ISSN 2278-3091

Volume 8, No.1.4, 2019

International Journal of Advanced Trends in Computer Science and Engineering

Available Online at http://www.warse.org/IJATCSE/static/pdf/file/ijatcse1881.42019.pdf

https://doi.org/10.30534/ijatcse/2019/1881.42019

Optimization of Idea Mining Model based on Text Position Weight

Azreen Azman[1*], Mostafa Alksher[2], Eissa Alshari[3], Razali Yaakob[1], Shyamala Doraisamy[1]

[1]Universiti Putra Malaysia, Malaysia, {azreenazman,razaliy,shyamala}@upm.edu.my
[2]Elmergib University, Libya, alksher@yahoo.com
[3]IBB University, Yemen, alsherai2002@gmail.com

ABSTRACT

Predictable behaviour of any system is important in order to ensure that its performance is always optimal. Researchers have used many techniques to model system's performance based on the data collected in many different runs to predict the ideal setting for any system. In idea mining, the probabilistic weights of the position of idea in a text needs to be set for optimal setting. In this paper, an experiment with a total of 10,000 runs with different randomly assigned weights is conducted for the idea mining model in order to discover the optimal setting. Based on the mean average score (MAP) score produced in each run, a prediction of the optimal weight is discovered by using the curve fitting based on the least squares method and the artificial neural network (ANN) model. Based on the findings, the ANN model appears to be more suitably fit as compared to the least squares method, which suggests that the data is nonlinear in nature.

Key words : Idea mining, Text position, Optimization, Curve fitting, Artificial Neural Network

1. INTRODUCTION

Innovation has become the key to the success of many organizations or nations in order to be competitive in the real world. It is driven by the capability of its member or citizen to generate an interesting idea and making it work. Brainstorming has been used as an effective idea generation technique for decades [18]. However, it is both expensive and challenging creative process to discover interesting ideas in order to solve a problem or to assist in decision making. In the process, textual resources such as scientific publications and the Web have been utilized as the source for the idea [28], [29].

Idea mining (IM) has been explored as a new task in information retrieval (IR) to automatically extract interesting idea from text [2]. It is an automatic process to identify new and innovative idea from unstructured text by using text mining techniques [17]. Through idea mining, a solution to the present problem can be discovered by analysing huge number of texts in order to help in decision making. Conventionally, a manual process of idea discovery will take longer time and huge effort with the possibility of only a few of them relevant to the current decision problem.

Abstract in a scientific paper has a consistent information structure. Commonly, the abstract starts with a discussion on the back- ground of the study, follows by a brief explanation of the method used and the description of obtained results at the end [15]. Previously, abstract has been considered as an ideal representation of a research paper in which the gist or the main idea of the paper is summarized in the abstract [19]. It was suggested that the most important part of the abstract that contains idea is at the beginning of the text [31]. A similar suggestion is explained in [4] that an author will start writing his/her idea at the introduction (the beginning) before emphasizing it by restating the idea at later stage of the text.

Therefore, many researchers in the area of idea mining have successfully used the abstract of scientific paper or patent as the text resource for identifying idea [1], [20], [29]. As an approximation of the entire content of the paper or pattern, the abstract provides a more manageable text for computation of idea mining model. A further investigation to the feasibility of using text position as an additional feature for idea identification is conducted through a preliminary study [3]. This investigation attempted to discover the position of the piece of text that have a higher likelihood to contain the idea. In this study, the abstract is equally partitioned into 3 sections, the introduction (*int*), the body (*bod*) and the conclusion (*con*). Based on the study, it is found that most idea are located in either the introduction or the conclusion.

As such, the measurement of the idea of the text should be weighted based on its location. For instance, a probabilistic weight can be assigned to int, bod and con, referring to the sections in the abstract, and if the text is located at a specific section, the final idea mining score can be multiplied with the weight. Therefore, the weights for int, bod and con can be assigned to reflect the importance of each section in the final idea mining score.

It is problematic to determine the ideal weight for int, bod and con. The combination of weights can be set randomly and used in the experiment to record the Mean Average Precision (MAP) score for each combination. Such method will find the combination of weights that produces the best MAP score. As the weights are empirically set, it is importance to discover the model to predict the ideal weights in order to consistently produce the best performance for the idea mining model. As such, this paper attempts to investigate the prediction model based on curve fitting method and an artificial neural network (ANN) model.

The remaining of the paper is structured as follows: A review of related work is presented in the following section. Section 3 elaborates the methods for the optimization of the idea mining measurements. Then, the experimental setup, results of utilizing fitting mathematical models and ANN model are discussed in detail in Section 4. Finally, the conclusion of the work is given in Section 5.

2. RELATED WORK

2.1 Idea Mining Models

Idea mining has been growing as an emerging area in information retrieval as more organisation or company focus on innovation. As more techniques being proposed for this task, they are still limited to the context of technological idea within research papers or patents. In particular, most researchers focused only abstracts of those papers or patents as the textual resources for discovering new idea [10], [13], [30]. Aggregation or concatenation of abstracts has been used in different researches on idea mining [1], [29].

The problem of idea mining has been investigated in many different settings. Online text has been used as the textual resources in automatically detect idea based on text extraction [10]. Similarly, an approach based on web mining has been investigated in searching for new idea from the Web by automatically construct search keywords based on idea mining model [28]. The model is derived from the earlier proposed idea mining model based on heuristics [29]. An approach based on crowdsourcing and text mining for discovering new idea has been proposed in [12] which requires involvement of many users. In this paper, the model in [29] is adopted, which is given by the following equation;

$$m_{idea} = \begin{cases} m_b + m_{fq} + m_{fu} + m_c & (p \neq q) \\ 0 & (p = q) \end{cases} \qquad (1)$$

The authors in [27] proposed a rule in the form of sub-measures of idea mining, which refer to the balancing between terms occur most in the new text with the terms that similar and occur in both documents m_b. Furthermore, the ranking of the filtered text is retrieved based on the most frequent known and unknown terms occurred m_{fq} and m_{fu}. However, the mc sub-measure has not being considered since it sets to a standard value ($m_c = 0$).

2.2 Idea Mining Model Based on Text Position

A modified idea mining model to capture the term position within text phrases is utilized [3], in which a term position measure is incorporated into the idea mining model, proposed in [29], to enhance the performance. This modified model assigned as a probability weight to be incorporated into the less considered sub-measure which enables calculating the importance of the position for the candidate ideas.

Authors in [3] performed an experimental investigation by implementing 10000 combinations and produce a new set of random weights based on rules assigned for discovering the best ranking. The findings of these experiments have shown that the highest MAP obtained when integrating position measure highly indicated ideas in conclusion sections more than other sections.

2.3 Curve Fitting

Curve fitting has been widely used for many research problems over the years [16]. With the aim to calculate the values of parameter that minimize the total errors for data points set. The mathematical function representing the data will be the one that best fits the data points series. As such, the curve produced by the method will model the approximation of system's characteristics [5]. The curve fitting methods have been used in many applications such as prediction, growth analysis, pattern recognition, image processing, finance *et cetera*.

Currently, there are several curve fitting methods capable of modelling complex and non-linear functions. In the case of noisy data, a rational fraction polynomial method is suitable for global curve fitting problem due to the fact that it is based on the least squared error principle [25]. However, it may not be the case if it involves large number of varying parameters, which can lead to inaccurate results. As it is the case of optimisation, a genetic algorithm approach has been used to solve the problem [14]. Also, a gradually imposed arbitrage restriction method can be used on the fitting of a sequence of yield curves to control the errors as a result of the fitting [7].

2.4 Artificial Neural Networks

In most cases, curve fitting method is not suitable for non-linear data points. In [5], the authors conclude that the Artificial Neural Network (ANN) can be used to provide fitting calibration of the curves for a multivariate system. For the problem of making direct prediction of the width and the location based on spectral data, ANN has shown to be effective and it is used to calculate the position based on a number of measurements data [8]. ANN has demonstrated to be faster than any conventional iterative techniques [11].

By using ANN to make prediction of wind speed hourly, [21] developed a method to predict deformation of upsetting processes. The method combines finite element and ANN, and also views the changes in the deformation in upsetting.

Furthermore, in order to provide sufficient smooth non-linear equations, Levenberg- Marquardt type algorithms are utilised in the problems of unconstrained multi-objective optimisation [26].

3. METHODOLOGY

In order to find the optimal value for int, bod and con, 10,000 different combinations of parameter values are randomly generated. Each combination is used in the experiment and the respected MAP score is recorded. A total of 50 selected abstracts from re- search papers are used as the *New Text* and the abstracts of the papers referenced by the *New Text* are used as the *Collection Text*. The test collection is based on the prior study in idea mining [1]. The performance of the model is measured by computing the Mean Average Precision (MAP) score of the ranked idea produced by the model.

The optimal combination that produces the best MAP score can be determined from the experiment, which is 0.19 for *int*, 0.13 for *bod* and **0.67** for ***con***. The process is as depicted in Figure 1, and the idea mining model based on text position is derived from the model in [1], [29].

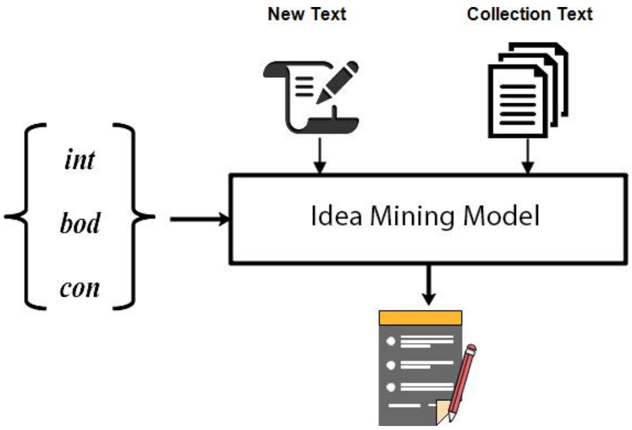

Figure 1: Idea Mining Based on Random Position Weights

Based on the data collected, this paper aims to compute a model to predict the optimal values for int, bod and con in order to get the best MAP score. Two methods are used for the analysis, which are the least squared method for curve fitting and the ANN model.

3.1 Curve Fitting: The Least Squares Method

In this paper, the least squares method is used to generate a curve that best represents the data points to be used for predicting the optimal value of int, bod or con in order to produce the best MAP score, independently. It works by determining the coefficient value that will minimize the value of Chi-square. The Chi-square is defined as:

$$x^2 = \sum_i \left(\frac{y-y_i}{\sigma_i}\right)^2 \qquad (2)$$

where y is a fitted value for a given point, y_i is the measured data value for the point and σ_i is an estimate of the standard deviation for y_i. In other words, the best values of the coefficients are obtained when the sum of the squared distances $\Sigma(y - y_i)^2$ of the fit to the data is minimized while giving data points more weight during this minimization process that have smaller errors (σ). The sum-of-squares $\Sigma(y - y_i)^2$ is the sum of the squares as depicted in Figure 2. The Least-Squares method during curve-fitting aims to minimize this sum.

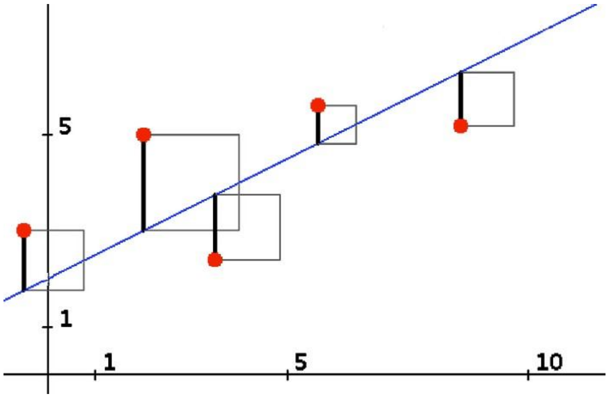

Figure 2: Example of Least-Squares method

The least squares method is capable of producing the line curve that best fitted to data, which uses linear algebra and simple calculus [22]. It is in the form of $y = ax + b$ where,

x, *y* are the coordinates of any point on the line
a is the slope of the line
b is the *y*-intercept (where the line crosses the *y*-axis)

3.2 ANN Model

One limitation of the curve fitting based on the least squares method is that it models the parameters (*int*, *bod*, *con*) independently. The model can only predict the best value for int, bod or con separately and not in combination. As such, the optimal values of all parameter may not be reached.

Alternatively, the artificial neural networks model is capable to solve this repetitive non-linear curve fitting problem. The model is able to determine the optimal values of all parameters in com- bination and has shown to be much faster. In a real-time setting, a special purpose hardware can used to implement the model for high speed processing [8].

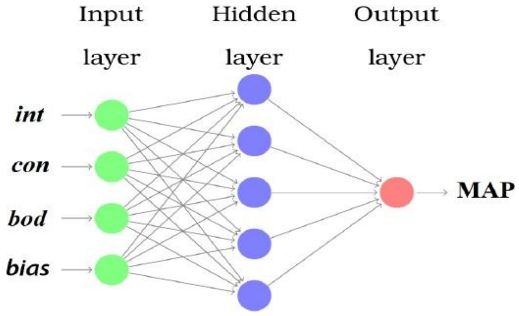

Figure 3: Example of the ANN architecture

The neural network is very flexible and versatile tool. In general, the network consists of the input layer, hidden layer and the output layer as depicted in Figure 3. The hidden layer, placed between input and output layer, is the main architecture to capture the non- linearity of the model. The input layer consists of nodes for int, bod, con and bias, while the output layer consists of the node for the MAP score. In addition, several other parameters need to be set as part of the architecture, which are the training algorithms, maximum training time, performance measure, the number of neurons in a layer, epochs, gradient and validation checks.

4. RESULTS AND DISCUSSION

4.1 The Least Squares Method

In the present research scenario, curve fitting models are used to determine the values of the parameters that make the function match the data as much as possible, so that the best values of the coefficients are those that reduce the value of the Chi-square, which able to model the behaviours of data as an Equation 2. In addition, it is to find a mathematical function to examine the behaviour of the probabilistic weight (*int*, *bod*, *con*) to predict values that is close to the observed mean curve of MAP based on idea mining measurement. In this discussion, only the result for parameter *con* is considered as it has shown to be more influential to the performance of the idea mining model.

As shown in Table 1, the mathematical fitting curve used three equations to estimate the MAP with the maximum variant of the best curve equation R^2 = 0.70 in the *con* parameter curve.

Table 1: Curve Fitting Statistics

Number of points	10,000
Degrees of Freedom	9996
Reduced Chi-Squares	3.38512E-4
Residual Sum of Squares	3.38377
Adjusted R-Square	0.70845
Fit Status	Succeeded (100)

Based on Figure 4, it is difficult to find a mathematical equation that capable of calculating the value of MAP because of the absence of relations between the variables. For example, an attempt to find symmetry between 10,000 samples for one variable (*con*) is almost impossible mathematically because of the scattered points and the absence of mathematically readable relationships; an alternative should be suggested using optimization techniques.

Therefore, a network to predict the best MAP directly from the raw data in a single step process is designed. Thus, neural network fitting is another technique that can improve the R^2 by optimizing the input parameter's weight by providing a prediction matrix. It provides methods for solving the iterative nonlinear curve fitting issues, as it is shown in [6] works that ANN is capable of fitting curves to a multivariate.

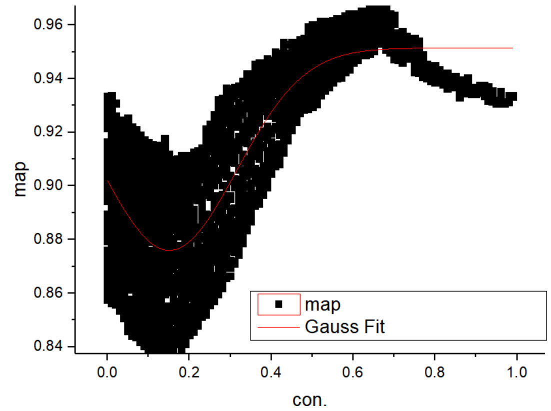

Figure 4: The Plot of Fitted Curve for con

4.2 ANN Model

In this experiment, it is believed that this network is robust enough where the input vectors (tested data) are partitioned randomly into three subsets. The first subset is the training set which is used to update weights and biases according to output values of network and targets. The second subset is the validation set which is used to terminate the training before over-fitting. Lastly, the testing subset is used to predict future performance of the network. In general, the training set makes up approximately 80% of full dataset, with validation and testing (10%, 10%) respectively.

In order to map the random inputs correctly to outputs, the standard technique Levenberg-Marquardt (LM) for solving nonlinear least squares problems and defined as:

$$w_{i+1} = w_i - (J_i^T \cdot J_i + \lambda_i I)^{-1} \cdot (2J_i^T fke_i), i = 0,1, \ldots \qquad (3)$$

In more details, Figure 5 represents the process of adopting the Levenberg-Marquardt algorithm in the training of the ANN model. The first step is to consider the loss, then compute its gradient vector and calculate Hessian approximation to adjust the damping parameter for reducing the loss at each iteration.

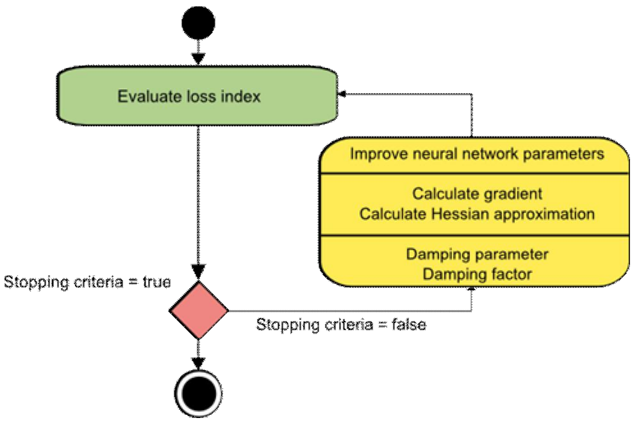

Figure 5: The training process of a neural network with the Levenberg-Marquardt algorithm [24]

The curves are plotted in Figure 6 which displays the network outputs on the targets for training, validation, and test sets. It can be seen from Figure 6 that the fit of the network is reasonably good for all data. Figure 6 depicts the curve fitting plots based on the data points for all training, validation, test and also the overall performance. It shows a strong correlation between them. For this experiment, the *R* value is very close to 1 (approximately 0.996), which means that there is strong correlation between the targets and the outputs.

5. CONCLUSION

The main goal of this study is to optimize predictive performance by implementing the previously introduced approach of iterative 10,000 weight combinations. This paper investigated how to find a mathematical function to examine the behaviour of the probability of positions to predict values that fit the observed mean curve of MAP based on idea mining measurement. This study has used mathematical fitting method for identifying the best fitting of the probability of positions for MAP prediction. These experiments confirmed that the likelihood for the conclusion section to accommodate the ideas is high. The research has also shown that neural network model could be used to optimize the matrix weight for predicting the MAP by given the probability of position (*int*, *bod*, *con*). The scope of the experiments was limited in 10,000 of weights combinations, therefore, there could be experiments with more iterative samples [23].

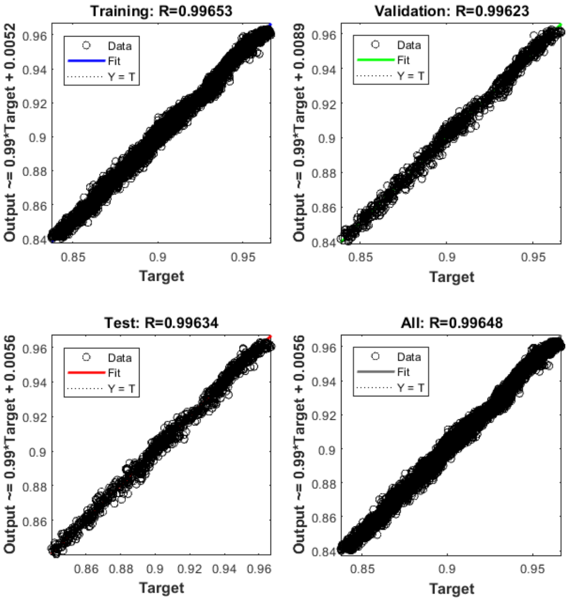

Figure 6: Plots for ANN Model

ACKNOWLEDGEMENT

This work is supported by the Universiti Putra Malaysia under the Putra Grant (UPM/800-3/3/1/GP/2017/9588800).

REFERENCES

[1] M. A. Alksher, A. Azman, R. Yaakob, R. A. Kadir, A. Mohamed, and E. Alshari. **A Framework for Idea Mining Evaluation**, In *16th International Conference on New Trends in Intelligent Software Methodology Tools, and Techniques*, SoMeT 2017. IOS Press.

[2] M. A. Alksher, A. Azman, R. Yaakob, R. A. Kadir, A. Mohamed, and E. M. Alshari. 2016. **A review of methods for mining idea from text.** In *Information Retrieval and Knowledge Management (CAMP), 2016 Third International Conference on.* IEEE, 88–93. https://doi.org/10.1109/INFRKM.2016.7806341

[3] M. A. Alksher, A. Azman, R. Yaakob, R. A. Kadir, A. Mohamed, and E. M. Alshari. 2018. **Feasibility of using the position as a feature for idea identification from text**. In *Information Retrieval and Knowledge Management (CAMP), 2018 Fourth International Conference on.* IEEE, 69–74. https://doi.org/10.1109/INFRKM.2018.8464819

[4] I. Atanassova, M. Bertin, and V. Larivière. 2016. **On the composition of scientific abstracts.** *Journal of Documentation* 72, 4 (2016), 636–647. https://doi.org/10.1108/JDOC-09-2015-0111

[5] C. Balasubramanyam, M. S. Ajay, K. R. Spandana, A. B. Shetty, and K. N. Seetharamu. 2014. **Curve fitting for coarse data using artificial neural network**. *WSEAS Transaction on Mathematics* 13 (2014), 406–415.

[6] I. M. Barbosa, E. del M. Hernandez, M.L.C.C. Reis, and O.A.F. Mello. 2006. **Measurement uncertainty contribution to the calibration curve fitting of an aerodynamic external balance using MLP artificial neural network**. In *XVIII Imeko World Congress, Metrology for Sustainable Development.*

[7] P. A. Bekker and K. E. Bouwman. 2009. **Arbitrage smoothing in fitting a sequence of yield curves**. *International Journal of Theoretical and Applied Finance* 12, 05 (2009), 577–588. https://doi.org/10.1142/S0219024909005373

[8] C. M. Bishop and C. M. Roach. 1992. **Fast curve fitting using neural networks**. *Review of scientific instruments* 63, 10 (1992), 4450–4456. https://doi.org/10.1063/1.1143696

[9] S. Ceri, A. Bozzon, M. Brambilla, E. D. Valle, P. Fraternali, and S. Quarteroni. 2013. A**n introduction to information retrieval**. In *Web information retrieval.* Springer, 3–11.

[10] K. Christensen, S. Nørskov, L. Frederiksen, and J. Scholderer. 2017. **In search of new product ideas: Identifying ideas in online communities by machine learning and text mining**. *Creativity and Innovation Management* 26, 1 (2017), 17–30.

[11] I. DiMatteo, C. R. Genovese, and R. E. Kass. 2001. **Bayesian curve-fitting with free-knot splines**. *Biometrika* 88, 4 (2001), 1055–1071. https://doi.org/10.1093/biomet/88.4.1055

[12] T. C. Dinh, H. Bae, J. Park, and J. Bae. 2015. **A framework to discover potential ideas of new product development from crowdsourcing application**. arXiv preprint arXiv:1502.07015 (2015).

[13] Y. Geum and Y. Park. 2016. **How to generate creative ideas for innovation: a hybrid approach of WordNet and morphological analysis**. *Technological Forecasting and Social Change* 111 (2016), 176–187.

[14] M. Gulsen, A. E. Smith, and D. M. Tate. 1995. **A genetic algorithm approach to curve fitting**. *International Journal of Production Research* 33, 7 (1995), 1911–1923. https://doi.org/10.1080/00207549508904789

[15] Y. Guo, A. Korhonen, M. Liakata, I. S. Karolinska, L. Sun, and U. Stenius. 2010. **Identifying the information structure of scientific abstracts: an investigation of three different schemes.** In *Proceedings of the 2010 Workshop on Biomedical Natural Language Processing*. Association for Computational Linguistics, 99–107.

[16] S. Gupta. 2015. **A regression modeling technique on data mining.** *International Journal of Computer Applications* 116, 9 (2015).

[17] A. Hotho, A. Nürnberger, and G. Paaß. 2005. **A brief survey of text mining.** In *Ldv Forum*, Vol. 20. 19–62.

[18] B. M. Kudrowitz and D. Wallace. 2013. **Assessing the quality of ideas from prolific, early-stage product ideation.** *Journal of Engineering Design* 24, 2 (2013), 120–139. https://doi.org/10.1080/09544828.2012.676633

[19] B. Liu, X. An, and J. X. Huang. 2015. **Using term location information to enhance probabilistic information retrieval.** In *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval.* ACM, 883–886.

[20] H. Liu, J. Goulding, and T. Brailsford. 2015. **Towards Computation of Novel Ideas from Corpora of Scientific Text.** In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer, 541–556.

[21] H. M. Majd, M. Poursina, and K. H. Shirazi. 2009. **Determination of barreling curve in upsetting process by artificial neural networks.** In *WSEAS International Conference. Proceedings. Mathematics and Computers in Science and Engineering.* WSEAS.

[22] S. J. Miller. 2006. **The method of least squares.** Mathematics Department Brown University (2006), 1–7.

[23] K. Molugaram and G. S. Rao. 2017. **Chapter 5 - Curve Fitting.** In *Statistical Techniques for Transportation Engineering*, Kumar Molugaram and G. Shanker Rao (Eds.). Butterworth-Heinemann, 281 – 292. https://doi.org/10.1016/B978-0-12-811555-8.00005-2

[24] Ö. Çelik, A. Teke, and H. B. Yıldırım. 2016. **The optimized artificial neural network model with Levenberg–Marquardt algorithm for global solar radiation estimation in Eastern Mediterranean Region of Turkey.** *Journal of cleaner production* 116 (2016), 1–12. https://doi.org/10.1016/j.jclepro.2015.12.082

[25] M. H. Richardson and D. L. Formenti. 1985. **Global curve fitting of frequency response measurements using the rational fraction polynomial method**. In *Proceedings of the Third International Modal Analysis Conference.* sn, 390–397.

[26] P. K. Shukla. 2009. **Levenberg-Marquardt Algorithms for Nonlinear Equations,** *Multi-objective Optimization, and Complementarity Problems*. (2009).

[27] D. Thorleuchter. 2008. **Finding new technological ideas and inventions with text mining and technique philosophy.** In *Data Analysis, Machine Learning and Applications*. Springer, 413–420.

[28] D. Thorleuchter and D. V. den Poel. 2013. **Web mining based extraction of problem solution ideas.** *Expert Systems with Applications* 40, 10 (2013), 3961–3969.

[29] D. Thorleuchter, D. V. den Poel, and A. Prinzie. 2010. **Mining ideas from textual information.** *Expert Systems with Applications* 37, 10 (2010), 7182–7188.

[30] H. Wang and Y. Ohsawa. 2013. **Idea discovery: A scenario-based systematic approach for decision making in market innovation.** *Expert Systems with Applications* 40, 2 (2013), 429–438.

[31] J. Zhao, J. X. Huang, and S. Wu. 2012. **Rewarding term location information to enhance probabilistic information retrieval.** In *Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval.* ACM, 1137–1138.