ANALYZING AUTHORS AND ARTICLES USING KEYWORD EXTRACTION, SELF-ORGANIZING MAP AND GRAPH ALGORITHMS

Tommi Vatanen, Mari-Sanna Paukkeri, Ilari T. Nieminen and Timo Honkela

Adaptive Informatics Research Centre, Helsinki University of Technology,
P.O.Box 5400, FIN-02015 TKK, FINLAND,
E-mail: first.last@tkk.fi

ABSTRACT

In order to analyze the scientific interests and relationships of the participants of the International and Interdisciplinary Conference on Adaptive Knowledge Representation and Reasoning 2008 (AKRR'08) conference, we have developed SOMPA environment (Self-Organizing Maps of Papers and Authors). SOMPA is a software with web-based interface for collecting and analyzing information on authors and their papers. It produces graphical output including graphs and maps. The program extracts keywords for the papers using Likey keyphrase extraction utility. Keywords are used to draw a self-organizing map which is the main end result of SOMPA.

1. INTRODUCTION

As AKRR'08 is an interdisciplinary conference, the participants may not know each others' research areas very well beforehand. The aim of the work reported here is to provide means for the conference participants to familiarize themselves with each others' research interests and topics. An earlier similar work was [1] in which a self-organizing map of Workshop on Self-Organizing Maps 1997 (WSOM '97) abstracts was created. The main methodological differences and extensions in comparison with the WSOM'97 map are 1) an automatic keyphrase extraction method called Likey has been used, 2) the map includes both contributions to the conference as well as other scientific articles by the participants and closely related articles, and 3) the data input is based on a web-based system[1]. Moreover, this work also includes a graph that shows the coauthoring relationships between the participants and a collection of related researchers.

2. DATA COLLECTION

An important factor in proper data analysis is extensive material. To courage people to contribute data collection we have devoted lots of time for developing a pleasant web user interface for SOMPA. The usability of the web page can have a huge impact on the behaviour and interest of the users[2].

We also implemented BibTeX importing in SOMPA. Because BibTeX formatting has several inconsistent practices, extensive regular expression subtituting and parsing needed to be done. If author provides links for his BibTeX entries through a URL field, SOMPA is also able to fetch the documents and include them on the SOM. Without links, documents are still useful for drawing connection graphs.

[1] http://cog.hut.fi/sompa/sompa.cgi

3. SELF-ORGANIZING MAP

The main product of Sompa is a self-organizing map [3] of all authors and articles with keywords. This chapter describes the process of creating the SOM.

3.1. Preprocessing

SOMPA has to go through a long preprocessing procedure, because original texts extracted mainly from portable document format (PDF) files have many elements that do not belong to the actual interesting content such as formatting instructions, variable names, etc. First everything before the abstract and after the beginning of the reference list is removed. Several regular expression substitutions are used to remove in-text references, variable names and other irrelevant expressions. Mathematical formulas are removed by a heuristic algorithm.

3.2. Keyword extraction

We use Likey keyphrase extraction utility to extract keywords from the text [4]. As a reference we use Europarl corpus. Because Likey is language independent, it provides a possibility to extract keywords from articles written in other languages also. By default Likey extracts also keyphrases longer than unigrams, but for the SOM creation, we extract only single keywords.

A total of one hundred keywords are extracted for every article. This seems to provide mostly reasonable keywords, according to qualitative evaluation.

After extraction we stem the keywords for better correspondence between documents. For example, words *discontinuous* and *discontinuities* have a common stem *discontinu*. If two keywords have a common stem, they most probably have a similar meaning [5]. Stemming also reduces dimensions from the SOM input matrix.

Table 1. Kohonen number for some researchers

Erkki Oja	1
Samuel Kaski	1
Włodzisław Duch	2
Eero Castrén	2
Marie Cottrell	3
José Príncipe	3
Patrick Letrémy	4
Philippe Grégoire	4

3.3. Keyword weighting

Weights for keywords in different articles are calculated using modified *tf.idf* -method,

$$\textit{weight(i,j)} = tf_{i,j} \cdot idf_i \tag{1}$$

where i is keyword, j is document and term frequency $tf_{i,j}$ is "normalized" by dividing it by the total number of words in the corresponding document:

$$tf_{i,j} = \frac{n_{i,j}}{\sum_k n_{k,j}} \tag{2}$$

where $n_{i,j}$ is frequency of keyword i in document j and $\sum_k n_{k,j}$ is total number of words in document. We take the logarithm of the document frequency to nullify keywords that occur in all documents:

$$idf_i = \log \frac{|D|}{|\{d_j : t_i \in d_j\}|} \tag{3}$$

where j, $|D|$ is number of documents in the database and $|\{d_j : t_i \in d_j\}|$ is number of documents in which keyword i appears.

3.4. SOM input matrix

For the SOM, we still need to do some preprocessing to simplify the input matrix. Trivially the keywords found in only one document are ignored. Second, oneletter keywords are ignored completely and twoletter keywords are filtered with a twoletter acronym whitelist (AI, AC, AV, etc.). The ignored words are variable names in equations with a high probability. Finally keywords are scanned for all author names in the database to be ignored. This is because it turned out that the surname of the article's main author was very often found in the keyword list.

Besides creating input vectors for articles, we also calculated vectors for the authors closely related to AKRR'08 themes. To obtain the *tf* values for the keywords of an authors we treat the articles of the author as one large document, from which the *tf* values for the keywords are calculated.

Eventually, an article has an average of 58 keywords (out of 100) used in the SOM input vector. This resulted vectors with 1290 keywords (features) in our sample material of 116 articles.

3.5. The interactive SOM

The produced SOM on the SOMPA web page is two and half dimensional. The colouring is calculated by projecting the data vectors using Principal Component Analysis (PCA) and heuristic `som_colorcode` function of Matlab SOM Toolbox [6].

We have implemented several interactive properties on the SOM. Users can trace the locations of the articles and authors as well as distribution of articles of a single author. Clicking on the cells displays contents of the cell and performs mutual keyword comparison if there are several articles or authors in the cell.

Visit the SOMPA web site[2] to experiment with the interactive SOM.

3.6. The SOM of the conference talks

On the SOM in Figure 1 we have presented the relationships of the contributions in the two AKRR conferences (2005 and 2008) and the second European Symposium on Time Series Prediction (ESTSP'08). Definitions of the tags on the map can be found in the tables 2 and 3. The SOM was trained using the prevailing SOMPA database, which included 116 articles. The gray scale colouring of the map represents the topography of the map, darker tones standing for greater distance between the cells.

4. CONNECTION GRAPHS

Second important feature of SOMPA is author connection tracing. SOMPA uses basic graph algorithms to find connections between authors. Two authors are connected if they have shared papers or co-authors, or if their coauthors are connected recursively.

4.1. Distance counting

We use modified breadth-first search (BFS) to determine all shortest paths between two people in the database. Distance in this case is defined so that people have distance of one with their coauthors. If the shortest distance between a co-author and person A is k, then the distance between the author and person A is $k+1$. The obtained results are based on the database material, and doesn't exclude the possibility of the "real distance" being shorter.

4.2. Kohonen number

We introduce Kohonen number honoring the academician Teuvo Kohonen. The Kohonen number is a way of describing the "collaborative distance" between an author and Kohonen.

With help of the bibliography of SOM papers [7, 8, 9] we can have extensive network of papers related to research topics of Kohonen. Table 1 shows a preliminary Kohonen number for a selection of researchers.

Figure 2 illustrates the graph drawing capabilities of SOMPA. It shows connections between selected authors, Kohonen in the middle. On the graph, only edges between people with consecutive Kohonen numbers are drawn.

[2] http://cog.hut.fi/sompa/sompa.cgi

5. FUTURE WORK

We intend to expand our database considerably by the AKRR'08 conference. This should improve the quality of the maps and make Kohonen number tracing more consistent.

To improve the keyword extraction, a method taking advantage of the structure of the scientific paper could be used. The sentence structure of the English language could be also taken into account. On the other hand, using Likey with a reference corpus collected from scientific articles would probably improve the results.

6. REFERENCES

[1] Krista Lagus, "Map of WSOM'97 abstracts—alternative index," in *Proceedings of WSOM'97, Workshop on Self-Organizing Maps, Espoo, Finland, June 4-6*, pp. 368–372. Helsinki University of Technology, Neural Networks Research Centre, Espoo, Finland, 1997.

[2] Jakob Nielsen, "Usability for the masses," *Journal of Usability Studies*, vol. 1, 2005.

[3] Teuvo Kohonen, *Self-Organizing Maps*, (Springer Series in Information Sciences, 30). Springer, 3nd edition, 2001.

[4] Mari-Sanna Paukkeri, Ilari T. Nieminen, Matti Pöllä, and Timo Honkela, "A language-independent approach to keyphrase extraction and evaluation," in *Proceedings of the 22nd International Conference on Computational Linguistics, Coling'08*, 2008.

[5] Martin Porter, "An algorithm for suffix stripping," *Program*, vol. 14, no. 3, pp. 130–137, 1980.

[6] Juha Vesanto, Johan Himberg, Esa Alhoniemi, and Juha Parhankangas, "Self-organizing map in matlab: the som toolbox," in *In Proceedings of the Matlab DSP Conference*, 1999, pp. 35–40.

[7] Samuel Kaski, Jari Kangas, and Teuvo Kohonen, "Bibliography of self-organizing map (SOM) papers: 1981–1997," .

[8] Merja Oja, Samuel Kaski, and Teuvo Kohonen, "Bibliography of self-organizing map (SOM) papers: 1998-2001 addendum," *Neural Computing Surveys*, vol. 1, pp. 1–176, 1998.

[9] M. Pöllä, T. Honkela, and T. Kohonen, "Bibliography of self-organizing map (SOM) papers: 2002-2005 addendum," *Neural Computing Surveys, forthcoming*, 2007.

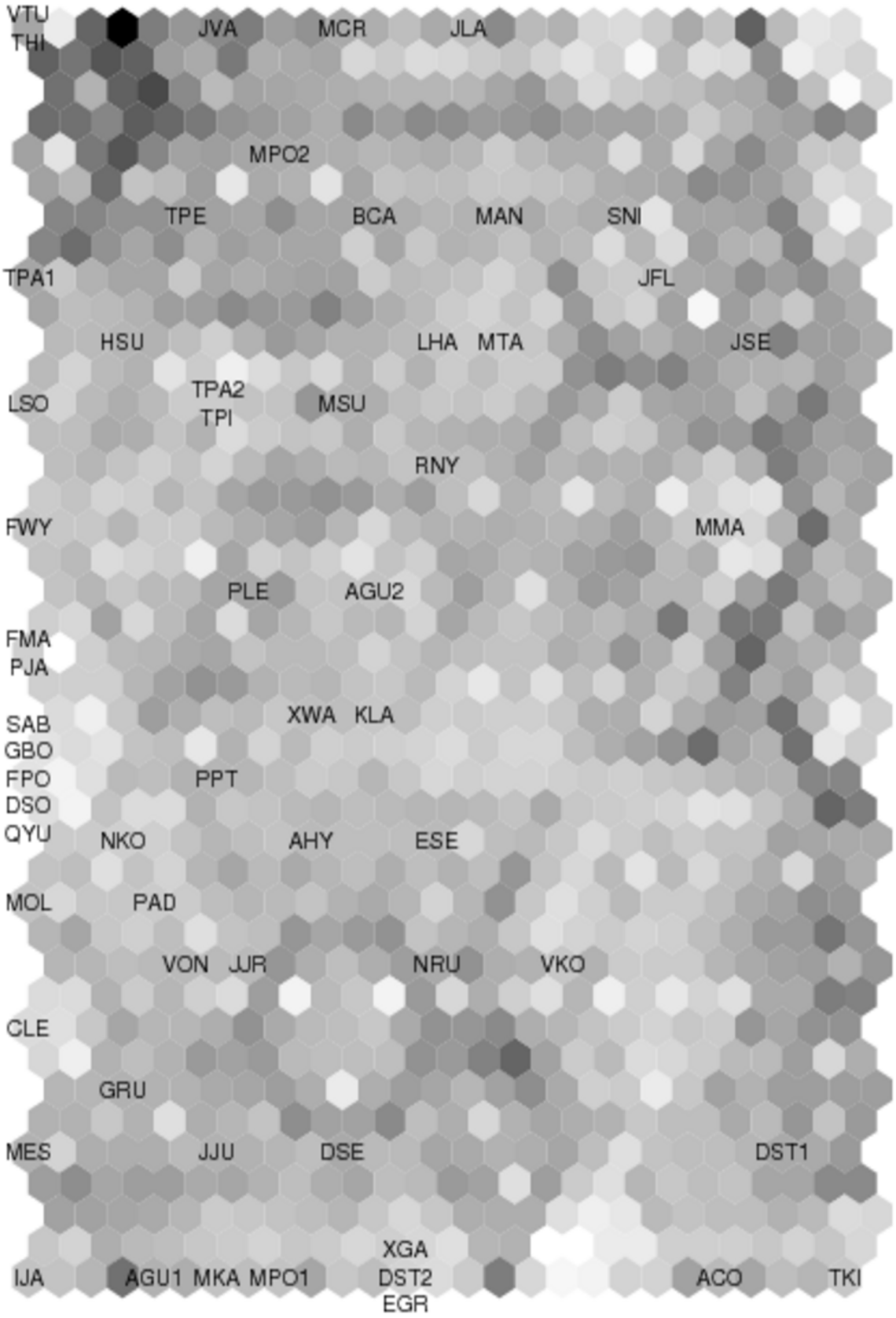

Figure 1. The self-organizing map of the conference talks

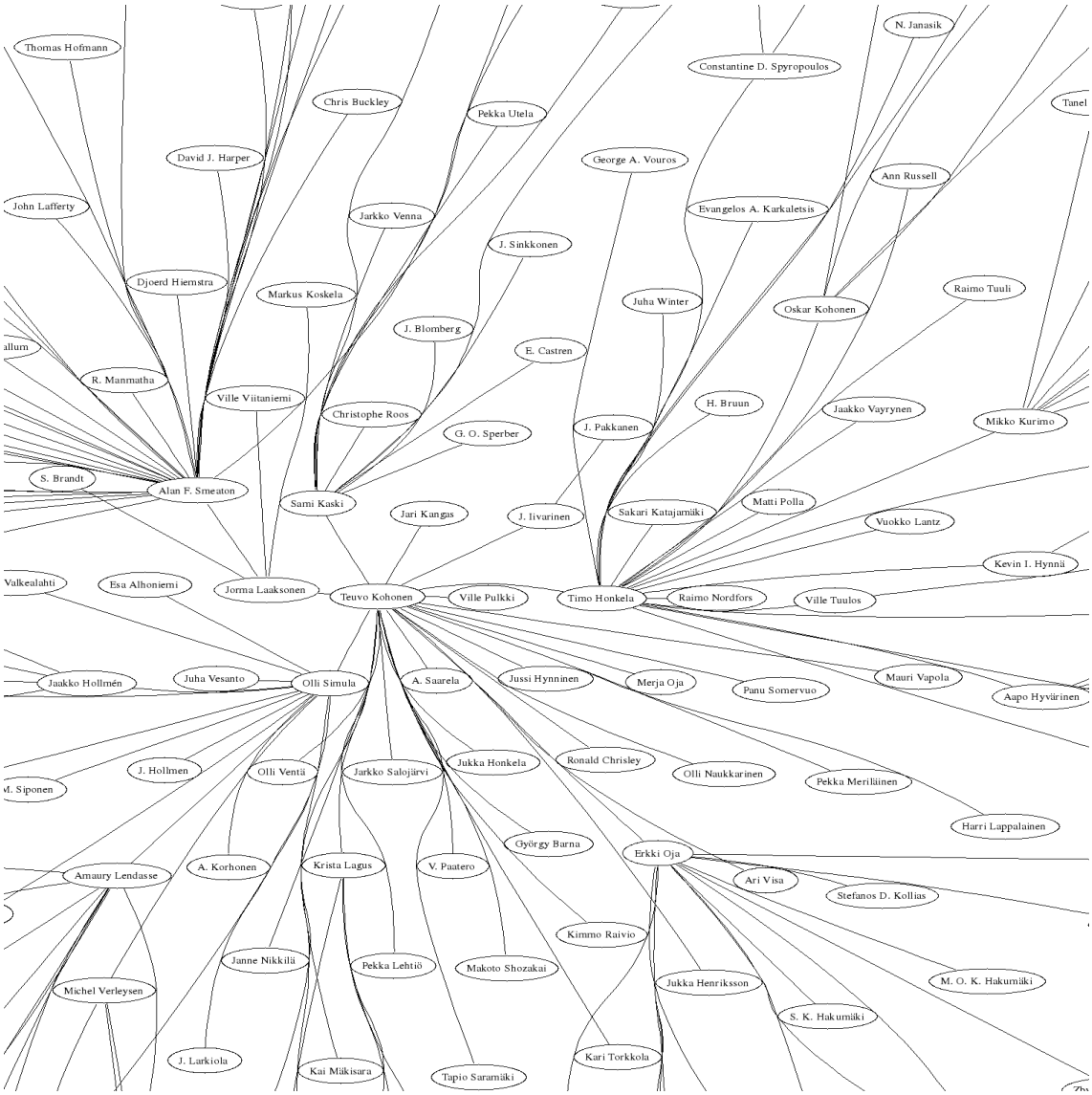

Figure 2. A graph illustrating connections of selected people and academician Teuvo Kohonen

Table 2. Contributions to the AKRR'08 and AKRR'05

ACO	Andrew Coward, Tom Gedeon: Physiological Representation of Concepts in the Brain (AKRR'05)
AHY	Aapo Hyvärinen, Patrik Hoyer, Jarmo Hurri, Michael Gutman: Statistical Models of Images and Early Vision (AKRR'05)
BCA	Basilio Calderone: Unsupervised Decomposition of Morphology a Distributed Representation of the Italian Verb System (AKRR'08)
EGR	Eric Grégoire: About the Limitations of Logic-Based Approaches to the Formalisation of Belief Fusion (AKRR'05)
DST1	Dimitrios Stamovlasis: A Catastrophe Theory Model For The Working-Memory Overload Hypothesis - Methodological Issues (AKRR'08)
DST2	David Stracuzzi: Scalable Knowledge Acquisition Through Memory Organization (AKRR'05)
HSU	Hanna Suominen, Tapio Pahikkala, Tapio Salakoski: Critical Points in Assessing Learning Performance via Cross-Validation (AKRR'08)
JFL	John Flanagan: Context Awareness in a Mobile Device: Ontologies versus Unsupervised/Supervised Learning (AKRR'05)
JLA	Jorma Laaksonen, Ville Viitaniemi, Markus Koskela: Emergence of Semantic Concepts in Visual Databases (AKRR'05)
JSE	Jan Sefranek: Knowledge Representation For Animal Reasoning (AKRR'08)
JVA	Jaakko Väyrynen, Timo Honkela: Comparison of Independent Component Analysis and Singular Value Decomposition in Word Context Analysis (AKRR'05)
LHA	Lars Kai Hansen, Peter Ahrendt, Jan Larsen: Towards Cognitive Component Analysis (AKRR'05)
KLA	Krista Lagus, Esa Alhoniemi, Jeremias Seppä, Antti Honkela, Paul Wagner: Independent Variable Group Analysis in Learning Compact Representations for Data (AKRR'05)
MAN	Mark Andrews, Gabriella Vigliocco, David Vinson: Integrating Attributional and Distributional Information in a Probabilistic Model of Meaning Representation (AKRR'05)
MCR	Mathias Creutz, Krista Lagus: Inducing the Morphological Lexicon of a Natural Language from Unannotated Text (AKRR'05)
MMA	Michael Malý: Cognitive Assembler (AKRR'08)
MPO1	Matti Pöllä, Tiina Lindh-Knuutila, Timo Honkela: Self-Refreshing SOM as a Semantic Memory Model (AKRR'05)
MPO2	Matti Pöllä: Change Detection Of Text Documents Using Negative First-Order Statistics (AKRR'08)
MTA	Martin Takac: Developing Episodic Semantics (AKRR'08)
NRU	Nicolas Ruh, Richard P. Cooper, Denis Mareschal: A Reinforcement Model of Sequential Routine Action (AKRR'05)
PLE	Philippe Leray, Olivier François: Bayesian Network Structural Learning and Incomplete Data (AKRR'05)
SNI	Sergei Nirenburg, Marjorie McShane, Stephen Beale, Bruce Jarrell: Adaptivity In a Multi-Agent Clinical Simulation System (AKRR'08)
THI	Teemu Hirsimäki, Mathias Creutz, Vesa Siivola, Mikko Kurimo: Morphologically Motivated Language Models in Speech Recognition (AKRR'05)
TKI	Toomas Kirt: Search for Meaning: an Evolutionary Agents Approach (AKRR'08)
TPA1	Tapio Pahikkala, Antti Airola, Jorma Boberg, Tapio Salakoski: Exact and Efficient Leave-Pair-Out Cross-Validation for Ranking RLS (AKRR'08)
TPA2	Tapio Pahikkala, Sampo Pyysalo, Jorma Boberg, Aleksandr Mylläri, Tapio Salakoski: Improving the Performance of Bayesian and Support Vector Classifiers in Word Sense Disambiguation using Positional Information (AKRR'05)
TPE	Tatjana Petkovic, Risto Lahdelma: Multi-Source Multi-Attribute Data Fusion (AKRR'05)
TTE	Tommi Tervonen, Jose Figueira, Risto Lahdelma, Pekka Salminen: An Approach for Modelling Preferences of Multiple Decision Makers (AKRR'05)
VKO	Ville Könönen: Hierarchical Multiagent Reinforcement Learning in Markov Games (AKRR'05)
VTU	Ville Tuulos, Tomi Silander: Language Pragmatics, Contexts and a Search Engine (AKRR'05)
XGA	Xiao-Zhi Gao, Seppo Ovaska, Xiaolei Wang: Re-editing and Censoring of Detectors in Negative Selection Algorithm (AKRR'08)
XWA	Xiaolei Wang, Xiao-Zhi Gao, Seppo Ovaska: A Simulated Annealing-Based Immune Optimization Method (AKRR'08)

Table 3. Contributions to the ESTSP'08

AGU1	Alberto Guillen, L.J. Herrera, Gines Rubio, Amaury Lendasse, Hector Pomares, Ignacio Rojas: Instance or Prototype Selection for Function Approximation using Mutual Information
AGU2	Alberto Guillen, Ignacio Rojas, Gines Rubio, Hector Pomares, L.J. Herrera, J. Gonzlez: A New Interface for MPI in MATLAB and its Application over a Genetic Algorithm
CLE	Christiane Lemke, Bogdan Gabrys: On the benefit of using time series features for choosing a forecasting method
DSE	D.V Serebryakov, I.V. Kuznetsov: Homicide Flash-up Prediction Algorithm Studying
DSO	Dušan Sovilj, Antti Sorjamaa, Yoan Miche: Tabu Search with Delta Test for Time Series Prediction using OP-KNN
ESE	Eric Séverin: Neural Networks and their application in the fields of corporate finance
FMA	Fernando Mateo, Amaury Lendasse: A variable selection approach based on the Delta Test for Extreme Learning Machine models
FMO	Federico Montesino Pouzols, Angel Barriga: Regressive Fuzzy Inference Models with Clustering Identification: Application to the ESTSP08 Competition
FWY	Francis Wyffels, Benjamin Schrauwen, Dirk Stroobandt: Using reservoir computing in a decomposition approach for time series prediction
GBO	Gianluca Bontempi: Long Term Time Series Prediction with Multi-Input Multi-Output Local Learning
GRU	Gines Rubio, Alberto Guillen, L.J. Herrera, Hector Pomares, Ignacio Rojas: Use of specific-to-problem kernel functions for time series modeling
IJA	Indir Jaganjac: Long-term prediction of nonlinear time series with recurrent least squares support vector machines
JJR	José B. Aragão Jr., Guilherme A. Barreto: Playout Delay Prediction in VoIP Applications: Linear versus Nonlinear Time Series Models
JJU	José Maria P. Júnior, Guilherme A. Barreto: Multistep-Ahead Prediction of Rainfall Precipitation Using the NARX Network
LSO	Luís Gustavo M. Souza, Guilherme A. Barreto: Multiple Local ARX Modeling for System Identification Using the Self-Organizing Map
MES	Marcelo Espinoza, Tillmann Falck, Johan A. K. Suykens, Bart De Moor: Time Series Prediction using LS-SVMs
MKA	M. Kanevski, V. Timonin, A. Pozdnoukhov, M. Maignan: Evolution of Interest Rate Curve: Empirical Analysis of Patterns Using Nonlinear Clustering Tools
MOL	Madalina Olteanu: Revisiting linear and non-linear methodologies for time series prediction - application to ESTSP08 competition data
MSU	Mika Sulkava, Harri Mäkinen, Pekka Höjd, Jaakko Hollmén: Automatic detection of onset and cessation of tree stem radius increase using dendrometer data and CUSUM charts
NKO	Nikolaos Kourentzes, Sven F. Crone: Automatic modelling of neural networks for time series prediction - in search of a uniform methodology across varying time frequencies
PAD	Paulo J. L. Adeodato, Adrian L. Arnaud, Germano C. Vasconcelos, Rodrigo C.L.V. Cunha, Domingos S.M.P. Monteiro: Exogenous Data and Ensembles of MLPs for Solving the ESTSP Forecast Competition Tasks
PJA	Philippe du Jardin: Bankruptcy prediction and neural networks: the contribution of variable selection methods
PPT	Piotr Ptak, Matylda Jabłońska, Dominique Habimana, Tuomo Kauranne: Reliability of ARMA and GARCH models of electricity spot market prices
QYU	Qi Yu, Antti Sorjamaa, Yoan Miche, Eric Séverin: A methodology for time series prediction in Finance
RNY	Roar Nybo: Time series opportunities in the petroleum industry
SAB	Syed Rahat Abbas, Muhammad Arif: Hybrid Criteria for Nearest Neighbor Selection with Avoidance of Biasing for Long Term Time Series Prediction
TPI	Tapio Pitkäranta: Kernel Based Imputation of Coded Data Sets
VON	Victor Onclinx, Michel Verleysen, Vincent Wertz: Projection of time series with periodicity on a sphere