, disciplines and subjects of research in order to transform the original corpus of bibliographic data into more homogenous subsets to be analysed independently. The topic generated on these filtered corpora would have a higher semantic coherence and provide separate dimensions for the classification of documents and for the assessment of the relevance of the topics.

The CCA method, which makes use of the interactions between scientists through the citation mechanism, allowed for the identification of more specific research fronts. These research fronts emerge only if a sufficient mass of citations aggregate around a nucleus of highly cited documents and therefore cannot be considered as representative of all the topics that are present in the corpus. The temporal analysis of the distribution of citations shows how research fronts are continuously emerging and evolving as new ideas start to supersede old ones. The CCA method allows detecting research fronts from the past or which are expanding; on the other hand it is more difficult to detect new trends if they have not yet attracted a sufficient number of citations. Although the results of the CCA cannot be considered as representative of all research subjects, they give a complementary and more specific view in respect of the global map and topics generated with the LSA and TM methods.

This study was not aimed at identifying research needs, but rather to explore the research topics and trends and in a wider sense the behaviour of the scientific community network. The amount of publications reflects the interest of the scientific community, which, despite giving useful indications, should not be directly and automatically associated with the quality of research and the relevance of research in addressing the sector's needs. On the other hand, the methods applied in this study could well be tools for monitoring research output. Delanghe et al. (2011) describes for example how bibliometric methods can serve as output indicators for policy purpose in the European Commission's Directorate—General for Research.

In the design of research projects, the applied methods could help to identify appropriate and promising areas of collaboration and interdisciplinary set up, as shown with the linkage of some research fronts in salmon (genomics—sea-lice—impacts on wild salmon,—epidemiology of salmon anaemia virus—analysis methods for emerging diseases in salmon).

In general terms the paper reconfirms the key research areas on aquaculture already identified by other qualitative studies (FEUFAR 2008) and provides in addition a quantitative evaluation of their relevance across the more recent scientific literature.

We can summarise below the main advantages and limitations that we experienced from the application of the three methods in reference to the analysis of the aquaculture literature.

The LSA provided the most general view of the entire research area on the basis of word associations but was lacking the possibility of quantifying the relevance of topics. A spatial representation of words or documents on the basis of their similarity values was useful to get an overall picture of the research area but did not offer by itself a way of clustering the terms into topics and assessing in quantitative terms their relevance across the examined corpus.

The TM method offered in probabilistic terms a quantification of relevance both for the identification of topics and for the classification of documents, however, it proved to lose semantic coherence with an increasing number of topics. With few topics TM provided a robust system of classification of documents although it was too generic to allow the identification of the emergence of specific trends.

The CCA method offered a more specific representation of the research fronts, however, by being sensitive to the temporal shift of interest towards most recent papers it did not provide a consolidated picture of research themes. The temporal sensitivity of CCA method could be seen on one side as a limitation in identifying newest research fronts but on the other side, also as an opportunity for developing this approach further in the analysis of the dynamic evolution of the scientific knowledge and its progressing through the constant process of connections of well established ideas with more recent ones.

References

Asche, F. (2008). Farming the sea. *Marine Resource Economics, 23*(4), 527–547.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet allocation. *Journal of Machine Learning Research, 3*, 993–1022.

Chang, J., Boyd-Graber, J., Gerrish, S., Wang, C., & Blei, D. (2009). *Reading tea leaves: How Humans Interpret Topic Models*. In Neural Information Processing Systems (NIPS).

Chen, C. (2006). CiteSpace II: Detecting and visualizing emerging trends and transient patterns in scientific literature. *Journal of the American Society for Information Science and Technology, 57*(3), 359–377.

Delanghe, H., Sloan, B., & Muldur, U. (2011). European research policy and bibliometric indicators, 1990–2005. *Scientometrics, 87*(2), 389–398.

EATIP *European aquaculture technology and innovation platform*. (2011). Retrieved 25 October 2011 from http://www.eatip.eu/.

FAO. *The state of world fisheries and aquaculture*. (2010). Retrieved 25 October 2011 from http://www.fao.org/docrep/013/i1820e/i1820e00.htm.

FEUFAR *The future of European fisheries and aquaculture research Final Report*. (2008). Retrieved 25 October 2011 from http://www.feufar.eu.

Griffiths, T. L., & Steyvers, M. (2004). Finding scientific topics. *Proceedings of the National Academy of Sciences of the United States of America, 101*(1), 5228–5235.

Gruen, B., & Hornik, K. (2011). Topicmodels: An R package for fitting topic models. *Journal of Statistical Software, 40*(13), 1–30.

Landauer, T., Foltz, P. W., & Laham, D. (1998). Introduction to Latent Semantic analysis. *Discourse Processes, 25*, 259–284.

Li, M., Wang, J., & Chen, J. (2008). A fast agglomerate algorithm for mining functional modules in protein interaction networks. BioMedical engineering and informatics: New development and the future. In *Proceedings of the 1st International Conference on BioMedical Engineering and Informatics, BMEI 2008* (pp. 1–603) Hainan.

Sci2 Team. (2009). *Science of Science (Sci2) Tool*. Indiana University and SciTech Strategies, http://sci2.cns.iu.edu.

Small, H. (2006). Tracking and predicting growth areas in science. *Scientometrics, 68*(3), 595–610.

Small, H., & Griffith, B. C. (1974). The structure of scientific literatures I: Identifying and graphing specialties. *Science Studies, 4*(1), 17–40.

Steyvers, M. (2007). Probabilistic topic models. In: T. Landauer, D McNamara, S. Dennis, and W. Kintsch (Eds.), *Latent Semantic Analysis: A Road to Meaning*. Hillsdale: Erlbaum.

Thomson. (2008). *Research front methodology*, Retrieved 25 October 2011, from http://esi-topics.com/RFmethodology.html.

Wild, F. (2005). *lsa: Latent Semantic Analysis*. R package version 0.57.

Wild, F., Stahl, C., Stermsek, G., & Neumann, G. 2005. Parameters driving effectiveness of automated essay scoring with LSA. In *Proceedings of the 9th International Computer Assisted Assessment Conference (CAA)*, pp. 485–494.