

Andrzej Sokołowski

Akademia Ekonomiczna w Krakowie

Krzysztof Jajuga

Akademia Ekonomiczna we Wrocławiu

Eugeniusz Gatnar

Akademia Ekonomiczna w Katowicach

IFCS-2006 LUBLANA – KILKA REFLEKSJI

1. IFCS – kilka faktów

International Federation of Classification Societies (IFCS) to międzynarodowe towarzystwo naukowe, będące federacją narodowych, regionalnych bądź „lingwistycznych” towarzystw naukowych. Celem IFCS jest działalność naukowa zorientowana na teorię, algorytmy i zastosowania w zakresie analizy danych i klasyfikacji. Towarzystwo powstało w roku 1985 w Cambridge (Wielka Brytania). Informacje o federacji zawarte są na stronie internetowej: www.classification-society.org.

Założycielami IFCS były towarzystwa naukowe obejmujące następujące obszary geograficzne: CSNA – Północna Ameryka, BCS – Wielka Brytania, GfKI – Niemcy, Austria, Szwajcaria niemieckojęzyczna, JCS – Japonia, SFC – Francja, Walonia, Szwajcaria francuskojęzyczna, SIS – Włochy. Z czasem do IFCS dołączyły następujące towarzystwa naukowe: VOC z Holandii i Flandrii, SKAD z Polski, CLAD z Portugalii, KCS z Korei Płd., IPRCS z Irlandii (również Północnej), SoCCCAD z Ameryki Środkowej i Karaibów.

Przez pewien czas członkiem IFCS było również towarzystwo jugosłowiańskie (JSK), które przestało funkcjonować w połowie lat dziewięćdziesiątych. Obecnie w skład IFCS wchodzi 12 towarzystw naukowych, a zainteresowanie wejściem do IFCS wyrażają towarzystwa naukowe reprezentujące takie kraje, jak: Słowenia, Grecja, Hiszpania, Republika Południowej Afryki oraz Chiny.

Pracami IFCS kieruje rada (IFCS Council) oraz zarząd (IFCS Executive Committee), na czele którego stoi prezydent wybierany na dwuletnią kadencję. Godność

prezydenta piastowali, piastują i będą piastować znani profesorowie zajmujący się analizą danych, byli to w kolejności:

- 1986-1987 – Hans Bock (Niemcy),
- 1988-1989 – Robert Sokal (Stany Zjednoczone),
- 1990-1991 – John Gower (Wielka Brytania),
- 1992-1993 – William Day (Kanada),
- 1994-1995 – Allan Gordon (Wielka Brytania),
- 1996-1997 – Doug Carroll (Stany Zjednoczone),
- 1998-1999 – Chikio Hayashi (Japonia),
- 2000-2001 – Jean Paul Rasson (Belgia),
- 2002-2003 – Carlo Lauro (Włochy),
- 2004-2005 – Henk Kiers (Holandia),
- 2006-2007 – David Hand (Wielka Brytania),
- 2008-2009 – Frank McMorris (Stany Zjednoczone).

Najważniejszym wydarzeniem w działalności IFCS jest organizowana zazwyczaj raz na dwa lata konferencja, gromadząca wszystkich czołowych przedstawicieli dziedziny analizy danych. Do tej pory odbyło się 10 konferencji: w 1987 r. – w Akwizgranie, w 1989 r. – w Charlottesville, w 1991 r. – w Edynburgu, w 1993 r. – w Paryżu, w 1996 r. – w Kobe, w 1998 r. – w Rzymie, w 2000 r. – w Namur, w 2002 r. – w Krakowie, w 2004 r. – w Chicago, w 2006 r. – w Lublanie.

Następna konferencja odbędzie się w roku 2009 w Dreźnie.

2. IFCS-2006 – Lublana

Dziesiąta konferencja IFCS miała miejsce w Lublanie w dniach 25-29 lipca 2006 r. Instytucją organizującą był Wydział Nauk Społecznych (Faculty of Social Sciences) Uniwersytetu w Lublanie. W konferencji wzięło udział około 200 osób z blisko 25 krajów. Z Polski w konferencji wzięło udział 5 osób, które przedstawiły 4 referaty.

Jeśli chodzi o ogólną liczbę referatów w poszczególnych kategoriach, to jest ona następująca: 1 referat plenarny, tzw. *keynote*, 1 referat wygłoszony na zaproszenie prezydenta IFCS, 1 referat wygłoszony przez prezydenta IFCS, 8 referatów gości zaproszonych, 52 referaty przedstwione w tzw. zaproszonych sesjach (15 sesji) oraz 77 referatów zgłoszonych przez uczestników (20 sesji).

Struktura geograficzna autorów referatów jest następująca: Francja – 18, Niemcy, Słowenia – 16, Włochy, Japonia – 12, Holandia – 11, Stany Zjednoczone – 10, Belgia – 8, Wielka Brytania, Portugalia – 7, Kanada – 6, Polska – 4, Hiszpania – 3, Brazylia, Austria, Rosja – 2, Szwajcaria, Korea, Irlandia, Kostaryka, Portoryko, Izrael, RPA – 1. Oprócz tego odbyły się dwa tzw. krótkie kursy, mające charakter tutorski.

W sumie w trakcie konferencji miało miejsce 15 sesji zaproszonych; oto ich tematy podane w oryginalnym brzmieniu: Design and Choice of Similarities and Dissimilarities, Clustering and Classification of Microarray Gene Expression Data

(2 sesje), Classification and Data Analysis in Economics, Musicology, Exploratory and Confirmatory Analysis of Interval Data, Classification of Complex Data, Dissimilarity Analysis and Clustering, Latent Class Analysis and Classification Issues, Blockmodeling, Multidimensional Scaling of Asymmetric Relationships Data (2 sesje), Symbolic Data Analysis, Spatial Classification, Web Mining.

Z kolei referaty zgłoszone przez uczestników zostały zgrupowane w 19 sesjach. Ich tematyka – podana w oryginalnym brzmieniu – była następująca: Cluster Analysis, Multidimensional Scaling, Network Analysis, Optimization, Data Analysis in Medicine, Graphs, Similarities for Mixed Variables, Clustering Temporal Data, Two-Mode Clustering, Data Analysis in Biology, Multivariate Analysis, Application in Social Sciences, Dissimilarities, Non-hierarchical Clustering, Statistics (2 sesje), Text Analysis, Clustering, Principal Components Analysis.

Miała miejsce również 1 sesja plakatowa.

Pewna część referatów przedstawionych na konferencji została opublikowana w książce pod redakcją Batagelja, Bocka, Ferligoj i Ziberny (2006 r.).

3. Analiza tytułów referatów wygłoszonych na konferencji IFCS-2006

W zbiorze streszczeń wydanych na konferencję IFCS-2006 zamieszczono krótkie opisy referatów. Przygotowano plik danych zawierających te tytuły. Jego fragmenty zawiera tab. 1.

Tabela 1. Początek pliku danych z tytułami referatów wygłoszonych na konferencji IFCS-2006

Dane: Ifcs2006 (1 zmn. * 142 prz.)	
	1 Tytuł
1	Mixed Principal Component Analysis
2	Improving the Performance of Principal Components for Classification of Gene Expression Data Through Feature Selection
3	Nonlinear Principal Component Analysis for Representing Inter-Variable Relationships by Trajectories
4	Double Hierarchical Mixture Model for Microarray Data
5	Survival Combination Scores: A Case of a Revolving Store Card
6	Hierarchical Clustering for Boxplot Variables
7	Model Selection in Latent Class Metric Conjoint Analysis: Sequential and Simultaneous Approach
8	Comparing of Clustering Approaches for Mixed Variables
9	Partition-Valued Random Variables
10	Phylogenetic Trees, Additive Tree Metrics and Filiation of Manuscripts: The Case of the Benares Glose
11	Analyzing the Structure of U.S. Patent Network
12	Three Dimensional Blockmodeling
13	Combining Data and Network Analysis
14	Revealing Hidden Relevance in Complex Laddering Networks
15	Spectral Clustering and Multidimensional Scaling: A Unified View
16	Evaluation Of Allocation Rules under Some Coast Constraints
17	Bayesian Models for Safety Design to Prevent Foreign in Children Body Injuries
18	Paired-Hierarchies and Associated Dissimilarities
19	Simultaneous Model-Based Clustering of Data Arising from Different Populations
20	Dependence in Interval-Valued Observations
21	A Tree-Based Similarity for Evaluating Concept Proximities in an Ontology
22	Class Prototypes for Complex (Interval) Data
23	Crisp Partitions Induced by a Fuzzy Set
24	Asymmetric Multidimensional Scaling with External Information: Some Possible Approaches
25	A Novel Mixture-Model Cluster Analysis with Genetic EM Algorithm and Information Complexity as the Fitness Function

Źródło: opracowanie własne.

Analizę przeprowadzono przy wykorzystaniu modułu *Text Mining & Document Retrieval* z programu *STATISTICA* (ver. 7). Program wyszukuje rdzenie wyrazów powtarzających się najwięcej razy (najczęściej występujących w tytułach), a następnie sporządza macierz binarną o liczbie wierszy równej liczbie referatów. Kolumny odpowiadają kolejnym rdzeniom wyrazów. Oczywiście, jedynka oznacza, że w danym tytule występuje konkretny wyraz.

Efekty wyszukiwania przedstawione są w tab. 2.

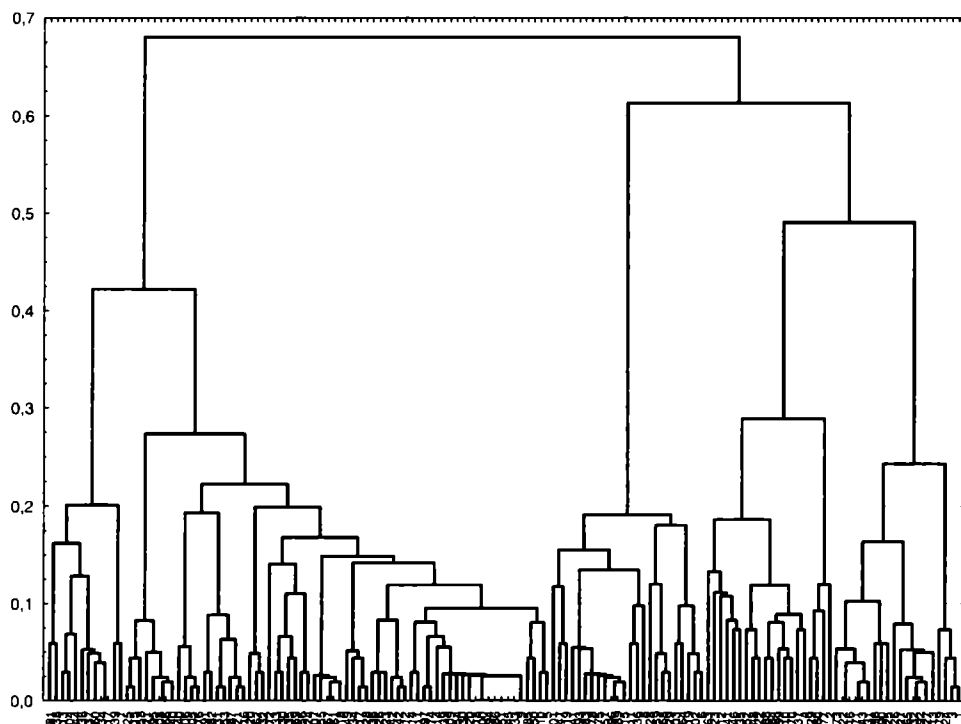
Tabela 2. Słowa najczęściej występujące w tytułach referatów

Słowo	Liczba referatów	Słowo	Liczba referatów	Słowo	Liczba referatów
<i>Cluster</i>	37	<i>compared</i>	5	<i>two-mode</i>	4
<i>Data</i>	36	<i>gene</i>	5	<i>biclustering</i>	3
<i>Analysis</i>	24	<i>group</i>	5	<i>binary</i>	3
<i>Model</i>	14	<i>mixed</i>	5	<i>blockmodeling</i>	3
<i>Classification</i>	13	<i>network</i>	5	<i>case</i>	3
<i>Method</i>	13	<i>optimal</i>	5	<i>combining</i>	3
<i>Approach</i>	12	<i>principal</i>	5	<i>complex</i>	3
<i>Used</i>	9	<i>simultaneous</i>	5	<i>dependence</i>	3
<i>Class</i>	8	<i>assessing</i>	4	<i>detection</i>	3
<i>Distance</i>	7	<i>associated</i>	4	<i>efficient</i>	3
<i>Hierarchical</i>	7	<i>different</i>	4	<i>empirical</i>	3
<i>New</i>	7	<i>expression</i>	4	<i>latent</i>	3
<i>Variable</i>	7	<i>framework</i>	4	<i>mining</i>	3
<i>Algorithm</i>	6	<i>local</i>	4	<i>mixture</i>	3
<i>Comparison</i>	6	<i>measures</i>	4	<i>model-based</i>	3
<i>Component</i>	6	<i>partition</i>	4	<i>perspective</i>	3
<i>Dissimilarity</i>	6	<i>scaling</i>	4	<i>representation</i>	3
<i>Evaluation</i>	6	<i>selection</i>	4	<i>research</i>	3
<i>Microarray</i>	6	<i>set</i>	4	<i>science</i>	3
<i>Similarity</i>	6	<i>structure</i>	4	<i>series</i>	3
<i>Study</i>	6	<i>symbolic</i>	4	<i>social</i>	3
<i>Application</i>	5	<i>time</i>	4	<i>spatial</i>	3
<i>Asymmetric</i>	5	<i>tree</i>	4		

Źródło: opracowanie własne.

Referaty podzielono na grupy na podstawie słów występujących w tytułach. Wykorzystano metodę Warda, a w niej niezgodność procentową jako miarę odległości. Otrzymano dendrogram przedstawiony na rys. 1.

Obiektami w tej analizie są referaty. Odległością pomiędzy dwoma referatami jest procent wyrazów (z listy podanej w tab. 2), które występują tylko w jednym z porównywanych referatów. Ze względu na relatywnie dużą liczbę analizowanych referatów na dendrogramie przedstawionym w pracy można tylko ocenić strukturę zbioru, a niemożliwe jest zidentyfikowanie nazw obiektów opisanych przy osi poziomej.



Rys. 1. Dendrogram Warda
Źródło: opracowanie własne.

Oczywisty wydaje się podział referatów na pięć grup. Zdecydowaliśmy się nazwać je następująco (jak zwykle w wynikach analizy skupień kolejność grup nie ma znaczenia; tutaj jest taka, jak pokazano na wykresie):

- grupa 1 – „Modelowanie” – 12 referatów (8,5%),
- grupa 2 – „Różności” – 66 referatów (46,4%),
- grupa 3 – „Grupowanie” – 24 referaty (16,9%),
- grupa 4 – „Zastosowania” – 20 referatów (14,1%),
- grupa 5 – „Analiza” – 20 referatów (14,1%).

W wyjaśnieniu takich nazw grup mogą pomóc charakterystyki grup pokazane w tab. 3.

Prezentujemy również ciekawe zestawienie najpopularniejszych wyrazów w tytułach (tab. 4).

Grupa pierwsza referatów to prace poświęcone modelowaniu. Podobnie wyraźne wyróżnienie poprzez najpopularniejszy wyraz dotyczy grup trzeciej i czwartej. Jest zrozumiałe, że na konferencji IFCS mamy do czynienia z grupami referatów poświęconych grupowaniu, zastosowaniom, modelowaniu i analizom. Te zagadnienia to prawie połowa wszystkich referatów. Na konferencjach IFCS zawsze jednak widać różnorodność problematyki (w tym roku przykładem może być sekcja

Tabela 3. Charakterystyki grup referatów

Wyszczególnienie	Grupa 1	Grupa 2	Grupa 3	Grupa 4	Grupa 5
Wszystkie tytuły zawierają ten wyraz	<i>model</i>		<i>cluster</i>	<i>data (19/20)</i>	<i>analysis</i>
Ten wyraz jest tylko w tytułach tej grupy		<i>blockmodeling</i> <i>case</i> <i>evaluation</i> <i>group</i> <i>optimal</i> <i>research</i> <i>science</i> <i>series</i> <i>set</i>	<i>model-based</i>	<i>biclustering</i> <i>expression</i>	
Co najmniej 2/3 wystąpień tego wyrazu jest w tej grupie	<i>binary</i> <i>efficient</i> <i>latent</i> <i>model</i>	<i>combining</i> <i>detection</i> <i>different</i> <i>dissimilarity</i> <i>distance</i> <i>measures</i> <i>mining</i> <i>partition</i> <i>perspective</i> <i>scaling</i> <i>similarity</i> <i>social</i> <i>spatial</i> <i>time</i> <i>tree</i>	<i>cluster</i> <i>hierarchical</i> <i>symbolic</i>	<i>gene</i> <i>microarray</i> <i>mixture</i>	<i>component</i> <i>principal</i>

Źródło: opracowanie własne.

Tabela 4. Najpopularniejsze wyrazy w grupach (w nawiasach podano procent referatów w grupie zawierających dany wyraz)

Grupa 1	Grupa 2	Grupa 3	Grupa 4	Grupa 5
<i>Model</i> (100) <i>Class</i> (17) <i>New</i> (17)	<i>classification</i> (12) <i>approach</i> (11)	<i>cluster</i> (100) <i>method</i> (25) <i>hierarchical</i> (25) <i>new</i> (17) <i>approach</i> (13) <i>used</i> (13) <i>variable</i> (13) <i>distance</i> (8)	<i>data</i> (95) <i>method</i> (20) <i>class</i> (20) <i>used</i> (15)	<i>analysis</i> (100) <i>classification</i> (15)

Źródło: opracowanie własne.

poświęconą zastosowaniu metod analizy danych w muzykologii). Ta różnorodność znajduje swój wyraz w referatach grupy drugiej i to też świadczy o sile konferencji niezamykającej się w wąskiej problematyce badawczej.

Literatura

Batagelj V., Bock H.-H., Ferligoj A., Ziberna A. (red.) (2006), *Data Science and Clasification*, Springer, Berlin.

IFCS-2006 LUBLANA – SOME REMARKS

Summary

In the paper we present some facts about International Federation of Classification Societies and the remarks on the papers presented in IFCS conference which took place in Lublana.