Trend Topic Analysis for Wind Energy Researches: A Data Mining Approach Using Text Mining

Yunus Eroğlu* and Serap U. Seçkiner

Department of Industrial Engineering, Faculty of Engineering, Gaziantep University, 27310, Şehitkamil, Gaziantep, Turkey

Abstract: This study reviews and analyses the recent research and development and trends in the applications of wind energy and it also discusses and summarizes the topic. We show the usage and the influence of text mining on the different aspects of wind energy systems especially for hot topics and trends of wind energy area. Text mining provides the state of the art in this area that will be a good guidance for future research work. The main results achieved from the study have shown that the text mining technique are adequate for serving as a proof of concept and as a test-bed for deriving requirements for the development of more generally applicable text mining tools and services within wind energy science.

Keywords: Wind energy research, text mining, concept extraction, clustering.

1. INTRODUCTION

In the 2020 and 2030 climate-energy packages, the European Union committed to lower greenhouse gas emissions by 20% with respect to the rate 40% in 1990. By 2030 it plans to reach a share of renewable energy of 20% by 2020 and 27% by 2030. In regards of wind energy, there is now 128.8 GW of installed wind energy capacity in the EU: approximately 120. 6 GW onshore and just over 8 GW offshore. 11,791.4 MW of wind power capacity was installed in the EU-28 during 2014, an increase of 3.8% compared to 2013 installations. The European Union power sector continues its move away from fuel oil, coal and gas with each technology continuing to decommission more than it installs. The wind power capacity installed by the end of 2014 would, in a normal wind year, produce 284 TWh of electricity, enough to cover 10.2% of the EU's electricity consumption. Therefore, more variable renewable energy sources in the electricity system in a well-functioning energy system will require many changes not only in terms of new technologies (e.g. smart energy management systems, energy storage) but also in terms of infrastructures, interconnection between Members States, regulatory environment, harmonization of standards, and new business models from energy production to final consumption.

Massive growth on wind energy literature is the main reason for a need in a systematic approach to review wind energy publications. Currently, systematic reviewing is mostly performed manually and it has many problems. The main problem is the proliferation of textual information. While the quantity of potentially relevant literature expands by several thousand papers per week, it is obvious that no individual can read and manage them all [1, 2]. If a reviewer wants to search literature, they have been accustomed to sacrificing specificity in searches in order to ensure not having missed any relevant studies. Titles, abstracts or even full texts are downloaded and screened manually. This is the most time-consuming part of the process and can involve tens of thousands of publications. Complex systematic reviews can take more than a year to complete with up to half time being spent searching and screening hits. This is problematic because policy-makers and practitioners often need to know the state of research evidence over a much shorter timescale than current methods allow. It can lessen the likelihood that research evidence will be used at all, with consequential dangers for people affected by policies or practices developed in the absence of a firm evidence base [2]. Although more efforts are focusing on wind energy area for improving the performance, there is a need for a systematic review for available literature of wind energy applications. For that reason, in order to identify the relevant studies and its related frameworks, policies, future challenges and prospects in the worldwide, we used the text mining technique.

The organization of the study is as follows: In Section 2, the literature in the area of bibliometrics and text mining is reviewed. Section 3 describes details of the proposed text-mining framework. In Section 4, experimental results illustrate potential application of text mining and clustering technique and Anova used in Statistica 10.0 are presented. Finally, the conclusions are presented in Section 5.

*Address correspondence to this author at the Department of Industrial Engineering, Faculty of Engineering, Gaziantep University, 27310, Şehitkamil, Gaziantep, Turkey; Tel: +90 342 3172600; E-mail: eroglu@gantep.edu.tr, erogluyunus@hotmail.com

E-ISSN: 1929-6002/16

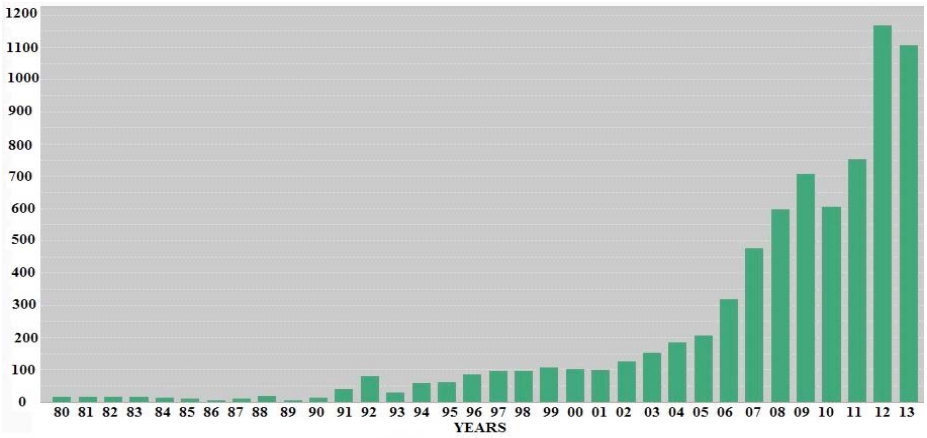

Figure 1: Published papers per year (Thomson Routers – ISI Web of Science, 2014).

2. LITERATURE

The literature review is conducted utilizing databases including Thomson Reuters – ISI Web of Science database. Figure **1** shows the statistics on wind energy publications between 1980 and 2013. It can be observed that there is certain increase in publications especially after 2001. The citations of wind energy oriented studies are over 12000 by 2014. Country/territory based publication counts until the end of 2013 show that USA has the most wind energy publication with 1353 publications followed by Peoples Republic of China with 783 publications. Table **1** presents Country/Territory rankings on number of publication all over the world until the end of the year 2013.

Text mining is a very useful methodology to develop a query for comprehensive information retrieval from published literature [3, 4]. It is also an important and fascinating area of modern analytics with its extremely iterative process [5]. Text mining is about inferring structure from sequences representing natural language text and may be defined as the process of extracting information from unstructured text data [6]. It can be used to identify technical themes in large unstructured text databases [7-12]. Text mining can also be used to enhance information retrieval [13, 14], to discuss potential discovery and innovation based on merging linkages between different working areas [15, 16], to uncover unexpected asymmetries from the literature [17, 18], to estimate global levels of effort in science and technology sub-disciplines [19-21], to help authors for increasing the citation of their works [3, 19, 20, 22], and to track myriad research impacts across time and application areas. An agent based text mining approach was investigated on financial news to assist the investors in deciding to buy and sell stocks in Taiwan market [23]. It has given a good example of gathering forecasts from financial newspapers texts in the field of agent based data mining.

With the large number of research articles in the literature, a manual review of the works would be tedious and time consuming [1, 5]. Thus, text mining can be used as a powerful tool to analyze current studies on a defined field. There are several studies on text mining application for literature analysis [1, 8, 12, 15, 24-40]. Delen and Crossland [31] investigated a survey analysis on management information field. They used text mining to identify trends on the topics and classify works that was presented in three major journals in the field of management information systems. Kostoff *et al*. [35] presented literature related to discovery method. They used links of different concepts which had not been linked before in order to produce novel, interesting, plausible, and intelligible knowledge. The evolution of literature related to discovery method was discussed in a recent study by Kostoff [38]. Ananiadou *et al*. [1] proposed a supporting systematic review using text mining methodology. They described how using text mining could produce a systematic literature review. Emerging topic detection is another field that considers literature review by text mining. Tu and Seng [39] presented a novel strategy to detect emerging topics in literature and investigated novelty index and published volume index. There are a few works on energy related text mining studies in the literature. Kostoff *et al*. [21] have prepared a report on science and technology for electric power sources by using database tomography. They suggested the combination of text mining and bibliometrics to extract useful information from large volumes of technical text. They also studied on deriving technical intelligence from a Power Sources databases derived from the Science Citation Index [27]. Yoon [41] investigated a

Table 1: Country/Territory Rankings on Number of Publication all over the World until the End of the Year 2013

Country/Territory	Publication Numbers	% of 7145 (Percentage of total Number of Publication)
USA	1353	18,94
A*	980	13,72
Peoples R China	783	10,96
B**	665	9,31
Canada	446	6,24
Germany	423	5,92
England	359	5,02
India	356	4,98
Spain	295	4,13
C***	285	3,99
Turkey	259	3,62
Japan	209	2,93
Denmark	198	2,77
France	188	2,63
Italy	162	2,27
Australia	160	2,24
Greece	149	2,09
Netherlands	145	2,03
Scotland	114	1,60
Iran	111	1,55
Romania	105	1,47
Portugal	102	1,43
Egypt	100	1,40

*A = countries which have number of publications between 100 and 50.
**B = countries which have number of publications between 50 and 10.
***C = countries which have number of publications less than 10.

diffusion approach to detect weak signals for long-term business opportunities for solar – cells by using text mining. On the wind energy area, there is only one study in the literature which aims to generate a domain ontology for wind energy field [40]. Thus, this study can be used to summarize the current wind energy literature as a whole and also be used as a tool to discuss hot topics in the field.

3. METHODOLOGY

The role of text mining is to investigate a systematic review on wind energy researches by analyzing huge number of papers, which consist of unstructured text data. Search and screening strategy of systematic reviewing for the domain of wind energy issues have been discussed below. Text Miner and Data Miner extensions of Statistica 10.0 have been used in the proposed systematic review for wind energy and there are approximately five major technical categories in the text mining process: Database Selection, Data Selection, Data Preparation, Analyzing, and Extracting Knowledge.

- *Database Selection:* Extensive wind energy searches are carried out in order to locate as many relevant researches as possible according to a query. These searches include electronic databases to collect and store published literature. Thomson Reuters – Web of Science database is a most widely-used electronic database by researchers and academicians on the World Wide Web.

- *Data Selection*: A query-based mining is applied for the topics including “Wind Energy” to select relevant studies. Collecting only relevant

Table 2: An Example Part of Stored Data

Type	Authors	Title	Source	Year	Abstract	Citations	Territory
J	Muljadi, E; Hess, HL; Thomas, K	Zero sequence method for energy recovery from…	IEEE TRANSACTIONS ON ENERGY CONVERSION	2001	An innovative power conversion system to convert energy from a variable-frequency…	6	USA
J	Mays, I	WREC 1996 – The status and prospects for wind energy…	RENEWABLE ENERGY	1996	This paper reviews developments in the Weibull distribution of wind energy in Europe over the past decade,….	1	England

documents provides highlight key evidence. Thus, reviews, editorials, news, patents, books, case results are eliminated from the search results. Published papers and conference proceedings are selected and this narrows the collected data to prepare a specific review.

- *Data Preparation*: Selected data is categorized by their authors' country/territories and stored in an Excel Sheet. The sheet contains following data about studies; Publication Type: Journal or Proceeding; Authors' Name; Title; Source: Published Journal or Conference; Publication Year; Authors' country/territories; Abstract; Citation Number. Studies having more than one author, who are from different country/territory may occur more than one times. These studies are categorized as Multinational in Country/Territory column. This correlates evidence from a plethora of resources and summarizes the results and creates meaningful contents in wind energy literatures.

- *Analyzing*: Text Miner and Data Miner extensions of Statistica 10.0 are used on prepared data. Section 3.2 gives details of Text Mining Technique.

- *Extracting Knowledge*: Text Mining summarizes the statistics and data for collected studies. Also, clustering analyses and ANOVA (analyses of variances) help to extract useful knowledge.

3.1. Data Gathering

Statistica 10.0 Text miner contains numerous options for accessing text documents in different formats, including.txt (text),.pdf (Adobe),.html,.xml (Web-formats), and Microsoft Office formats (e.g. .doc, .rtf). The program supports full "Web-crawling" capabilities, so that documents can be extracted from the Web, starting at a particular root Web page (URL). All documents linked to that particular page will be included, as well as the documents linked to those sub-documents, and so on, up to a user-specified level or depth. File names and URLs can also be stored in text variables, in *STATISTICA* data files. In this manner, the program can not only process actual text stored in text variables, but it also properly interprets references to text documents or URLs. Thus, numeric information and textual information (large documents) can be stored on a per-case (observation) basis and meaningful analyses can be performed on data files where for each observation numeric as well as unstructured textual information is available.

All of the available abstracts of journal papers and conference proceedings were collected from the year 1990 to end of 2013. The database is constructed with the Publication type, Authors' name, title, Source, Publication year, Authors' country/territory, abstract, and citation numbers of selected papers. Irrelevant editorial notes, research notes, patents, news, and reviews are not selected for database to include only relevant studies. The collected data is categorized by their authors' country/territories and stored in an Excel Sheet as shown in Table **2**.

Country/Territories of publications are classified as four groups according to the number of publications: more than 100, between 100 and 50, between 50 and 10, and less than 10. The first group is stored with their country/territory name; rest of them is stored with their class name in the data sheet respectively. Also, a multinational category is created for international collaborative publications in Country/Territory column. After collection and preparation of data, database is ready for Text Mining.

3.2. Text Mining

Text mining is a burgeoning methodology to semi-automatically extract information from unstructured text data and involves imposing structure upon text so that relevant information can be extracted from it [5, 31, 42, 43]. Statistica 10.0 is reported as one of the high satisfaction rated text mining software [44]. It has Text Miner and Data Miner tools as an optional extension in Predictive Analytics Solutions. Text mining has been applied for wind energy literature as follows;

- *Creating a database*: A database is created in an Excel Sheet and imported to a Statistica Sheet.

- *Selecting text variable to be mined*: Abstracts of the publications and proceedings are selected as variable.

- *Setting text mining stemming language parameters*: Parameter settings of stemming language of text mining have been shown in Figure **2**. Stemming language was English due to all abstracts are in English.

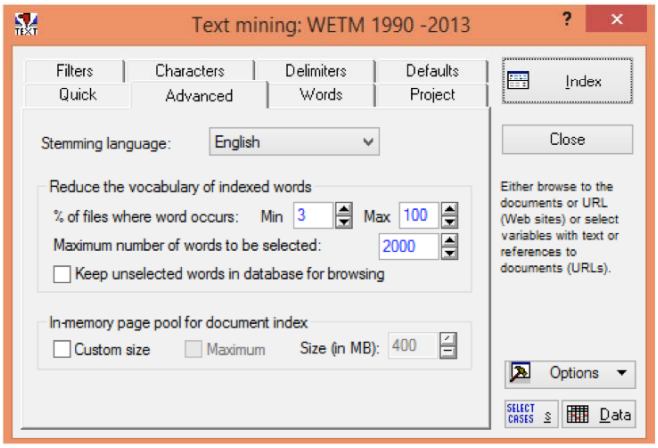

Figure 2: Stemming language parameters of text mining.

- *Giving the list of stopping words, phrases, and synonyms*: Stopping word list is the list of words which is not needed selecting for further analysis. A Standard English Stop List consists of pronouns, auxiliaries, articles, prepositions, conjunctions, adverbs, and some common words such as "now, never, between, into, etc." All word lists are unnecessary for academic literature text mining process. Besides, an academic abstract has some common words such as abstract, address, article, author, copyright, Elsevier, energy, ltd, paper, reserved rights, study, wind, wind energy, wind-energy and work. These are added to stop list as unnecessary items. The phrases' list is the list of phrases, which contains joint words such as converting system, variable speed, renewable energy, weibull distribution, three phases, wind power and wind turbine Also, some words have similar meanings/synonyms and they have to be considered in only one word (for instance; predict-forecast-estimate, these all three words combined as the word "forecast").

- *Setting the word processing/filtering parameters*: Figure **3** shows the parameter settings of word processing of our study.

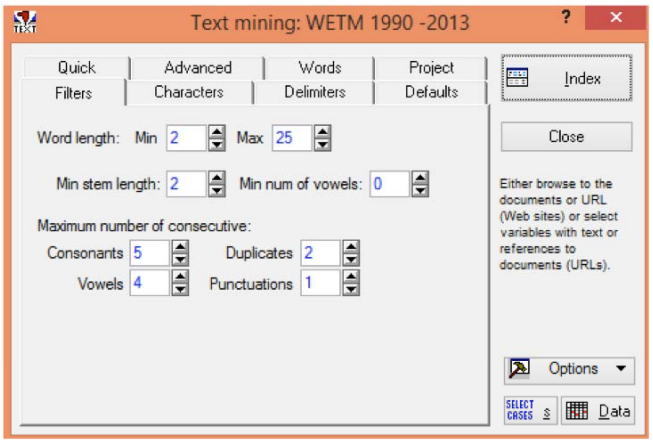

Figure 3: Word processing/filtering parameters of text mining.

- *Starting to index words*: Index button is used for indexing/selecting or filtering the words. Indexing process is shown in Figure **4** there were 7145 documents in the database, 473 words were selected, and 18954 words were not selected. Most frequent ten words are shown in Table **3**. "*System*" is the most repetitive word through the whole publications which is used 10792 times in a total of 4062 documents. The second one is "*power"* with 9542 repetitions in 3763 documents, followed by the words "*generation*", "*control*", and "*model*" and etc.

- *Concept Extraction*: There are four Text Mining methods to extract concepts: Raw Statistics, Binary Frequency, Logarithmic frequency and Inverse Document Frequency:

  - *Raw Statistics:* Gives only counts of words in all documents

  - *Binary Frequency:* Returns "1" for a word if it occurs in any document else "0"

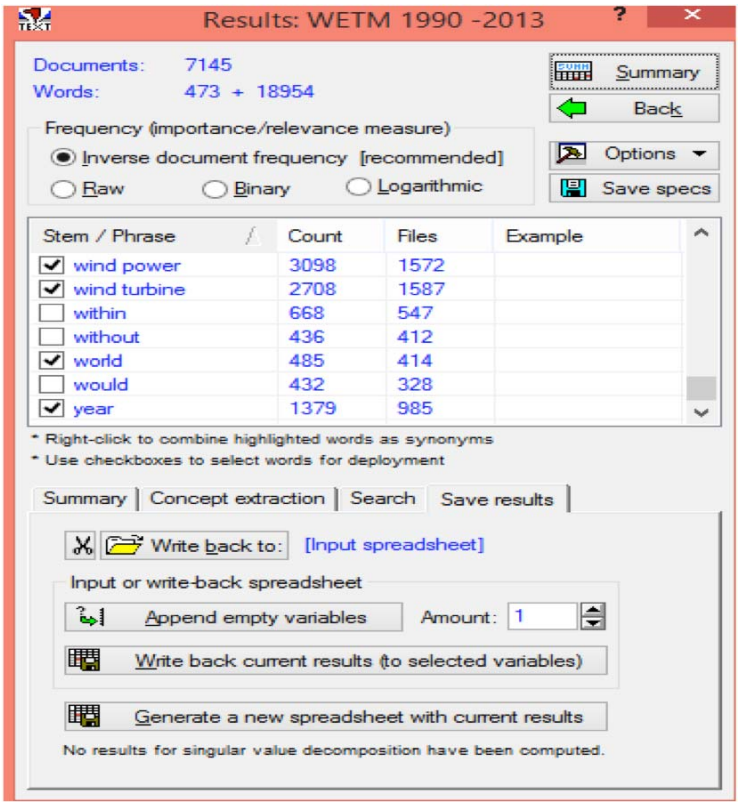

Figure 4: Word processing/filtering parameters of text mining.

Table 3: The Most Frequent Words

Word	Count	Number of Documents	Word	Count	Number of Documents
1. system	10792	4067	6. speed	4496	2081
2. power	9542	3763	7. result	4226	3247
3. generation	7914	3650	8. develop	4064	2432
4. control	6813	2128	9. electric	3958	2157
5. model	5618	2543	10. present	3625	2787

- *Logarithmic Frequency:* Computes frequency for a word by using formula as given in Equation 1:

$$F = 1 + \log(wrf) \quad for \quad wrf > 0 \tag{1}$$

where "*wrf*" stands for word raw frequency. If a word occurs "1" time in document *A* but "3" times in document *B*, then it is not necessarily reasonable to conclude that this word is "3" times as important as descriptor of document *B* as compared to *A*. Thus, logarithmic frequency is useful in these situations.

- *Inverse Document Frequency:* This is the relative document frequencies (*df*) of different words. A common and very useful transformation that reflects both the specificity of words (document frequencies) as well as the overall frequency of their occurrences (word frequencies) is the so-called inverse document frequency (for the *i'th* word and *j'th* document). The formulation of inverse document frequency (*idf*) is given in Equation 2, where *N* is the total number of documents, *wf* is the word frequency for all documents, *df* is the word frequency for current document;

$$idf(i,j) = \begin{cases} 0 & if\ wf_i = 0 \\ (1 + \log(wf_i)) \log \dfrac{N}{df_i} & if\ wf_i \geq 1 \end{cases} \tag{2}$$

In this study, Inverse Document Frequency has been used to extract concepts and to determine words' importance. Figure **5** gives the Scree Plot of extracted concepts that is used to decide on the number of singular values that are useful and informative, and that should be retained for subsequent analyses. Usually, the number of "informative" dimensions to retain for subsequent analysis is determined by locating the "elbow" in this plot. The points, which are above from elbow of the graph, have more importance than others [5]. In our study, there are 24 concepts extracted and the first three concepts have more importance than others because the elbow starts from the fourth concept.

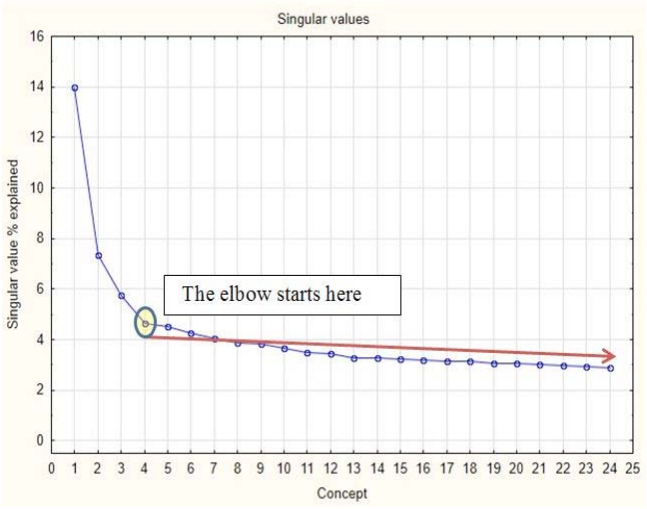

Figure 5: Scree Plot of extracted concepts.

Text mining process transforms unstructured text data into structured and analyzable data. All results from text mining process are stored in database. We applied some statistical analysis methods to the numeric summaries representing the texts. Clustering technique has been applied to identify clusters of similar documents and we also used Anova test in order to analyze the differences between group means and their associated procedures.

4. EXPERIMENTAL RESULTS

This section focuses on developing a *concept grouping* to construct a thesaurus from synonymous terms. For the purpose, a "word" is a sequence of letters separated by spaces, a "phrase" consists of one or more words, and a "term" is a phrase, which is extracted from title or abstract. Table **4** shows some promising results from the data set relevance to the importance value of indexed words. If the words in Tables **3** and **4** are compared, it is seen that the most important words are not totally the same with the most repetitive (frequent) words. It is important that text mining process does not only give the frequent words, it also selects the important ones from frequents.

According to the inverse document frequency analysis, as we see from Table **4**, control word is the most important word in control of wind energy systems. Also such as offshore, solar, blade, voltage, wind farm, storage, WEC, DFIG and forecast are top ten in the list. All have caption numbers according to their importance levels and also words are in the importance order as seen in the list. Results showed that the most important concept is related with *System Control Models* consequently including following words: *control*, *model*, *power*, *speed*, *system*, *forecast*, and *electric*.

Also we defined the most important words of extracted concepts. There are twenty-four concepts, which are extracted from text mining analysis (see Figure **5**). All concepts can be titled according to important words included in their lists. As seen in Table **5**, System control models as the first concept supports that developing new algorithms for control and management of energy systems, electric storage devices will be vital studies in the future.

Researches may probably involve modeling of electrical loads; power forecasting, development and testing of energy control systems. Also, electrical control systems, data observation in wind farm sites, storage of energy, solar systems, electrical systems,

Table 4: The Most Important Ten Words

Words	Importance (%)	Words	Importance (%)
1. Control	100	6. Wind farm	81,490206
2. Offshore	93,880264	7. Storage	80,12595
3. Solar	86,830387	8. WEC	76,725528
4. Blade	84,827511	9. DFIG	75,564418
5. Voltage	82,696561	10. Forecast	75,176318

Table 5: The Most Important Words of Extracted Concepts

Concept 1	**Concept 2**	**Concept 3**	**Concept 4**	**Concept 5**	**Concept 6**
System Control Models	***Electrical Control Systems***	***Data Observation***	***Storage Systems***	***Solar Systems***	***Electrical Systems***
control	control	Speed	storage	solar	DFIG
model	voltage	Data	hour	magnet	voltage
power	convert	Observe	hybrid	heat	reactive
speed	induct	Measure	battery	fuel	induct
system	DFIG	Forecast	distribute	flux	grid
forecast	WEC	Surface	month	temperature	FED
electric	reactive	Density	speed	field	site
Concept 7	**Concept 8**	**Concept 9**	**Concept 10**	**Concept 11**	**Concept 12**
Network-Grid Systems	***Renewable Energy Policies***	**Turbine Components**	**Solar Hybrid Systems**	**Aerodynamics**	**Offshore Systems**
network	policy	DFIG	hybrid	Blade	offshore
penetrate	forecast	Rotor	solar	Turbulence	control
integrate	track	Induct	DFIG	fluctuate	wave
grid	algorithm	Blade	battery	renewable energy	water
model	uncertainty	FED	photovoltaic	change	sea
stability	approach	Double	induct	wind turbine	strategy
storage	renewable energy	Stator	data	source	surface
Concept 13	**Concept 14**	**Concept 15**	**Concept 16**	**Concept 17**	**Concept 18**
***NWD**	***NWD**	**Market Researches**	***NWD**	**Electric Distribution Systems**	***NWD**
battery	monitor	Price	forecast	distribute	monitor
storage	offshore	Market	market	wave	rotor
market	structure	Emission	accuracy	load	algorithm
project	solar	Load	storage	rotor	reliable
policy	year	Frequency	wind power	voltage	cost
solar	industry	Forecast	machine	density	optimum
support	blade	Voltage	flux	probably	hybrid
Concept 19	**Concept 20**	**Concept 21**	**Concept 22**	**Concept 23**	**Concept 24**
***NWD**	***NWD**	***NWD**	**Island Plants**	**Climate Researches**	***NWD**
monitor	WEC	Heat	plant	change	wave
fault	machine	Storage	island	temperature	machine
measure	reliable	Water	station	climate	device
WEC	sea	Temperature	area	machine	rate
convert	change	Frequency	region	heat	frequency
layer	transmission	Air	magnet	DC	communicate
install	wave	Impact	water	capacity	hybrid

*Not well determined.

network-grid systems, renewable energy policies, turbine components, solar hybrid systems, aerodynamics, offshore systems etc. are concepts, which define problem environments and popular hot topics for wind energy. The scree plot (see Figure **5**) tells that first three concepts (System Control Models,

Electrical Control Systems and Data Observation) are more important than the others. All these concepts are not clearly identified by only looking up its important word lists. Especially after ten concepts, the important word list consists of directly irrelevant terms such as thirteenth concept: *battery*, *storage*, *market*, *project policy*, *solar*, and *support*. Thus, some concepts are labeled as *not well determined (concepts 13, 14, 16, 18, 19, 20, 21, and 24)*.

Similarity among terms in a set of documents receives much attention from bibliometrics and scientometrics communities. This study also focuses on terms' relationships to understand the cognitive structure of research domains. The ultimate aim is to provide useful intelligence in support of wind energy management. The text mining process starts by considering an occurrence matrix of documents by terms. If terms occur together more often in documents, these could reflect a strong relationship. K-means clustering algorithm has been applied to word importance values and six clusters have been found. Text mining analysis results gave four-hundred-seventy-three indexed words. Table **6** lists the members of clusters. As the words are clustered as their importance values, the rank of clusters also represents its importance rank. Furthermore, the rank of words gives the importance level of the term in the cluster. The most important seven words are listed in the first cluster. This cluster clarifies that these words have exactly more different importance levels than others. As the importance value of words decreases, the number of members of clusters increases. However, the terms in the sixth cluster have the least attention by wind energy researchers. The first cluster terms; control, offshore, solar, blade, voltage, wind farm, and storage deserve to be takeninto consideration.

As another further analysis, Anova statistical procedure has been applied. We analyzed the differences between group means and their associated

Table 6: Term Clusters Generated from K-Means Clustering

Cluster 1	Control, Offshore, Solar, Blade, Voltage, Wind Farm, Storage
Cluster 2	WEC, DFIG, forecast, speed, convert, wave, turbine, rotor, magnet, load distribute, grid, data, cost, model, optimum, induct, wind power
Cluster 3	wind turbine, source, renewable energy, method, potential, site, fault, measure, heat, plant, PMSG, battery, develop, frequency, fuel, reactive, surface, electric, generate, flow, power structure, permanent, market, operate, simulate, capacity, density, policy, hybrid, emission turbulence, water, strategy, project, machine, technic, stator, maximum, region, field, resource, current, active, design, convers, area, propos, impact, year, system
Cluster 4	Increase, perform, mw, effect, new, change, product, differ, problem, period, sea, flux, base, demand, approach, synchrony, parameter, height, gas, test, observe, invert, function, penetrate, present, price, plan, track, stability, velocity, DC, assess, connect, reduce, result, economy, fed, integrate, month, fluctuate, algorithm, annual, network, torque, local, analysis, efficient, doubly, accuracy, thermal, environ, climate, fossil, time, process, output, rate, station, comparison, grow, install, temperature, high, locate, variable, dynamic, layer, scheme, country, show, reliable, transmission, supply, hour, evaluate, monitor, probable, renew, value, require, industry, mean, apply, condition, input, compensative
Cluster 5	Average, uncertainty, state, quality, manage, first, factor, consider, drive, combine, scenariocycle, import, low, season, transient, remote, higher, direct, future, statist, select, public, mode, experiment, calculate, include, numeric, cause, possible, found, relate, nonlinear, risk, world, benefit, large, profile, side, small, challenge, disturb, provide, produce, force, determine, well, three, island, obtain, convent, main, number, inform, significant, harmony, similar, device, air, matlab, atmosphere, error, global, size, stochastic, robust, point, case, vector,meteorology, govern, valid, regular, utility, invest, switch, nation, depend, two, support, nature, photovoltaic, alternative, associate, level, one, investigate, module, rotate, electron, kw, unit, order, compute, discuss, intermit, improve, phase, mechanism, implement, many, additive, research, need, componentcapable, angle, give, characteristic, variety, achieve, maintenance, degree, exist, indication, avail, rang, scale, part, framework, total, loss, commune, weather, reduction, north, variable speed, solution, sustain, response consumption, limit, engine
Cluster 6	Feasible, examine, various, target, type, consist, represent, enhance, form, recent, identify, experiment, allow, carry, build, demonstration, contribute, role, final, posit, concept, rapid, expect, isolate, program, tool, couple, decrease, derive, capture, set, specify, lead, curve, influence, minim, near, conduct, accord, sever, sector, lower, ratio, extract, balance, suggest, option, general, typical, day, Simulink, concern account, interact, line, adapt, transfer, detail, less, occur, particular, find, verify, standard, linear, take, link, continue, interest, term, configure, constraint, affect, advantage, correlation, theory, meet, complex, introduce, suitable, variable, solve, element, real, construct, strong, amount, major, key, constant, issue, object, intern, consider, serial, respect, focus, aim, collect, describe, commercial, second, software, publish, critic, initial, toward, actual, last, enable, procedure, transform, example, involve, offer, behavior, large-scale, place, feature, appropriate, especial, maintain, explore, reach, fast, close, larger, four, regard, often, practice, follow, great, goal promise, employ, best, success, previous, deal, advance, space, aspect, tradition, report, exploit, short, better, long, reason, wide, around, illustrate, necessary, create, approximate, common, science, situate, single, simple, purpose, correspond, avoid, complete, review, help, much, principle, consequent, establish, play, good, adopt, way, ensure, equip, incorporate.

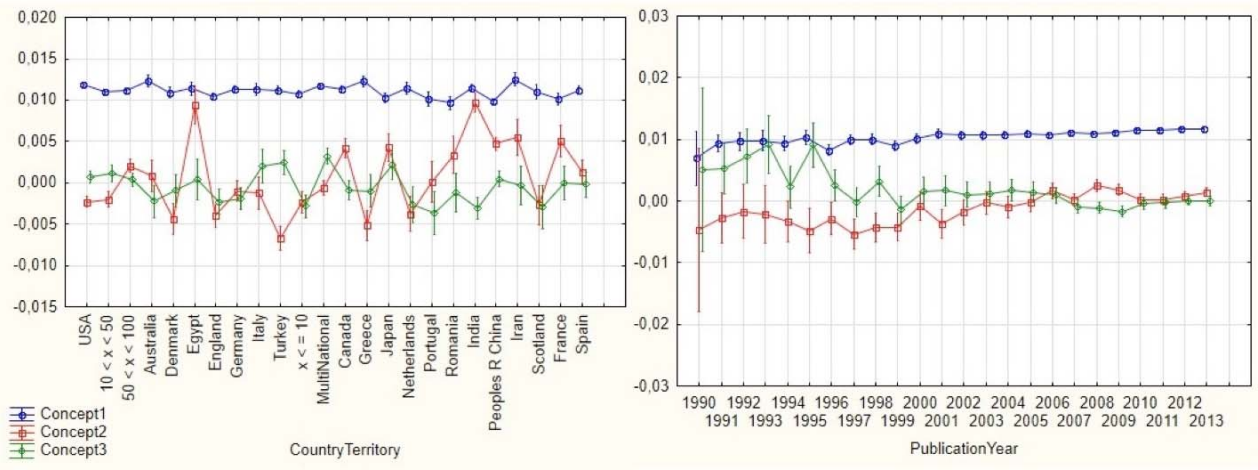

Figure 6: Anova for the three concepts *vs*. Country/Territory and Publication Year.

procedures. It is used to explain wind energy research trend differences between countries/territories and years. Level of significance in all results was less than 0.01 and vertical bars of Anova graphs denote 95 % confidence intervals. Figure **6** can light the way for new technologies of wind energy systems. Anova analysis was used for all extracted concepts versus countries and Publication Years. One can conclude on countries research/interest areas and trends of hot topics. Due to importance level, Anova test was used for only System Control Models (Concept 1), Electrical Control Systems (Concept 2), and Data Observation (Concept 3) according to countries/territories and publication years. Concept 1 has the highest mean of importance value; it has almost no significant variants according to countries. This means that researches on *System Control Models* have nearly the same importance all over the world. For concept 2, we can conclude that Egypt, India, Iran, France, Japan, Canada, and China give more attention to Concept 2. In contrast, Turkey, Greece, Denmark, England, and Netherlands give less attention to concept 2. While the Concept 1 and 2 have an increasing importance, *Data Observation (*Concept 3) studies lose its importance especially after 1995. This can be explained by the improvement of the energy market. Today, many countries have wind farms. They are mostly interested in improving productivity, efficiency, and profitability of current wind energy market. Therefore, decrease on the interest of feasibility studies, which includes *Data Observation,* should be considered to be normal. However, Italy, Turkey, Japan, and Multinational oriented publications are more interested in *Data Observation* studies are more than other countries/territories.

The most important word *control* has an increasing trend. India, Iran, China, Egypt, and France are mostly interested in *control* studies. On the other hand; Turkey, Denmark, England, Greece, and Scotland gives less attention to *control* studies (see Figure **7**). This word occurs both in the first and second concepts. Thus, its importance is clear through all wind energy publications. Technological improvements on equipment of wind energy sector (sensors, blades, brake systems, etc.) lead to increase on *control*-focused researches as a result of the more sensitive unites, the more *control*. Thus, it can be forecasted from the positive trend on *control* studies that wind energy technology has still positive trend. *Offshore* wind energy plants are preferred to on-land plants because of some advantages; It has better wind speeds available, Huge wind turbines can be constructed *offshore* so electricity supplied per turbine is higher than on-land, Usually construction is weaker. Figure **8** shows that *offshore* is one of the most trending topics of wind energy publications. Countries such as Denmark, Netherlands, and Germany, which have offshore wind energy plants, give more importance on *offshore* studies. This figure supports that if one country has technology, it gives more attention to its problems. Also, it is clear that researches on offshore technologies have been increasing since 1996.

However, solar word is the third important word; its importance has been decreasing drastically since 2005. But, researches on solar are increasing in Turkey, Multinational publications, Category C countries (having less than 10 publications in total), and Category B countries (having publications between 10 and 50 in total) (see Figure **9**). Blade related publications have in a steady state condition. Japan and USA have many blade researches (see Figure **10**). Because of having leading technologies all over the

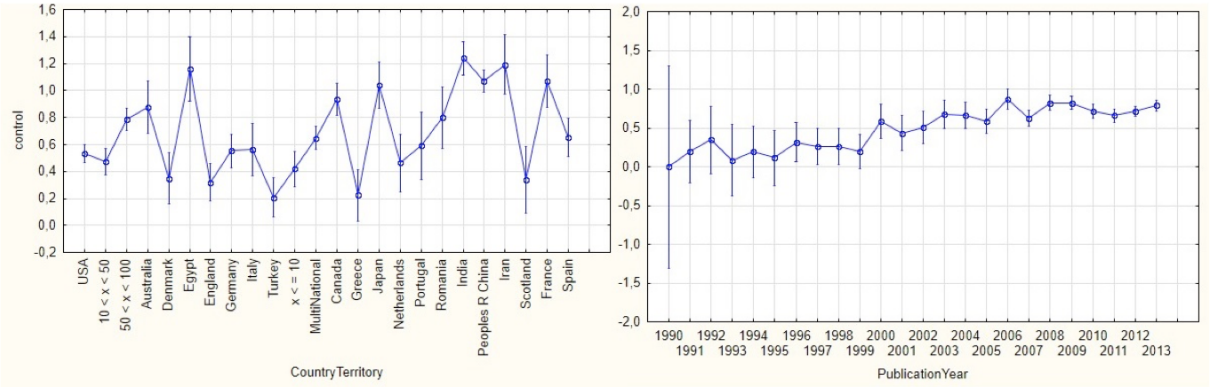

Figure 7: Anova for the word *control vs.* Country/Territory and Publication Year.

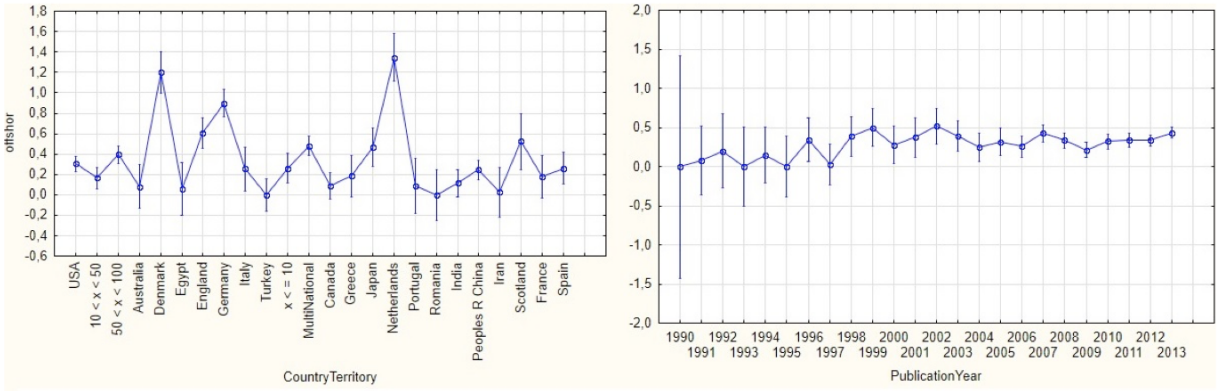

Figure 8: Anova for the word *offshore vs.* Country/Territory and Publication Year.

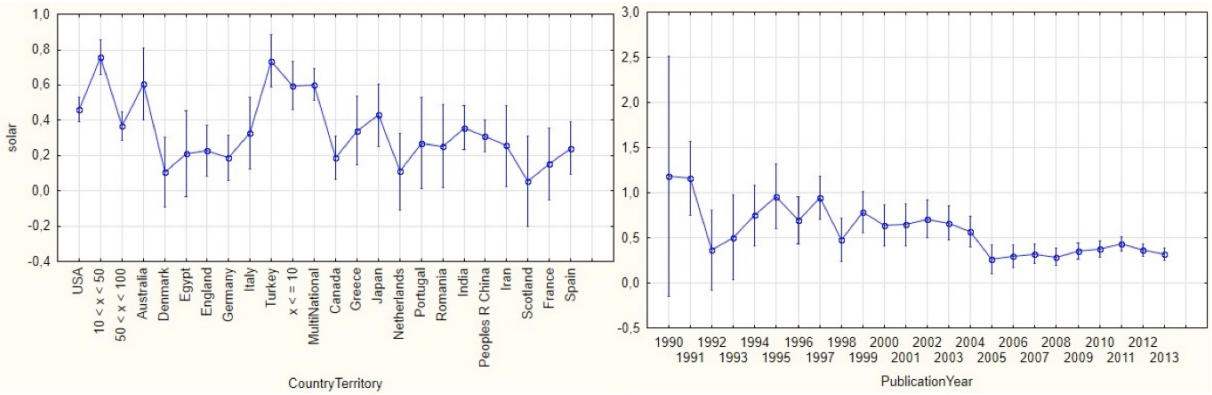

Figure 9: Anova for the word *solar vs.* Country/Territory and Publication Year.

world, it is not surprising that Japan and USA are interested in wind turbine blade technology. Aerospace industry is one of the leading sectors of England and the results explain why their interest is in blade technology. Egypt and India have significantly different interests in voltage related studies. This can be explained by having incapable or weak electrical transformation and network devices. Because the demand of electricity increases dramatically, this topic has also an increasing importance especially after 2000 (Figure **11**). Understanding of the physics of wind as a primary resource and wind energy technology must be improved. Especially for an improved design of large-scale wind rotors, a better understanding of the underlying physics is needed.

Wind turbine technology and needs for renewable energy sources grow annually. This causes increase in installed *wind farms* all over the world. A *wind farm* may also be located offshore. Especially, Spain and Scotland have many *wind farm* studies (see Figure **12**). Because today's technology is incapable to *store* wind

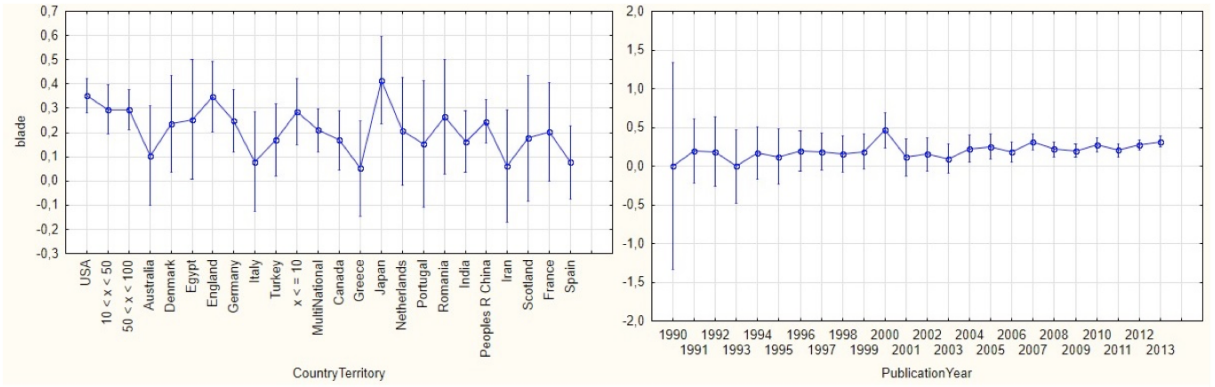

Figure 10: Anova for the word *blade vs.* Country/Territory and Publication Year.

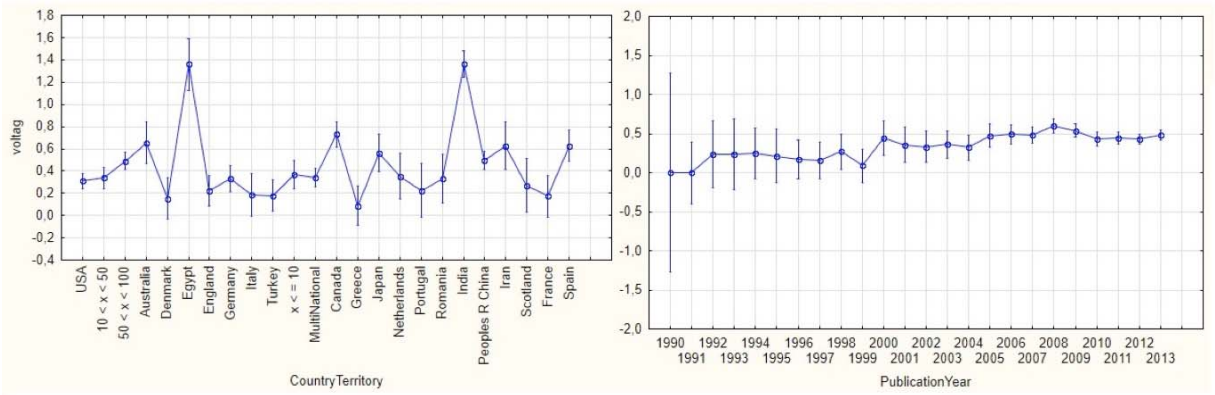

Figure 11: Anova for the word *voltage vs*. Country/Territory and Publication Year.

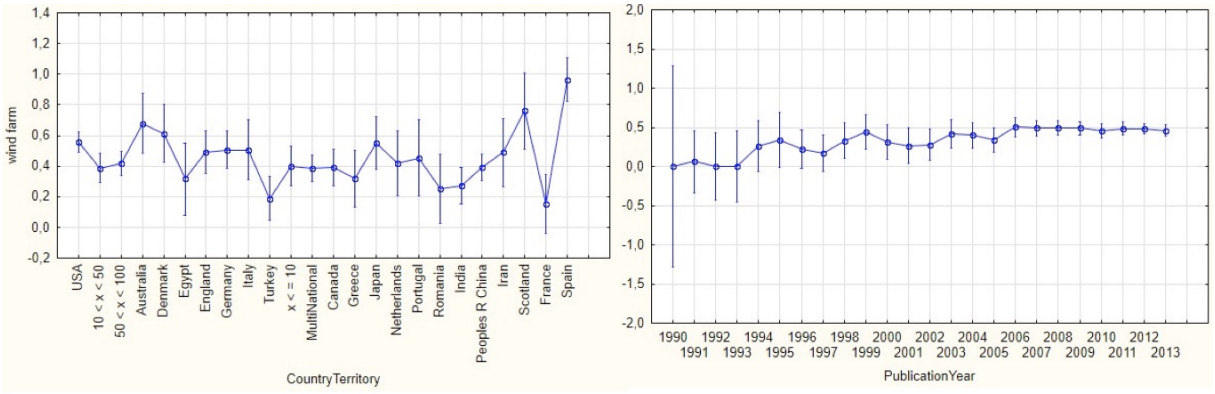

Figure 12: Anova for the word *wind farm vs.* Country/Territory and Publication Year.

energy, one of the main problems on renewable energy plants is that once you produce energy, you have to use it. So, *storage* researches are one of the hot topics. Australia, Canada, and Germany have many publications about *storage* (see Figure **13**).

Figure **14** shows an interesting story about citation numbers. If the publication is prepared with a collaboration of more than one country, it may have more citations. Also, if researcher's origin is England, Turkey, or Greece, the citations are higher than the others. Especially, citations, in 1995, 2000, and 2004, are more than other years.

Following results can be summarized from Anova analysis:

- If a country has offshore wind energy facilities, offshore-related studies are important (Denmark, Netherlands, and Germany).
- Wind turbine manufacturers are interested in new wind energy technology to produce more

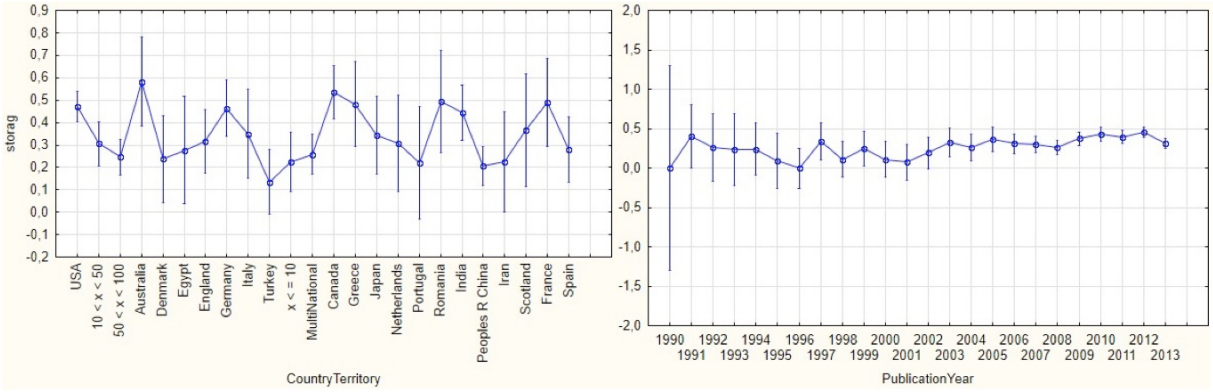

Figure 13: Anova for the word *storage vs*. Country/Territory and Publication Year.

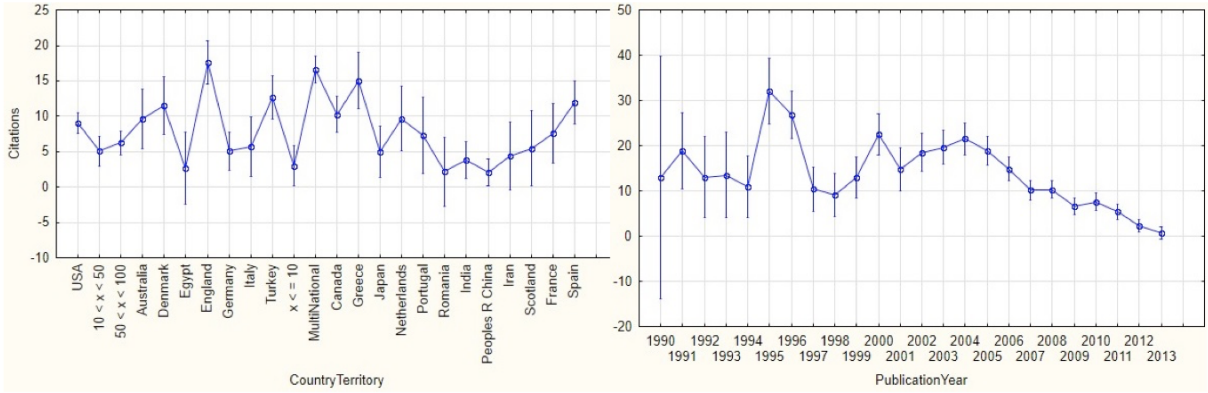

Figure 14: Anova for citation numbers *vs.* Country/Territory and Publication Year.

energy. Developed countries such as Japan and the USA give more attention to *blade* technology, which has similarities with aerospace industry. Especially, England is a leader country for blade technology.

- Citations are related with authors' origin.
- Collaborative studies are cited more than others

5. CONCLUSION

This study reviews and analyses the recent research and development and trends in the applications of wind energy and it also discusses and summarizes the topic. 7145 publications between the publication years from 1990 to the end of 2013 have been stored from a most- commonly-used web based database, Thomson Reuter - ISI Web of Knowledge. Wind is a key low-carbon technology. We have learned from this study that key challenges to tackle are to reduce further cost of wind energy, especially for offshore, and to further increase wind energy level of deployment. Cost reductions are to be achieved in the wide range of technologies and components that compose an offshore wind system: new advanced turbines and substructures as well as installation, operation, and maintenance and decommissioning. To further increase wind energy deployment possibilities, research is still also needed to reduce the environmental and social impact and to increase the public acceptance of on-shore wind. Control strategies and systems for new and/or large rotors and wind farms (on- and offshore), new innovative substructure concepts, including floating platforms, to reduce production, installation and O&M costs for water depths and demonstrating and testing of new nacelle and rotor prototypes and control strategies and system can be studied in the upcoming years. The proposed text mining framework for wind energy studies follows the steps: The process starts with retrieving the relevant documents from the appropriate databases. Then the data is extracted and cleaned to remove noises and errors. This cleaned data is then fed to the analysis process. This study adds to this body of knowledge by implementing a text clustering process. The last step is the representation/visualization of the results. The underlying motivation driving the research is to create a text-mining framework that can extract wind energy challenges from electronic text sources. This knowledge is a prime requirement for successful wind

energy management. The text-mining framework can help:

- identify technology infrastructure of wind energy that are the main players in R&D concerning a target technology)
- discover overlapping or similar research activities among countries
- identify and categorize the main research areas and sub-areas in a large body of technical literature of wind energy
- identify emerging technologies from related or disparate technical literature.

REFERENCES

[1] Ananiadou S, Rea B, Okazaki N, Procter R, Thomas J. Supporting systematic reviews using text mining. Soc Sci Comput Rev 2009; 27: 509-23. http://dx.doi.org/10.1177/0894439309332293

[2] Chalmers I. Trying to do more good than harm in policy and practice: the role of rigorous, transparent, up-to-date evaluations. Ann Am Acad Pol Soc Sci 2003; 589: 22-40. http://dx.doi.org/10.1177/0002716203254762

[3] Kostoff RN, Tshiteya R, Pfeil KM, Humenik JA, Karypis G. Science and Technology Text Mining: Electric Power Sources 2004. http://dx.doi.org/10.1016/S0010-9452(08)70885-2

[4] Kostoff RN, Buchtel HA, Andrews J, Pfeil KM. The hidden structure of neuropsychology: text mining of the Journal Cortex: 1991-2001. Cortex 2005; 41: 103-15.

[5] Miner G, Elder J, Hill T, Nisbet R, Delen D, Fast A. Practical Text Mining and Statistical Analysis for Non-structured Text Data Applications. 1st ed. Academic Press 2012.

[6] Witten IH. Adaptive text mining: inferring structure from sequences. J Discrete Algorithms 2004; 2: 137-59. http://dx.doi.org/10.1016/S1570-8667(03)00084-4

[7] Hearst MA. Untangling text data mining. In: Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics on Computational Linguistics [Internet]. Stroudsburg, PA, USA: Association for Computational Linguistics; 1999 [cited2014 Mar 13]. p. 3-10. http://dx.doi.org/10.3115/1034678.1034679

[8] Losiewicz P, Oard DW, Kostoff RN. Textual data mining to support science and technology management. J Intell Inf Syst 2000; 15: 99-119. http://dx.doi.org/10.1023/A:1008777222412

[9] Zhu D, Porter AL. Automated extraction and visualization of information for technological intelligence and forecasting. Technol Forecast Soc Change 2002; 69: 495-506. http://dx.doi.org/10.1016/S0040-1625(01)00157-3

[10] Drake M. Encyclopedia of library and information science. 2nd Ed. CRC Press 2003.

[11] Ghazinoory S, Ameri F, Farnoodi S. An application of the text mining approach to select technology centers of excellence. Technol Forecast Soc Change 2013; 80: 918-31. http://dx.doi.org/10.1016/j.techfore.2012.09.001

[12] Jun S, Park S-S, Jang D-S. Document clustering method using dimension reduction and support vector clustering to overcome sparseness. Expert Syst Appl 2014; 41: 3204-12. http://dx.doi.org/10.1016/j.eswa.2013.11.018

[13] Kostoff RN, Eberhart HJ, Toothman DR. Database tomography for information retrieval. J Inf Sci 1997; 23: 301-11. http://dx.doi.org/10.1177/016555159702300404

[14] Greengrass E. Information Retrieval: A Survey [Internet]. University of Maryland; 2000 [cited 2014 Mar 13]. 224 p. Available from: http://www.csee.umbc.edu/csee/research/cadip/readings/IR.report.120600.book.pdf

[15] Swanson DR, Smalheiser NR. An interactive system for finding complementary literatures: a stimulus to scientific discovery. Artif Intell 1997; 91: 183-203.

[16] Shavinina LV. The International Handbook on Innovation. Elsevier 2003; p.1202.

[17] Goldman JA, Chu WW, Parker DS, Goldman RM. Term domain distribution analysis: a data mining tool for text databases. Methods Inf Med 1999; 38: 96-101.

[18] Kostoff RN. Bilateral asymmetry prediction. Med Hypotheses 2003; 61: 265-6. http://dx.doi.org/10.1016/S0306-9877(03)00167-1

[19] Kostoff RN, Green KA, Toothman DR, Humenik JA. Database tomography applied to an aircraft science and technology investment strategy. J Aircr 2000; 37: 727-30. http://dx.doi.org/10.2514/2.2659

[20] Viator JA, Pestorius FM. Investigating trends in acoustics research from 1970–1999. J Acoust Soc Am 2001; 109: 1779-83. http://dx.doi.org/10.1121/1.1366711

[21] Kostoff RN, Shlesinger MF, Malpohl G. Fractals text mining using bibliometrics and database tomography. Fractals 2004; 12: 1-16. http://dx.doi.org/10.1142/S0218348X04002343

[22] Kostoff RN, Shlesinger MF, Tshiteya R. Nonlinear dynamics text mining using bibliometrics and database tomography. Int J Bifurc Chaos 2004; 14: 61-92. http://dx.doi.org/10.1142/S0218127404009089

[23] Huang C-J, Liao J-J, Yang D-X, Chang T-Y, Luo Y-C. Realization of a news dissemination agent based on weighted association rules and text mining techniques. Expert Syst Appl 2010; 37: 6409-13. http://dx.doi.org/10.1016/j.eswa.2010.02.078

[24] Kostoff RN, del Río JA, Humenik JA, García EO, Ramírez AM. Citation mining: Integrating text mining and bibliometrics for research user profiling. J Am Soc Inf Sci Technol 2001; 52: 1148-56. http://dx.doi.org/10.1002/asi.1181

[25] Kostoff RN, Tshiteya R, Pfeil KM, Humenik JA. Electrochemical power text mining using bibliometrics and database tomography. J Power Sources 2002; 110: 163-76. http://dx.doi.org/10.1016/S0378-7753(02)00233-1

[26] Kongthon A. A text mining framework for discovering technological intelligence to support science and technology management [Internet] [Ph.D.]. Georgia Institute of Technology; 2004 [cited 2014 Jan 21]. Available from: http://202.28.199.34/multim/3126708.pdf

[27] Kostoff RN, Tshiteya R, Pfeil KM, Humenik JA, Karypis G. Power source roadmaps using bibliometrics and database tomography. Energy 2005; 30: 709-30. http://dx.doi.org/10.1016/j.energy.2004.04.058

[28] de Miranda Santo M, Coelho GM, dos Santos DM, Filho LF. Text mining as a valuable tool in foresight exercises: A study on nanotechnology. Technol Forecast Soc Change 2006; 73: 1013-27. http://dx.doi.org/10.1016/j.techfore.2006.05.020

[29] Kostoff RN, Koytcheff RG, Lau CGY. Global nanotechnology research literature overview. Technol Forecast Soc Change 2007; 74: 1733-47. http://dx.doi.org/10.1016/j.techfore.2007.04.004

[30] Malheiros V, Hohn E, Pinho R, Mendonca M. A Visual Text Mining approach for Systematic Reviews. In: First International Symposium on Empirical Software Engineering and Measurement, 2007 ESEM 2007. 2007; pp. 245-54. http://dx.doi.org/10.1109/esem.2007.21

[31] Delen D, Crossland MD. Seeding the survey and analysis of research literature with text mining. Expert Syst Appl 2008; 34: 1707-20. http://dx.doi.org/10.1016/j.eswa.2007.01.035

[32] Kajikawa Y, Yoshikawa J, Takeda Y, Matsushima K. Tracking emerging technologies in energy research: Toward a roadmap for sustainable energy. Technol Forecast Soc Change 2008; 75: 771-82. http://dx.doi.org/10.1016/j.techfore.2007.05.005

[33] Kim H, Lee JY. Exploring the emerging intellectual structure of archival studies using text mining: 2001-2004. J Inf Sci 2008; 34: 356-69. http://dx.doi.org/10.1177/0165551507086260

[34] Kostoff RN. Literature-Related Discovery (LRD): Introduction and background. Technol Forecast Soc Change 2008; 75: 165-85. http://dx.doi.org/10.1016/j.techfore.2007.11.004

[35] Kostoff RN, Briggs MB, Solka JL, Rushenberg RL. Literature-related discovery (LRD): Methodology. Technol Forecast Soc Change 2008; 75: 186-202. http://dx.doi.org/10.1016/j.techfore.2007.11.010

[36] Liu JS, Kuan C-H, Cha S-C, Chuang W-L, Gau GJ, Jeng J-Y. Photovoltaic technology development: A perspective from patent growth analysis. Sol Energy Mater Sol Cells 2011; 95: 3130-6. http://dx.doi.org/10.1016/j.solmat.2011.07.002

[37] Thomas J, McNaught J, Ananiadou S. Applications of text mining within systematic reviews. Res Synth Methods 2011; 2: 1-14. http://dx.doi.org/10.1002/jrsm.27

[38] Kostoff RN. Literature-related discovery and innovation — update. Technol Forecast Soc Change 2012; 79: 789-800. http://dx.doi.org/10.1016/j.techfore.2012.02.002

[39] Tu Y-N, Seng J-L. Indices of novelty for emerging topic detection. Inf Process Manag 2012; 48: 303-25. http://dx.doi.org/10.1016/j.ipm.2011.07.006

[40] Küçük D, Arslan Y. Semi-automatic construction of a domain ontology for wind energy using Wikipedia articles. Renew Energy 2014; 62: 484-9. http://dx.doi.org/10.1016/j.renene.2013.08.002

[41] Yoon J. Detecting weak signals for long-term business opportunities using text mining of Web news. Expert Syst Appl 2012; 39: 12543-50. http://dx.doi.org/10.1016/j.eswa.2012.04.059

[42] Miller TW. Data and text mining: a business applications approach. Upper Saddle River, N.J.: Pearson Prentice Hall 2005.

[43] Romero C, Ventura S. Educational data mining: A survey from 1995 to 2005. Expert Syst Appl 2007; 33(1): 135-46. http://dx.doi.org/10.1016/j.eswa.2006.04.005

[44] Rexer Analytics. Rexer Analytics 6th Data Miner Survey - 2013 [Internet]. 2014 [cited 2014 Oct 22]. Available from: http://www.rexeranalytics.com/Data-Miner-Survey-Results-2013.html

Received on 20-10-2015 Accepted on 03-02-2016 Published on 25-07-2016

DOI: http://dx.doi.org/10.6000/1929-6002.2016.05.02.2