Gender Statistical Analysis Applied for Identifying Style Patterns in English Academic Writing

Mădălina ZURINI
Bucharest University of Economic Studies, Romania
madalina.zurini@csie.ase.ro

The present paper addresses the problem of writing style patterns in the context of English Academic Writing. Stylometric analysis is used in order to extract the main characteristics obtained from the evaluation of articles written in well-known scientific journals such as Elsevier and Springer. The objective of the paper is to establish a pattern description of articles written in the same domain depending on the gender of the authors. Relevant prior written work upon the current subject reveal different characteristics of writing style of authors from different cultural orientation and gender. The paper describes the main characteristics taken into account for the clustering model when it comes to title, abstract and chapters' construction within the analyzed articles. A short description of the algorithms and tools for clustering and space reduction is presented for further selecting the best combination for the proposed model. An additional statistical layer is added to the current clustering algorithms and space reduction for obtaining statistical proven results of usage. An aggregated structure model is conducted as a result of characteristics selection and processing for future work usage in gender analysis of scientific articles writing. Conclusions and withdrawn along with the future directions extracted from the current work. A database structure is proposed formed out of statistical calculated percentage of papers depending on the author gender. The relevance of the work can be well used as a guide line in writing scientific articles as the main musts in scientific writing are presented.

Keywords: *Stylometry, Gender analysis, Clustering algorithms, Space reduction, Feature selection*

1 Introduction

Stylometry is a very well-spread current field of interest among researchers as using and integrating in the research this domain can result in great conclusions such as object orientation to a particular group. Analyzing the writing style can reveal the author identity, [1], or even the gender of the author of the text paper, [2]. Analysing the text by the concept of Bow, Bag Of Words, and other types of feature collection from a text paper, using stylometry it can also be concluded in a percentage of native language vs. non-native language writers, [3]. While in other domains such as social media, the method of evaluating the text using stylometry can conclude in proper results, in the area of writing scientific papers, just the fact that an author should comply to the writing pattern provided by the submission guideline and editors generates eye related text papers.

But, within this well-common article structure, the writing style of an author can still reveal unique features just by using an in-deeper analysis. This is also what the current research paper aims to succeed, that is selecting that set of characteristics that best suits grouping scientific articles to be written by male or female. Stylometry is an important area in itself, as systems for scientific stylometry would give sociologists new tools for analysing academic communities in scientific journals, and new ways to resolve the nature of collaboration in specific articles [15]. Authors might also use these tools, e.g., to help ensure a consistent style in multi-authored papers [16] or to determine sections of a paper needing revision.

The originality of the research lies within the selection method for the n-dimension space of representation of the clustered and supervised grouped objects. Even though each scientific article written in well-known journals should comply with a set of rules that are sometimes

DOI: 10.12948/issn14531305/22.1.2018.07

very strict or exact, there is still room for in the middle approach of the authors that signed the papers.

The present paper is organized with other four chapters besides the introduction, starting from a wider view of the stylometric analysis of English academic writing that is found in chapter 2. In chapter 3, a comparison of unsupervised clustering algorithms in done in terms of optimization of time and results. A metric of evaluating the correctness of the classification is also proposed and will be used for evaluating the current proposed model.

Combining the stylometry analysis, along with clustering and classification algorithms and tools, a framework for establishing the gender of the authors of English written scientific papers is proposed in chapter 4. This framework can also be expended in other areas of interest and analysis of scientific articles as it can reveal the tendency of well-written articles that were published in renowned and high-ranked journals. Conclusions and future work are withdrawn in chapter 5.

2 Statistical stylometry analysis in English Academic Writing

Information retrieval, automated learning techniques as well as statistical processing of natural language deal with, among other things, extracting content from text documents. The extracted content is then used in areas such as: classifying the text in specific fields, clustering documents to obtain similar document sets from the content point of view, or assigning the author of the paper. Depending on how the document is represented, a representation based on the terms in which the document is composed, the representation of its terms and meanings, the representation based on the document's features, the results of the text-processing can be used to organization, classification or search within a collection of documents.

By combining the level of the author, the document represented by the text level and that of the membership panel, adding generative models, a series of questions find answers, such as: the subject that an author deals with in a document, the authors who would could have written an anonymous document, the intersection of domains, referring to the same set of similar features of documents in different assignment domains.

Adding this approach and a level that treats stylometry in document analysis can help improve the assignment of the author, the domains, and the topics being treated. The level of stylometry also generates a description of the diversity of the author's vocabulary, leading to a model assignment of the documents of a set of authors, advancing in the area of cultural orientation analysis. The assignment of the author, or the ownership of a document to a specific author, the quest to deduce an author's writing characteristics from the character set of documents written by that author as presented in [4] is a long debated issue and a wide range of applications.

A general approach to the document-author-domain link is presented as an initial research point of building an extension in cultural orientation. To this end, the semantic analysis is presented and introduced into the model using WordNet lexical ontology. WordNet is used for documents written in English, providing the possibility to identify the contextual meanings of polysemantic words. It is demonstrated that, by transforming the word level into a word level, an improvement in the performance of information retrieval techniques in text documents is obtained, subsequently generating performance in document representation as well as in processing it using automated grading techniques, such as supervised and unsupervised.

Most previous papers on classification of documents in multiple categories use data sets with relatively few assignment categories and many instances of training [5]. In [6], models based on multinomial or multivariate distributions of Bernoulli type are presented as vast methods encountered in document representation. In current research, k-means spherical algorithms, along with the properties involved in document clustering, are used in special cases of generative models. Genetic patterns for text usually associate a multinomial with each association class or domain [7] and [8].

DOI: 10.12948/issn14531305/22.1.2018.07

The statistical analysis of style, stylometry, as described in [9], proceeds from the demonstrated presumption that an author's style has certain characteristics that cannot be identified by direct manipulation. Thus, these are the components underlying the methods of identifying the author of anonymous works or not. An author's style may vary over time due to the various areas in which it writes or personal development. Generally, stylometry should identify features that do not depend on these changes, but are sufficient to characterize and differentiate one author from another.
Two different patterns for extracting stylometry from text documents involve the use of distinctive description features of the style that characterizes a specific author and patterns that focus on extracting the general semantic content of a document rather than the details of the style of the person who wrote the document.
The probabilistic generative model reduces the process of writing a text document to a series of simple probabilistic steps. The first step is to generate those probabilistic models by retrieving the document into a set of words and their number of occurrences in the text. The plurality of words can be made up of the first words most commonly used. Choosing the optimal value for the n variable is made based on maximizing the percentage of information retained in the context of minimizing the number of word-type features used to describe and model the manipulated objects.
In order to extract that set of writing style features that define at maximum the lexical, semantic and cultural components of an author through the specialized work written by him, the initial set of characteristics is defined as the set on which they are run different combinations. Thus, the set of writing style features consists of the components:

- average word length;
- the average length of the sentences, measured in number of words;
- the number of link words relative to the total number of words identified in the documents analyzed;
- the frequency of use of special signs;
- Type-Token vocabulary wealth;
- the semantic vocabulary semantics;
- Frequency of speech parts.

Also, besides the set of seven writing style features, two features are described that describe the semantic component, characteristics that are calculated using the WordNet ontology.
In this respect, the following variables are used:

- the word on position i in the set of words that make up the analyzed document;
- Contextual meaning that is returned for the word using the WSD component, Word Sense Disambiguation, available in the WordNet ontology;
- the weight of the context of the word, weight which is taken from the WordNet lexical ontology and which is calculated using a set of training;
- ISC is the indicator of contextual meanings used by an author, on average in the specialized works that have him as the author;
- IPSC is the weighted indicator of contextual meanings used by an author, on average in the specialized works that have it as the author, weighted with the probability of occurrence of the meanings in the WordNet ontology.

The two proposed indicators, ISC and IPSC, complement the initial set of characteristics by integrating the analysis of the use of the common or non-polar meanings of polysemantic words. ISC, Contextual Context Indicator, is calculated based on the formula:

$$ISC = \frac{\sum_{i=1}^{n} s(w_i)}{n}$$

where:

- n represents the cardinality of the set of words extracted from the analyzed document, the set of words being not reduced by eliminating the redundant words, due to the possibility of using multiple-sense words within the same document.

On the other hand, IPSC, the Weighted Contextual Meaning Indicator, derives from the ISC indicator, but is improved by integrating the probabilities of occurrence of each contextual sense using the formula:

DOI: 10.12948/issn14531305/22.1.2018.07

$$IPSC = \frac{\sum_{i=1}^{n} s(w_i) \times \frac{1}{p(s(w_i))}}{n}$$

IPSC is a variable inversely proportional to the percentage of occurrence of contextual meanings of polysemanic words:

$$\begin{cases} p(s(w_i)) \to 1 \Rightarrow \frac{1}{p(s(w_i))} \to 0 \Rightarrow IPSC \to 0 \\ p(s(w_i)) \to 0 \Rightarrow \frac{1}{p(s(w_i))} \to \infty \Rightarrow IPSC \to \infty \end{cases}$$

A value of the IPSC variable of 0 indicates the use of common contextual meanings, while a value of the indicator $IPSC \to \infty$ leads to an interpretation that the author usually uses the less contextual meanings of polysemantic words.

Table 1 contains the initial set of writing style characteristics separated in the two areas of interest: lexical features and semantic features.

Table 1 Set of writing style features categorized in the two analysis directions

Lexical characteristics	Semantic characteristics
Average word length	Contextual Meter Indicator
Average word length, measured in number of words	Weighted Contextual Meaning indicator
The number of link words relative to the total number of words identified in the documents analyzed	The richnessh of the Type - Token vocabulary
Frequency of use of special signs, {, ; . !? @#$%& ∗ (){}[]}	Vocabulary semantic wealth
	Frequency of parts of speech

In order to choose the set that best characterizes from the point of view of the cultural affiliation of the authors of the specialized works is defined the set of combinations with NC cardinality, which can be generated from the nine characteristics listed in Table 1, so:

$$NC = C_9^1 + C_9^2 + \cdots + C_9^9 = 2^9 - 1 \; combinations$$

The choice of the optimal combination is performed using the objective function that fulfils the cluster formation conditions in the non-supervised classification:

- minimizing inter-cluster dispersion;
- maximizing intra-cluster dispersion.

Thus, the combination of the use of the writing style features is chosen so that the groups of objects formed are as compact and at the same time spaced apart.

3 Clustering and metric space reduction algorithms and tools

Found at the intersection of fundamental domains such as computer science, knowledge theory, decision theory, geometry, probabilistic theory and statistical mathematics, the theory of form recognition knows in present applications of which sphere is widely spread from the anthropological research down to hardware and software design, [10]. Theory of

DOI: 10.12948/issn14531305/22.1.2018.07

form recognition is defined by the rules, principles, methods, decision and analysis tools used with the aim of identifying the membership of objects, units, phenomena, actions, processes the different sets with well-defined individuality.

The theory of form recognition is divided in two categories: supervised and unsupervised learning. The classifiers are part of the supervised learning and use information about the membership of an object in order to classify new objects in one of the defined sets, while unsupervised learning, represented by the clustering analysis, groups the initial set of objects according to their characteristic, generating partitions of the initial set of objects.

Data clustering, also known as clustering analysis, is defined by Webster, Merriam-Webster Online Dictionary, as being o statistical classification technique for discovering if the individuals from a population can be divided in different groups by quantitative comparing of their characteristics. The objectives of clustering analysis are given by:

- the understanding of structures, in order to generate hypothesis upon the data or to detect abnormalities;
- the natural classification, for identifying the degree of similarity among forms;
- the compression of data, as a method of data organization and summary in structures given by the clusters.

The evaluation of cluster methods is done at the:

- local level, for the evaluation of the results obtained by a particular clustering method that varies depending on the input parameters, like the number of cluster, optimization function or the finishing point, resulting a number of local optimum equal to the number of different clustering methods used;
- global level, for the evaluation of the results of the local optimum and choosing the one that maximizes the optimization function, resulting a global optimum that characterizes the best the initial set of objects.

Starting from the initial set of objects, X, each object has a domain membership associated to it, domains that are initial set to, $T_1, T_2, \dots, T_k$, where k represents the total number of classes the objects are assigned in.

$$T_i = \{x_j | j \in \{1,2, \dots, m\}, t(x_j) = i\}$$

where:

- $t(x_j)$ represents the class in which object x_j is assigned from the set of objects X.

Based on the partitions $T_1, T_2, \dots, T_k$ of the initial X set, the correlation between objects matrix is formed, *MCP*, in terms of objects' interaction from the same partition, with $MCP \in \mathcal{M}_{m\times m}\{0,1\}$, thereby:

$$mcp_{ij} = \begin{cases} 1, t(x_i) = t(x_j) \\ 0, t(x_i) \neq t(x_j) \end{cases}$$

After applying each clustering method proposed, the result is given by the number of resulted clusters, k, and a set of partitions, $C_1, C_2, \dots, C_k$. With these partitions, the correlation between the objects operator is applied, resulting the correlation matrix of the objects clustered, *MCC*, with $MCC \in \mathcal{M}_{m\times m}\{0,1\}$, defined by:

$$mcc_{ij} = \begin{cases} 1, c(x_i) = c(x_j) \\ 0, c(x_i) \neq c(x_j) \end{cases}$$

where:

- $c(x_j)$ represents the cluster in which the object x_j is clustered.

The matrix of cluster evaluation, *MEC*, by comparing the initial grouping in k sets with the sets given by the clustering analysis, is computed out of the *MCP* and *MCC* matrixes:

$$mec_{ij} = \begin{cases} 1, (mcp_{ij} + mcc_{ij})\%2 = 0 \\ 0, (mcp_{ij} + mcc_{ij})\%2 = 1 \end{cases}$$

where:

- x*%y* returned the rest of dividing *x* to *y*.

Using the clustering evaluation matrix, the Indicator for Cluster Evaluation, *ICE*, is calculated, and measures the degree of clustering correctness reported to the prior grouping of the analyzed objects, thereby:

$$ICE = \frac{\sum_{i=1}^{m}\sum_{j=1}^{m} mec_{ij}}{m^2} \times 100$$

DOI: 10.12948/issn14531305/22.1.2018.07

The *ICE* metric takes values in the [0%; 100%] interval. *ICE=0%* if no cluster corresponds according to the objects' interactions from the same group with the initial grouping, and *ICE=100%* if all the clusters are identical to the initial *k* sets.

4 Proposed model for domain oriented pattern construction in gender clustering

Combining the algorithms and tools presented in the current research paper, figure 1 contains the major steps which should be followed in order to gain a pattern description of the analyzed research papers

.

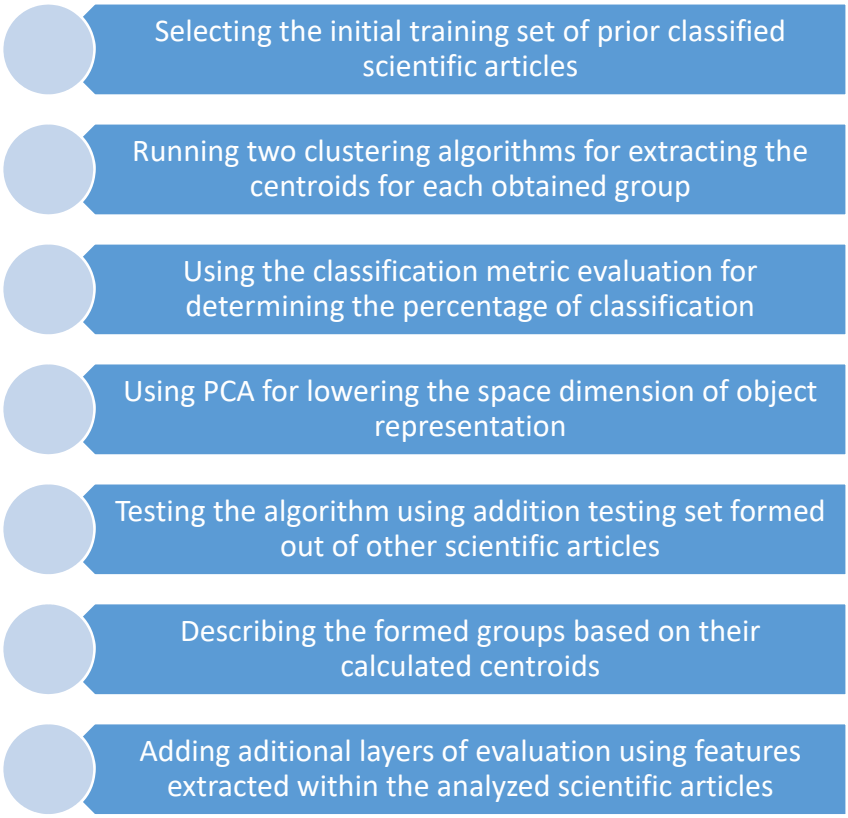

Fig. 1. Methodology for domain oriented pattern construction in gender clustering

Step 1, Selecting the initial training set of prior classified scientific articles, imply is determining the dimension of desired classification. If a wide range of domain oriented scientific articles are inserted within the training set, a big number of articles are needed for each proposed domain group. For determining the domain orientation of the article, WordNet lexical ontology is used. Under lemmatizing and stemming algorithms, the desired parts of the articles are divided into words and the token is extracted using stemming algorithms. This involves the space dimension of the article seen as a Bag Of Words. Using the similarity metric proposed within WordNet lexical ontology, each token is computed and each article is classified within the closest domain set. This technique does not involve prior manual classification, as the running algorithm can obtain the results.

Following step 2, with the run of two different clustering algorithms, DBSCAN and k-Means, each group of domain articles are computed in their centroid representation. This representation, again done under Bag of Words space-dimension, represent the mean of each domain in terms of words used. For this step, a prior reduction of the common words can be done, extracting from the analysis the words that are probabilistic found similar within each analyzed domain.

Using the results from step 2, the centroids, Principal Component Analysis is conducted again in order to diminish the space dimension

DOI: 10.12948/issn14531305/22.1.2018.07

of the representation of the objects. Selecting the best features that characterize the articles in the terms of the current classification. The objective of this step is to reduce the time consumption for running the whole model for generating pattern in English written scientific articles. Similar results were obtained also by studies such as [17] and [18].

For proper testing of the obtained centroids, additional testing using cross-fold technique are done within step 3, Testing the algorithm using additional testing set formed out of other scientific articles. The technique involves putting aside a percentage of articles from the initial test that were prior classified and used just for the testing phase. This concludes in a more rigorous testing as cross-fold is recommended to be used several times until a threshold is obtained.

In step 4, Describing the formed groups based on their calculated centroids, a text description is done for each domain group using the mean of each group.

For further stylometry analysis, additional characteristics can be added within the present classification, for step 5, such as: abstract length, percentage of common words, percentage of domain oriented tokens used. The title of each article can be analyzed separated from the abstract and other text parts of the article, as this combination of words are stronger related to the general article.

5 Conclusions and future work

The proposed model for domain oriented pattern construction in gender clustering and further classification combines an initial layer of stylometry analysis upon the writing style of authors male and female with the supervised and unsupervised clustering techniques available. Using an initial supervises classified scientific articles the best set of writing style features are extracted. The writing style features contains semantic and lexical analysis, while combining a lexical ontology such as WordNet in order to compute the results to each analyzed domain oriented text paper.

Creating a writing pattern for each domain oriented scientific paper can decrease the level of generality, thus creating more concrete patterns for evaluating the papers. A further division is done within this set of articles, the gender of the authors. Future work is focused on implementing the proposed model in order to obtain the desired patterns. Addition refinement may be needed for optimal results.

References

[1] Frederick Mosteller and David L. Wallace. Applied Bayesian and Classical Inference: The Case of the Federalist Papers. Springer-Verlag, 1984

[2] Shlomo Argamon, Moshe Koppel, Jonathan Fine, and Anat Rachel Shimoni. 2003. Gender, genre, and writing style in formal written texts. Text, 23(3), August.T.

[3] Xuan Le, Ian Lancashire, Graeme Hirst, and Regina Jokel. 2011. Longitudinal detection of dementia through lexical and syntactic changes in writing: A case study of three British novelists. Literary and Linguistic Computing, 26(4):435–461.

[4] Juola, P 2006 "Authorship attribution", Foundation and Trends in information Retrieval, Vol 1, no. 3, pp. 233-234

[5] Rubin, T, Chambers, A, Smyth, P & Steyvers, M 2012 "Statistical topic models for multi-label document classification", Machine Learning, Vol. 88, no. 1-2, pp. 157-208

[6] Zhong, S & Ghosh, J 2003 „A Comparative Study of Generative Models for Document Clustering", SIAM International Conference Data Mining Workshop on Clustering High Dimensional Data and Its Applications

[7] Eisenstein, J, Ahmed, A & Xing, E.P 2011 "Sparse additive generative models for text", in Proceedings of the 28th International Conference on Machine Learning, ICML-11, pp. 1041-1048

[8] Rosen-Zvi M 2004 "The author-topic model for authors and documents", in Proceedings of the 20th conference on Uncertainty in artificial intelligence AUAI Press., pp. 487-494

[9] Diederich, J, Kindermann, J, Leopold, E & Paass, G 2003 "Authorship attribution

DOI: 10.12948/issn14531305/22.1.2018.07

with support vector machine", Applied Intelligence, Vol. 19, no. 1-2, pp. 109-123, 2003

[10] B. Stein, N. Lipka and P. Prettenhofer, "Intrinsic plagiarism analysis", Language Resources & Evaluation, 2010, Vol. 45, No. 1, pp. 63-82

[11] Dominique Estival, Tanja Gaustad, Son-Bao Pham, Will Radford, and Ben Hutchinson. 2007. Author profiling for English emails. In Proc. PACLING, pages 263–272.

[12] E. Stamatatos, "Author identification: Using text sampling to handle the class imbalance problem", Inf. Process Manage, 2008, vol. 44, pp. 790-799

[13] S. Benno, K. Moshe and S. Efstathios, „Plagiarism analysis, authorship identification and near-duplicate detection", Proceedings ACM SIGIR Forum PAN'07, 2007, New York, pp. 68-71

[14] Data Mining Algorithms in R/Clustering/Hybrid Hierarchical Clustering, Available online at: https://en.wikibooks.org/wiki/Data_Mining_Algorithms_In_R/Clustering/Hybrid_Hierarchical_Clustering

[15] M-E. Osiceanu, „Considerații privind drepturile de proprietate intelectuală în știință, tehnică și artă sau între creație și plagiat", Available online at: http://api.ning.com/files/uPa7BpseSwF6lqvQmgiaPdijUqzZEL9nHLQz-kOJht94wzdjkfubWxs5cGMb-kITg3agVjj0s2dOhxhjn88Hy*72*M4OH2MIVb/Osiceanu_MEConsideraiiprivinddrepturiledeproprietateintelectuala_final.pdf

[16] Angela Glover and Graeme Hirst. 1995. Detecting stylistic inconsistencies in collaborative writing. In Writers at work: Professional writing in the computerized environment, pages 147–168.

[17] Shawndra Hill and Foster Provost. 2003. The myth of the double-blind review?: Author identification using only citations. SIGKDD Explor. Newsl. 5:179–184.

[18] Graeme Hirst and Ol'ga Feiguina. 2007. Bigrams of syntactic labels for authorship discrimination of short texts. Literary and Linguistic Computing, 22(4):405–417.

[19] Nikhil Johri, Daniel Ramage, Daniel McFarland, and Daniel Jurafsky. 2011. A study of academic collaborations in computational linguistics using a latent mixture of authors model. In Proc. 5th ACL-HLT Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities, pages 124–132.

[20] Oren Tsur and Ari Rappoport. 2007. Using classifier features for studying the effect of native language on the choice of written second language words. In Proc. Workshop on Cognitive Aspects of Computational Language Acquisition, pages 9–16.

[21] Dragomir R. Radev, Mark Thomas Joseph, Bryan Gibson, and Pradeep Muthukrishnan. 2009a. A bibliometric and network analysis of the field of computational linguistics. Journal of the American Society for Information Science and Technology.

[22] Delip Rao, Michael Paul, Clay Fink, David Yarowsky, Timothy Oates, and Glen Coppersmith. 2011. Hierarchical bayesian models for latent attribute detection in social media. In Proc. ICWSM, pages 598–601.

[23] Harald Baayen, Fiona Tweedie, and Hans van Halteren. 1996. Outside the cave of shadows: Using syntactic annotation to enhance authorship attribution. Literary and Linguistic Computing, 11(3):121–132

DOI: 10.12948/issn14531305/22.1.2018.07

Mădălina ZURINI is currently a teaching assistant in the field of Economic Informatics. She graduated the Faculty of Cybernetics, Statistics and Economic Informatics (2008) and a master in Computer Science in 2010. In 2013 she defended her PhD research with the title "*Spatial representations and knowledge processing using ontologies*". She published more than 20 articles in collaboration or as single author. Her fields of interest are data classification, artificial intelligence, data quality, algorithm analysis and optimizations.

DOI: 10.12948/issn14531305/22.1.2018.07