Identifying Emerging Research Related to Solar Cells Field using a Machine Learning Approach

Hajime Sasaki*[*1], Tadayoshi Hara[2], Ichiro Sakata[3]
[1]Policy Alternatives Research Institute, University of Tokyo, 7-3-1 Hongo, Bunkyo-ku, Tokyo, Japan
e-mail: sasaki@pari.u-tokyo.ac.jp
[2]Innovation Policy Research Center, Institute of Engineering Innovation, School of Engineering, University of Tokyo, Yayoi 2-11-16, Bunkyo-ku, Tokyo, Japan
e-mail: t.hara@ipr-ctr.t.u-tokyo.ac.jp
[3]Policy Alternatives Research Institute, University of Tokyo, 7-3-1 Hongo, Bunkyo-ku, Tokyo, Japan
Innovation Policy Research Center, Institute of Engineering Innovation, School of Engineering, University of Tokyo, Yayoi 2-11-16, Bunkyo-ku, Tokyo, Japan
e-mail: isakata@ipr-ctr.t.u-tokyo.ac.jp

Cite as: Sasaki, H., Hara, T., Sakata, I., Identifying Emerging Research Related to Solar Cells Field using a Machine Learning Approach, J. sustain. dev. energy water environ. syst., 4(4), pp 418-429, 2016,
DOI: http://dx.doi.org/10.13044/j.sdewes.2016.04.0032

ABSTRACT

The number of research papers related to solar cells field is increasing rapidly. It is hard to grasp research trends and to identify emerging research issues because of exponential growth of publications, and the field's subdivided knowledge structure. Machine learning techniques can be applied to the enormous amounts of data and subdivided research fields to identify emerging researches. This paper proposed a prediction model using a machine learning approach to identify emerging solar cells related academic research, i.e. papers that might be cited very frequently within three years. The proposed model performed well and stable. The model highlighted some articles published in 2015 that will be emerging in the future. Research related to vegetable-based dye-sensitized solar cells was identified as the one of the promising researches by the model. The proposed prediction model is useful to gain foresight into research trends in science and technology, facilitating decision-making processes.

KEYWORDS

Solar cells, Photovoltaic, Emerging research, Technology prediction, Citation network, Machine learning, Scientometrics, Innovation management.

INTRODUCTION

Analyzing trends in academic research can be very helpful when determining the direction of technical developments. This is particularly true in a field such as solar photovoltaic power, which uses technologies that have close linkages to scientific knowledge.

Many methods have been applied in various fields to produce technological forecasts by gathering experts and making a consensus. Recently, some weaknesses have been pointed out with these methods. One is that the individuals who create the forecast are increasingly dependent on the relevant knowledge-base; committee members could produce a useful forecast on their own. Another issue is the huge amount of related data. Few professionals can completely ascertain a comprehensive image of the field. The number of related research papers rapidly increases, so it is difficult for one person,

* Corresponding author

restricted by time and resource constraints, to perceive the contents of all available papers.

Researchers now need methods that can identify emerging research in advance from the vast amounts of available information. The large amounts of data and finely segmented research fields have necessitated such methods, spurring the development of machine-learning techniques. Among such techniques, some methods have been proposed to identify emerging research efforts that might eventually lead to great advances. Emerging research is one that might develop into remarkable and fruitful research activities, although it may not have been in the spotlight at the time of publication. In this paper, the prediction of emerging research was defined as advance identification of papers that might be cited very frequently at a later date.

Many earlier works have proposed methods for estimating and predicting emerging fields in science and technology. Winnink and Tijssen demonstrated the predictability of emerging fields in graphene research, which eventually developed into a paper that won a Nobel Prize [1]. Adams reported a correlation between the numbers of citations that arose in the literature 3-10 years after publication of a paper and those 1-2 years after its publication [2]. Goffman and Newill modeled the propagation of information similarly to the spread of plague [3]. Bettencourt *et al*. described the propagation of new fields using a Susceptible-Infected-Recovered (SIR) model that had been used to simulate a spreading plague [4]. Chen *et al*. assessed research papers related to structural holes of networks making use of a co-citation network and a joint-research network [5].

Kajikawa *et al*. collected papers related to solar photovoltaic power generation, constructed a landscape of academic knowledge and demonstrated that the field is divided into four clusters [6]. Lizin *et al*. described a landscape of academic knowledge related to patent data and compared it with an organic photovoltaic effect [7]. Sakata and Sasaki analyzed the publication trends in the field of solar photovoltaic power generation in several countries; their results showed that Asian countries keep up with global trends [8]. Shibata *et al*. analyzed bibliographic data from academic papers and patents, and discussed development prediction in fields that had sufficient research papers but few patents [9]. Consequently, many reports described a general landscape and reviews, but there were no attempts to predict the growth of citations in the field of solar cell. Therefore, we believe that our research is important to the field of solar cell.

Methods for predicting emerging research have been proposed by researchers in bibliometrics or library and information science. However, owing to the increasing influence of “big data”, these predictions are currently studied in the fields of computer science, data mining and information retrieval. Li and Tong considered predicting the number of citations as an optimization problem. For 500,000 papers in computer science, that study predicted the number of citations 10 years after publication based on the number of citations during the first 3 years after publication. Their results showed that the number of citations 3 years after publication is a useful predictor of later citations [10]. Dong *et al*. predicted the *h*-index of authors 5 years after the publication of their papers. The impact of a paper is defined using six factors: author, content, publisher, citation, co-authors and chronological order. The dataset used for that study included 2 million papers related to computer science [11]. Davletov *et al*. predicted citations 5 and 10 years after publication based on chronological data of citations a few years after publication, and structural information related to citation networks [12]. They used a dataset of 27,000 arXiv records for papers related to energy physics, 1.5 million AMiner records and 2 million CiteSeerX records related to computer science papers [13-15]. Their results show the importance of chronological citation data during the first 2 years after publication [12]. Chakraborty *et al*. classified chronological information related to the

number of citations a few years after publication into six patterns, and predicted the number of citations over 5 years based on the features of authors, academic societies and keywords [16]. Their dataset included 1.5 million data records of computer science papers, and their results demonstrated the particular importance of the number of citations of a paper's author and the number of citations 1 year after publication. Wang *et al*. examined a method that predicted future citations from chronological citations over the 5 years after publication, using the power law. Their dataset included bibliographic data from three journals: *Physical Review B, PNAS* and *Cell*. The citations of 90% of the papers matched the predictions for the 25 years after publication [17].

These prediction methods are based on chronological citation data for a few years after publication, particularly the number of citations and the degree of impact. However, our objective is the "early" prediction of emerging research. This research has tried to predict the growth of citations in the near future (3 years after publication) using chronological data for the year after publication. Our method differs from existing techniques in that it uses only topological features such as network indices without domain-specific information (e.g. keywords). Furthermore, it uniquely predicts an increase of citations in the near future using chronological data obtained shortly after publication. The authors extracted structural features at different granularities from large citation networks using clustering analysis. Our model represents a novel early prediction method, integrating structural features from citation networks.

METHODS

Construction of the prediction model

In this research, academic papers that had the terms "solar cell" or "photovoltaic" in their title, abstract or keywords were extracted from the Thomson Web of Science Core Collection database. Only journal papers related to the field were targeted. The information related to the target field was extracted including paper title, abstract, name of authors, year of publication and citation-related information from the dataset. From the extracted data, a citation network was created for each year, with cumulative papers as nodes and with cumulative citation relationships as links of the networks. From the created time expanded network, the features of the following classes were extracted in each paper of each year. Here, the constructed features are used to express learning data for predicting emerging research.

The features used in the prediction model were categorized into four classes: network, cluster, centrality and properties of citation. The network features represent the general features of the citation network. A cluster is defined as a set of papers that have many citations in the citation network, extracted by maximizing the modularity [18]. Centrality represents how central the paper is in terms of its position in the cited network. The degree of centrality can be represented using several methods [19-25]. The citation properties are the overall statistical properties: maximum, minimum, average and sum of the set of papers that a paper cites. The 63 features were used as presented in Table 1. These features were calculated for all of the papers in the largest connected component, and were used as explanatory variables. The result predicts if a paper will be emerging.

In this paper, emerging research was defined as "papers for which the incremental of citation 3 years after ($t_0 + 3$) publication are in the top 5% of all papers published in that same year (t_0) in the dataset". Based on this definition, a model was constructed that extracts the features of emerging research. For this purpose, a model used papers that are emerging 3 years after publication ($t_0 + 3$) as the training data and applied it to data 4 years later ($t_0 + 4 = t_1$). Data published in this year (t_1) is referred to as the prediction

target year data. To evaluate the performance of this model, the citation number from 3 years after the prediction target year (t_1 + 3) was used. Figure 1 shows the relationship between the training target period and prediction target period.

Table 1. Features used in prediction model

Class of feature	Name of feature	Description	Ref.
Network		Dataset in question and feature of network in the year in question	
	NW_NODES	Number of papers in a network	
	NW_EDGES	Number of citation links in a network	
	NW_MAXQ	Maximum of *Q*-values of clusters in a network	[18]
Cluster		Feature of the cluster to which a paper belongs	
	CL_QMAX	Maximum of *Q*-values of clusters to which a paper belongs	[18]
	CL_NODES	Number of nodes in the cluster to which a paper belongs	
	CL_RANK	Rank of the cluster to which a paper belongs	
Centrality		Network centrality of a paper	
	CNT_DEGRE	Degree centrality	[19]
	CNT_BETWE	Betweenness centrality	[20]
	CNT_CLOSE	Closeness centrality	[19]
	CNT_EIGEN	Eigenvector centrality	[21]
	CNT_NETWO	Network constraint	[22]
	CNT_CLUST	Clustering coefficient	[23]
	CNT_PAGER	Page rank	[24]
	CNT_HUBSC	Hub score	[25]
	CNT_AUTHOR	Authority score	[25]
Property of citation		Feature made as sum of features of paper sets that a paper cites	
	CITING_MAX-[feature]	Maximum of feature in question in cited paper sets that a paper cites	
	CITING_MIN-[feature]	Minimum of feature in question in cited paper sets that a paper cites	
	CITING_AVG-[feature]	Average of features in question in cited paper sets that a paper cites	
	CITING_SUM-[feature]	Sum of features in question in cited paper sets that a paper cites	

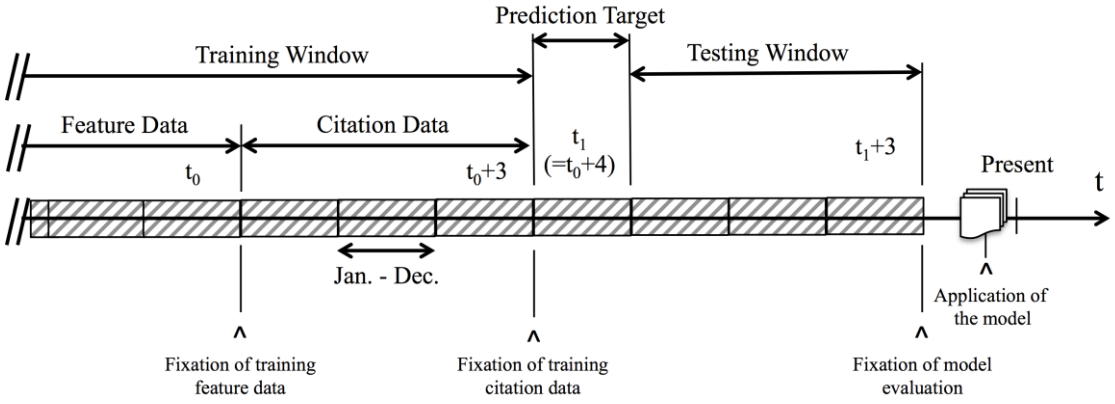

Figure 1. Model training and prediction

For example, if 2012 was the prediction target year (t_1), the model requires features data up to the year 2008 (t_0) and the correct data at t_0 + 3. This was called the "2008 model". We can apply the "2008 model" to the data for 2012 ($t_1 = t_0 + 4$) to calculate our prediction. This prediction model was evaluated by using data from the end of 2015 (t_1 + 3). Table 2 shows which data was used for each training and verification step.

The model was constructed by using a statistical machine learning method. Using knowledge from the data confirmation year, items that become emerging research were classed as "positive examples". Papers with citation numbers in the bottom 50% were classed as "negative examples". A model explains the response variable using the features (explanatory variables) calculated as shown above. This research chose logistic regression as a classifier model. The output of the model is probability of a binary response variable. LIBLINEAR from the analysis package was used [26].

Table 2. Model training year and corresponding verification year

Training window		Testing window	
Model training year t_0	Correct data (Training citation data) confirmation year $t_0 + 3$	Prediction year t_1	Prediction model evaluation year $t_1 + 3$
2003	2006	2007	2010
2004	2007	2008	2011
2005	2008	2009	2012
2006	2009	2010	2013
2007	2010	2011	2014
2008	2011	2012	2015

The authors randomly extracted the negative example with the same amount as positive example sets. This process was repeated to generate multiple data sets for each year, which were then used to construct the models. To predict the model performance, the average performance of multiple models was calculated for each year. Additionally, 5-fold cross validation was implemented for each model to avoid overfitting.

Evaluation of the prediction model

The *F*-value was used to evaluate the analytical model. The *F*-value is an index defined as the harmonic mean of precision and recall. Precision is the ratio of actually emerging papers to those predicted as emerging. Recall is the ratio of papers predicted as emerging to actually emerging papers. The *F*-value is extensively used to evaluate prediction models.

Prediction by model constructed

In this phase, the input data was papers published between January 1, 2015 and December 31, 2015 and the papers were determined by the model predicted to be in the top 5% of papers in 2018. The forecasted top 10 papers were examined in this research.

RESULTS

Dataset retrieval and feature creation

Papers that included the terms “solar cell” or “photovoltaic” in title or keywords were extracted from the Web of Science between January 1, 1900 and December 27, 2015. This resulted in 121,393 papers. The earliest was published in 1906. Figure 2 shows the number of publications after 1900. There was exponential growth after the 1990’s (more than 18,000 reports were published in 2015).

Figure 2. Number of publications for each year following 1900

Examining the network produced by direct citations of these papers, 112,430 papers were found that belonged to the largest connected network set. The number of annual publications was calculated as shown in Table 1 for papers in this largest connected component. The number of citations for all papers in the network was also calculated.

Model development

The negative examples were randomly extracted with the same amount as positive examples eight times to construct eight datasets for each year (corresponding to the prediction models). The precision of the results for each year was shown in Table 3. All the *F*-values exceeded 70, demonstrating a stable precision.

Table 3. Precision for each year

Year of prediction done	Year of model learning	Average of experimental values on balance sets of negative examples each year			
		Number of papers predicted	Number of emerging papers	Number of papers predicted as emerging paper	*F*-value
2007	2003	688	344	334.3	82.2
2008	2004	824	412	401.2	83.3
2009	2005	1,030	515	493.6	82.9
2010	2006	1,358	679	655.9	83.9
2011	2007	1,818	909	900.5	76.8
2012	2008	2,246	1,123	1,118.7	74.9

Table 4 shows the most important features for each models. They were ordered by descending importance.

Table 4. Five most important features for each model

Constructed by 2003 data for 2007 prediction		Constructed by 2004 data for 2008 prediction		Constructed by 2005 data for 2009 prediction	
CNT_DEGRE	4.8	CNT_PAGER	6.2	CNT_PAGER	6.9
CNT_PAGER	4.7	CNT_DEGRE	4.3	CNT_DEGRE	3.4
CNT_AUTHO	2.4	CITING_MAX-CNT_DEGRE	3.2	CITING_MAX-CNT_DEGRE	3.0
CITING_MAX-CNT_HUBSC	2.4	CNT_AUTHO	2.5	CNT_AUTHO	2.6
CITING_MIN-CNT_CLUST	2.2	CITING_MAX-CNT_CLOSE	2.1	CITING_SUM-CNT_DEGRE	2.2
Constructed by 2006 data for 2010 prediction		**Constructed by 2007 data for 2011 prediction**		**Constructed by 2008 data for 2012 prediction**	
CNT_PAGER	5.8	CNT_PAGER	6.5	CNT_PAGER	10.2
CNT_DEGRE	5.5	CNT_AUTHO	5.0	CNT_DEGRE	7.6
CNT_AUTHO	4.6	CNT_DEGRE	4.2	CNT_AUTHO	6.9
CITING_MAX-CNT_DEGRE	2.1	CITING_MAX-CNT_DEGRE	2.2	CITING_AVG-CNT_CLOSE	3.5
CITING_SUM-CNT_DEGRE	2.0	CITING_SUM-CNT_DEGRE	2.1	CITING_MIN-CNT_CLOSE	3.4

The feature with the highest weight in Table 4 was PageRank (CNT_PAGER) [21]. It is calculated using an algorithm that assesses the importance of a webpage and evaluates academic papers based on citation properties. This index identifies a paper cited by papers that are themselves frequently cited. Furthermore, it reduces the relative importance of papers that have citations that contain mutual citations. The next most important feature was degree centrality (CNT_DEGRE) [16]. The more a paper is cited in reference lists, the higher the index. The authority score (CNT_AUTHOR) is high for papers that represent bridges between clusters [22]. This sort of paper could generate a new, emerging research. The importance of the CITING_SUM-CL_RANK feature indicates that an increase in the number of clusters that include a paper increases its chance of becoming emerging. The sixth to ninth ranking features are based on features of papers in reference lists of papers.

Table 5 shows how the top 10 papers in 2012 that were predicted to become emerging have expanded their citations in 2014, 3 years later. Papers 1, 3, 4, 6, 7, 8 and 10 in Table 5 were considered emerging in 2014. That is, 70% of the 10 papers listed in Table 5 were in the top 5% for 2014.

Table 5. Top 10 predictions for 2012 (that is, predicted to become emerging)

Authors	Title	Journal	No. cited (in Dec. 2012)	No. cited (in Dec. 2015)	Ref.
Yip, H. L., Jen, A. K. Y.	Recent advances in solution-processed interfacial materials for efficient and stable polymer solar cells	Energy & Environmental Science	19	320	[27]
Chen, *et al*.	Morphology characterization in organic and hybrid solar cells	Energy & Environmental Science	5	180	[28]
Kumar, P., Chand, S.	Recent progress and future aspects of organic solar cells	Progress in Photovoltaics	1	68	[29]
Boucle, J., Ackermann, J.	Solid-state dye-sensitized and bulk heterojunction solar cells using TiO_2 and ZnO nanostructures: recent progress and new concepts at the borderline	Polymer International	13	50	[30]
Zhou, *et al*.	Rational design of high performance conjugated polymers for organic solar cells	Macromolecules	52	584	[31]
Ooyama, Y., Harima, Y.	Photophysical and electrochemical properties, and molecular structures of organic dyes for dye-sensitized solar cells	Chemphyschem	0	104	[32]
Mishra, A., Bauerle, P.	Small molecule organic semiconductors on the move: promises for future solar energy technology	Angewandte Chemie-International Edition	42	529	[33]
Li, *et al*.	Characterisation of electron transport and charge recombination using temporally resolved and frequency-domain techniques for dye-sensitised solar cells	International Reviews in Physical Chemistry	1	35	[34]
Dou, *et al*.	Tandem polymer solar cells featuring a spectrally matched low-bandgap polymer	Nature Photonics	0	865	[35]
Berger, *et al*.	The electrochemistry of nanostructured titanium dioxide electrodes	Chemphyschem	3	36	[36]

Prediction for papers published in 2015

Lastly, the papers published in 2015 were inputted into prediction model and the top 10 papers were listed as shown in Table 6.

Table 6. Top 10 predictions for 2015 (i.e. predicted to become emerging)

Authors	Title	Journal	Ref.
Calogero, *et al*.	Vegetable-based dye-sensitized solar cells	Chemical Society Reviews	[37]
Wu, *et al*.	Electrolytes in dye-sensitized solar cells	Chemical Reviews	[38]
Lu, *et al*.	Recent advances in bulk heterojunction polymer solar cells	Chemical Reviews	[39]
Bella, *et al*.	Aqueous dye-sensitized solar cells	Chemical Society Reviews	[40]
Cheng, *et al*.	Versatile third components for efficient and stable organic solar cells	Materials Horizons	[41]
Chueh, *et al*.	Recent progress and perspective in solution-processed Interfacial materials for efficient and stable polymer and organometal perovskite solar cells	Energy & Environmental Science	[42]
Liu, *et al*.	Functionalized graphene and other two-dimensional materials for photovoltaic devices: device design and processing	Chemical Society Reviews	[43]
Singh, *et al*.	Graphene-based dye-sensitized solar cells: a review	Science of Advanced Materials	[44]
Liang, *et al*.	ZnO cathode buffer layers for inverted polymer solar cells	Energy & Environmental Science	[45]
Albero, *et al*.	Efficiency records in mesoscopic dye-sensitized solar cells	Chemical Record	[46]

DISCUSSION

This paper proposed and evaluated a method that predicts whether a published paper will become an emerging one in the next 3 years. Table 5 shows that 70% of the top 10

predictions for 2012 were correct. The proposed model was sufficiently dependable; the *F*-values fluctuated around 70 for all of the years, and the precision and recall values suggest that the model was accurate.

PageRank was an important predictor; a paper that is cited by frequently cited papers is therefore more likely to become emerging. Furthermore, a higher degree of centrality indicates that a paper citing many papers in its reference list will be cited in years to come. As a result of these mechanisms, many review papers could have been predicted as emerging. However, all the papers that are cited frequently in their reference lists do not necessarily develop into emerging research. Determining these papers that are very likely to become emerging could facilitate estimates of future research and development trends.

Table 6 contains the predictions of the most important publications in 2018. One paper by Calogero *et al*. describes vegetable-based dye-sensitized solar cells. Vegetable dyes are sensitizers extracted from alga, flowers and fruit [37]. Vegetable-based dye-sensitized solar cells use these Dye-Sensitized Solar Cells (DSSC). Kay and Grätzel proposed this idea [47]. Because that paper was published, many researchers have tackled this idea. In fact, as of September 3, 2015, that report had been cited 733 times. This field should be carefully observed.

Perovskite solar cells have gained widespread attention. They are produced from cheap materials using a solution technique and so they are highly likely to be used extensively. Perovskite is a crystal structure of calcium titanate (perovskite, $CaTiO_3$). It was named after the Russian researcher Perovski, who first reported the structure. The first paper related to perovskite photoelectric conversion was published in 2009 [48]. The National Renewable Energy Laboratory (NREL) reported a value of 20.1% as the most efficient perovskite photoelectric conversion on February 17, 2015 [49]. Chueh *et al*. reviewed the latest developments in solution-processed interfacial layers, which have contributed to a marked improvement in the performance of polymer and perovskite solar cells [42]. Based on the results, we can assume that perovskite photovoltaics will become an emerging research field. Two important journals (Science and Nature) highlighted perovskite photovoltaics as one of the greatest breakthroughs of 2013 [50, 51].

At the end of 2018, we will be able to evaluate the predictions in Table 6. Because some of these papers deal with common themes, our method could help decision-making processes. This method becomes useful when private enterprises plan their research and development activities or when central governments make decisions related to science and technology policies.

CONCLUSIONS

This paper proposed a prediction model that uses large amounts of data to determine potential papers that will later become emerging in solar cells field. The authors succeeded to predict the growth of citations three years after publication by applying machine learning techniques to information derived from data one year after publication. The goal was to achieve an “early” prediction of emerging fields. Various features were used in the model. These features were not used in existing research. The authors used four classes of features: network, cluster, centrality and properties of citation.

Dataset contained papers that included “solar cell” or “photovoltaic” and were extracted from the Web of Science. They were published between January 1, 1900 and December 31, 2015. There were 121,393 papers in the dataset.

This paper could test the model results for 2007-2012 and found that the *F*-values were stable and greater than 70. This paper also forecast the results for 2018 using papers from 2015 and believes we found useful information regarding future solar power technologies.

The model predicted that a paper related to vegetable-based dye-sensitized solar cells would be emerging. Although the ideas in the paper are not new, this type of solar cell is remarkable because of the more efficient conversion. Our model predicted the incremental of citation of this paper would be remarkable in 3 years.

The conversion efficiency of perovskite solar cell power generation has increased by a factor of five and is approaching that of silicon-based solar cells (which are currently very extensively used). The cheap production cost of perovskite solar cells means that they are becoming popular. Future developments in perovskite solar cells will be very important. Many of the papers that the model forecasted to be most emerging consider common topics.

Among the papers published in 2015, an authoritative journal publisher has noted all of the forecasted top 10 papers. This demonstrates that the prediction is rational. We should test these predictions and propose guidelines to ascertain future trends, while constructing more stable models.

The rapidly increasing amount of information and complicated knowledge structures mean that it is difficult for private enterprises to manage research and development decisions, and for governments to develop science and technology policies. This model can be used to gain foresight into developing trends in science and technology, facilitating human decision making processes. The proposed model must be more important for the researchers in fields of sustainability to processes huge amounts of information in the field, analyzes it, and extracts papers that are expected to be valuable in the near future.

ACKNOWLEDGEMENT

This research was supported by grants from the Project of the NARO Bio-oriented Technology Research Advancement Institution (Integration research for agriculture and interdisciplinary fields).

REFERENCES

1. Winnink, J. J. and Tijssen, R. J., Early Stage Identification Of Breakthroughs at the Interface of Science and Technology: Lessons Drawn from a Landmark Publication, *Scientometrics*, Vol. 102, No. 1, pp 113-134, 2015, http://dx.doi.org/10.1007/s11192-014-1451-z
2. Adams, J., Early Citation Counts Correlate with Accumulated Impact, *Scientometrics*, Vol. 63, No. 3, pp 567-581, 2005, http://dx.doi.org/10.1007/s11192-005-0228-9
3. Goffman, W. and Newill, V. A., Generalization of Epidemic Theory, *Nature*, Vol. 204, pp 225-228, 1964, http://dx.doi.org/10.1038/204225a0
4. Bettencourt, L., Kaiser, D., Kaur, J., Castillo-Chavez, C. and Wojick, D., Population Modeling of the Emergence and Development of Scientific Fields, *Scientometrics*, Vol. 75, No. 3, pp 495-518, 2008, http://dx.doi.org/10.1007/s11192-007-1888-4
5. Chen, C., Chen, Y., Horowitz, M., Hou, H., Liu, Z. and Pellegrino, D., Towards an Explanatory and Computational Theory of Scientific Discovery, *Journal of Informetrics*, Vol. 3, No. 3, pp 191-209, 2009, http://dx.doi.org/10.1016/j.joi.2009.03.004
6. Kajikawa, Y., Ohno, J., Takeda, Y., Matsushima, K. and Komiyama, H., Creating an Academic Landscape of Sustainability Science: An Analysis of the Citation Network, *Sustainability Science*, Vol. 2, No. 2, pp 221-231, 2007, http://dx.doi.org/10.1007/s11625-007-0027-8

7. Lizin, S., Leroy, J., Delvenne, C., Dijk, M., De Schepper, E. and Van Passel, S., A Patent Landscape Analysis for Organic Photovoltaic Solar Cells: Identifying the Technology's Development Phase, *Renewable Energy*, Vol. 57, pp 5-11, 2013, http://dx.doi.org/10.1016/j.renene.2013.01.027
8. Sakata, I. and Sasaki, H., Scientific Catch-up in Asian Economies: A Case Study for Solar Cell, *Natural Resources*, Vol. 4, No. 1A, pp 134-141, 2013, http://dx.doi.org/10.4236/nr.2013.41A017
9. Shibata, N., Kajikawa, Y. and Sakata, I., Extracting the Commercialization Gap between Science and Technology – Case Study of a Solar Cell, *Technological Forecasting And Social Change*, Vol. 77, No. 7, pp 1147-1155, 2010, http://dx.doi.org/10.1016/j.techfore.2010.03.008
10. Li, L. and Tong, H., The Child is Father of the Man: Foresee the Success at the Early Stage, *Arxiv Preprint Arxiv:1504.00948*, 2015.
11. Dong, Y., Johnson, R. A. and Chawla, N. V., Will this Paper Increase your H-index?: Scientific Impact Prediction, *Proceedings of the 8th ACM International Conference on Web Search and Data Mining*, pp 149-158, January 31-February 6, 2015, http://dx.doi.org/10.1007/978-3-319-23461-8_26
12. Davletov, F., Aydin, A. S. and Cakmak, A., High Impact Academic Paper Prediction using Temporal and Topological Features, *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management*, pp 491-498, November 3-7, 2014, http://dx.doi.org/10.1145/2661829.2662066
13. Arxiv, http://Arxiv.Org, [Accessed: 17-March-2016]
14. Aminer – Open Science Platform, http://www.Arnetminer.org, [Accessed: 17-March-2016]
15. Citeseerx, http://Citeseerx.Ist.Psu.Edu, [Accessed: 17-March-2016].
16. Chakraborty, T., Kumar, S., Goyal, P., Ganguly, N. and Mukherjee, A., Towards a Stratified Learning Approach to Predict Future Citation Counts, *Proceedings of the 14th ACM/IEEE-CS Joint Conference on Digital Libraries*, pp 351-360, September 8-12, 2014, http://dx.doi.org/10.1109/jcdl.2014.6970190
17. Wang, D., Song, C. and Barabási, A. L., Quantifying Long-term Scientific Impact, *Science*, Vol. 342, pp 127-132, 2013, http://dx.doi.org/10.1126/science.1237825
18. Newman, M. E., Modularity and Community Structure in Networks, *Proceedings of the National Academy of Sciences*, Vol. 103, No. 23, pp 8577-8582, 2006, http://dx.doi.org/10.1073/pnas.0601602103
19. Freeman, L. C., Centrality in Social Networks Conceptual Clarification, *Social Networks*, Vol. 1, No. 3, pp 215-239, 1979, http://dx.doi.org/10.1016/0378-8733(78)90021-7
20. Freeman, L. C., A Set of Measures of Centrality Based on Betweenness, *Sociometry*, Vol. 40, No. 1, pp 35-41, 1977, http://dx.doi.org/10.2307/3033543
21. Bonacich, P., Technique for Analyzing Overlapping Memberships, *Sociological Methodology*, Vol. 4, pp 176-185, 1972, http://dx.doi.org/10.2307/270732
22. Burt, R. S., Structural Holes and Good Ideas, *American Journal of Sociology*, Vol. 110, No. 2, pp 349-399, 2004, http://dx.doi.org/10.1086/421787
23. Watts, D. J. and Strogatz, S. H., Collective Dynamics of 'Small-World' Networks, *Nature*, Vol. 393, pp 440-442, 1998, http://dx.doi.org/10.1038/30918
24. Brin, S. and Page, L., Reprint of: The Anatomy of a Large-Scale Hypertextual Web Search Engine, *Computer Networks*, Vol. 56, No. 18, pp 3825-3833, 2012, http://dx.doi.org/10.1016/j.comnet.2012.10.007
25. Guimera, R. and Amaral, L. A. N., Functional Cartography of Complex Metabolic Networks, *Nature*, Vol. 433, pp 895-900, 2005, http://dx.doi.org/10.1038/nature03288

26. LIBLINEAR – A Library for Large Linear Classification, https://www.Csie.Ntu.Edu.Tw/~Cjlin/Liblinear/https://www.Csie.Ntu.Edu.Tw/~Cjlin/Liblinear, [Accessed: 17-March-2016]
27. Yip, H. L. and Jen, A. K. Y., Recent Advances in Solution-processed Interfacial Materials for Efficient and Stable Polymer Solar Cells, *Energy and Environmental Science*, Vol. 5, No. 3, pp 5994-6011, 2012, http://dx.doi.org/10.1039/c2ee02806a
28. Chen, W., Nikiforov, M. P. and Darling, S. B., Morphology Characterization in Organic and Hybrid Solar Cells, *Energy and Environmental Science*, Vol. 5, No. 8, pp 8045-8074, 2012, http://dx.doi.org/10.1039/c2ee22056c
29. Kumar, P., and Chand, S., Recent Progress and Future Aspects of Organic Solar Cells, *Progress in Photovoltaics: Research and Applications*, Vol. 20, No. 4, pp 377-415, 2012, http://dx.doi.org/10.1002/pip.1141
30. Bouclé, J. and Ackermann, J., Solid-state Dye-sensitized and Bulk Heterojunction Solar Cells using Tio_2 and Zno Nanostructures: Recent Progress and new Concepts at the Borderline, *Polymer International*, Vol. 61, No. 3, pp 355-373, 2012, http://dx.doi.org/10.1002/pi.3157
31. Zhou, H., Yang, L. and You, W., Rational Design of High Performance Conjugated Polymers for Organic Solar Cells, *Macromolecules*, Vol. 45, No. 2, pp 607-632, 2012, http://dx.doi.org/10.1021/ma201648t
32. Ooyama, Y. and Harima, Y., Photophysical and Electrochemical Properties, and Molecular Structures of Organic Dyes for Dye-sensitized Solar Cells, *Chemphyschem*, Vol. 13, No. 18, pp 4032-4080, 2012, http://dx.doi.org/10.1002/cphc.201200218
33. Mishra, A. and Bäuerle, P., Small Molecule Organic Semiconductors on the move: Promises for Future Solar Energy Technology, *Angewandte Chemie International Edition*, Vol. 51, No. 9, pp 2020-2067, 2012, http://dx.doi.org/10.1002/anie.201102326
34. Li, L. L., Chang, Y. C., Wu, H. P. and Diau, E. W. G., Characterisation of Electron Transport and Charge Recombination using Temporally Resolved and Frequency-domain Techniques for Dye-sensitised Solar Cells, *International Reviews in Physical Chemistry*, Vol. 31, No. 3, pp 420-467, 2012, http://dx.doi.org/10.1080/0144235X.2012.733539
35. Dou, L., You, J., Yang, J., Chen, C. C., He, Y., Murase, S., ... and Yang, Y., Tandem Polymer Solar Cells Featuring a Spectrally Matched Low-bandgap Polymer, *Nature Photonics*, Vol. 6, No. 3, pp 180-185, 2012, http://dx.doi.org/10.1038/nphoton.2011.356
36. Berger, T., Monllor-Satoca, D., Jankulovska, M., Lana-Villarreal, T. and Gómez, R., The Electrochemistry of Nanostructured Titanium Dioxide Electrodes, *Chemphyschem*, Vol. 13, No. 12, pp 2824-2875, 2012, http://dx.doi.org/10.1002/cphc.201200073
37. Calogero, G., Bartolotta, A., Di Marco, G., Di Carlo, A. and Bonaccorso, F., Vegetable-based Dye-sensitized Solar Cells, *Chemical Society Reviews*, Vol. 44, No. 10, pp 3244-3294, 2015, http://dx.doi.org/10.1039/C4CS00309H
38. Wu, J., Lan, Z., Lin, J., Huang, M., Huang, Y., Fan, L. and Luo, G., Electrolytes in Dye-sensitized Solar Cells, *Chemical Reviews*, Vol. 115, No. 5, pp 2136-2173, 2015, http://dx.doi.org/10.1021/cr400675m
39. Lu, L., Zheng, T., Wu, Q., Schneider, A. M., Zhao, D. and Yu, L., Recent Advances in Bulk Heterojunction Polymer Solar Cells, *Chemical Reviews*, Vol. 115, No. 23, pp 12666-12731, 2015, http://dx.doi.org/10.1021/acs.chemrev.5b00098
40. Bella, F., Gerbaldi, C., Barolo, C. and Grätzel, M., Aqueous Dye-sensitized Solar Cells, *Chemical Society Reviews*, Vol. 44, No. 11, pp 3431-3473, 2015, http://dx.doi.org/10.1039/C4CS00456F

41. Cheng, P. and Zhan, X., Versatile Third Components for Efficient and Stable Organic Solar Cells, *Materials Horizons*, Vol. 2, No. 5, pp 462-485, 2015, http://dx.doi.org/10.1039/C5MH00090D
42. Chueh, C. C., Li, C. Z. and Jen, A. K. Y., Recent Progress and Perspective in Solution-processed Interfacial Materials for Efficient and Stable Polymer and Organometal Perovskite Solar Cells, *Energy and Environmental Science*, Vol. 8, No. 4, pp 1160-1189, 2015, http://dx.doi.org/10.1039/C4EE03824J
43. Liu, Z., Lau, S. P. and Yan, F., Functionalized Graphene and Other Two-dimensional Materials for Photovoltaic Devices: Device Design and Processing, *Chemical Society Reviews*, Vol. 44, No. 15, pp 5638-5679, 2015, http://dx.doi.org/10.1039/C4CS00455H
44. Singh, E. and Nalwa, H. S., Graphene-based Dye-sensitized Solar Cells: A Review, *Science of Advanced Materials*, Vol. 7, No. 10, pp 1863-1912, 2015, http://dx.doi.org/10.1166/sam.2015.2438
45. Liang, Z., Zhang, Q., Jiang, L. and Cao, G., Zno Cathode Buffer Layers for Inverted Polymer Solar Cells, *Energy and Environmental Science*, Vol. 8, No. 12, pp 3442-3476, 2015, http://dx.doi.org/10.1039/C5EE02510A
46. Albero, J., Atienzar, P., Corma, A. and Garcia, H., Efficiency Records in Mesoscopic Dye-sensitized Solar Cells, *The Chemical Record*, Vol. 15, No. 4, pp 803-828, 2015, http://dx.doi.org/10.1002/tcr.201500007
47. Kay, A. and Graetzel, M., Artificial Photosynthesis, 1st Photosensitization of Titania Solar Cells with Chlorophyll Derivatives and Related Natural Porphyrins, *The Journal of Physical Chemistry*, Vol. 97, No. 23, pp 6272-6277, 1993, http://dx.doi.org/10.1021/j100125a029
48. Kojima, A., Teshima, K., Shirai, Y. and Miyasaka, T., Organometal Halide Perovskites as Visible-light Sensitizers for Photovoltaic Cells, *Journal of the American Chemical Society*, Vol. 131, No. 17, pp 6050-6051, 2009, http://dx.doi.org/10.1021/ja809598r
49. National Center for Photovoltaics, NREL, http://www.Nrel.Gov/Ncpv, [Accessed: 17-March-2016]
50. Coontz, R. (N. D.), Science's Top 10 Breakthroughs of 2013, http://News.Sciencemag.Org/2013/12/Sciences-Top-10-Breakthroughs-2013, [Accessed: 17-March-2016]
51. 365 Days: Nature's 10, Ten People who Mattered this Year, http://News.Sciencemag.Org/2013/12/Sciences-Top-10-Breakthroughs-2013, [Accessed: 17-March-2016]

Paper submitted: 17.03.2016
Paper revised: 11.05.2016
Paper accepted: 14.05.2016