Semi-Supervised Clustering Algorithms for Grouping Scientific Articles

Diego Vallejo, Paulina Morillo, Cèsar Ferri
dvallejoh@ups.edu.ec, pmorillo@ups.edu.ec, cferri@dsic.upv.es

UNIVERSITAT POLITÈCNICA DE VALÈNCIA

Universidad Politécnica SALESIANA Ecuador

1.

Introduction

Grouping documents with size constraints

Introduction

- Scientific conferences are organised by sessions formed by papers with similar topics
- Schedule is configured by some size constraints
- This work presents some methods to automatise the configuration of sessions in scientific conferences

Title

Semi-Supervised Clustering Algorithms for Grouping Scientific Articles

Diego Vallejo-Huanga[1], Paulina Morillo[2], and Cèsar Ferri[3]

[1] Universidad Politécnica Salesiana, Department of Computer Science, Quito, Ecuador
dvallejoh@ups.edu.ec
[2] Universidad Politécnica Salesiana, Research Group IDEIAGEOCA, Quito, Ecuador
pmorillo@ups.edu.ec
[3] Universitat Politècnica de València, DSIC, València, Spain
cferri@dsic.upv.es

Abstract

Abstract

Creating sessions in scientific conferences consists in grouping papers with common topics taking into account the size restrictions imposed by the conference schedule. Therefore, this problem can be considered as semi-supervised clustering of documents based on their content. This paper aims to propose modifications in traditional clustering algorithms to incorporate size constraints in each cluster. Specifically, two new algorithms are proposed to semi-supervised clustering, based on: binary integer linear programming with cannot-link constraints and a variation of the K-Medoids algorithm, respectively. The applicability of the proposed semi-supervised clustering methods is illustrated by addressing the problem of automatic configuration of conference schedules by clustering articles by similarity. We include experiments, applying the new techniques, over real conferences datasets: ICMLA-2014, AAAI-2013 and AAAI-2014. The results of these experiments show that the new methods are able to solve practical and real problems.

Keywords

Keywords: Clustering with constraints, Size constraint, K-Medoids, Linear programming

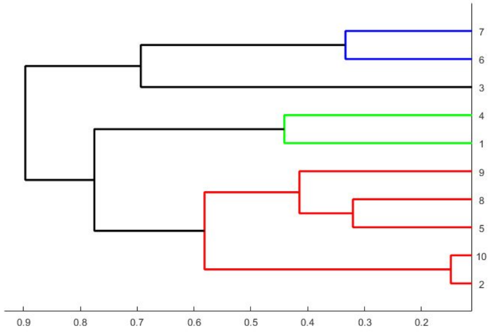
1 Introduction

Clustering

- Task of grouping a set of objects in such a way that objects in the same cluster are more similar to each other than to those in other groups
    - Exploratory data mining

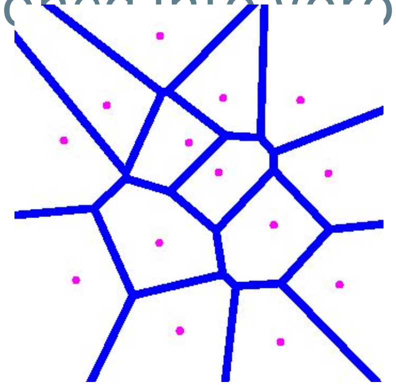
Hierarchical Clustering

- Clusters are built forming a hierarchy
  - **Agglomerative**: "bottom up" approach: each element forms a cluster, pairs of clusters are merged until creating just a single cluster
  - **Divisive**: "top down" approach: all observations start in one cluster, which is split recursively until having an element per group



Centroid-based clustering

- Divide *n* elements into *k* clusters
  - Each element belongs to the cluster with the nearest mean, serving as a prototype of the cluster.
  - Computationally difficult (NP-hard); there are efficient heuristic algorithms that can converge quickly to a local optimum.
  - Space is partitioned into Voronoi cells.



K-means

- *K* points are placed randomly in the space (centroids)
  a. Each object is assigned to the closer centroid
  b. After assigning all the objects, centroids are recalculated
  c. Steps *a* and *b* are repeated until centroids keep stable
- The performance of K-means depends on the initialization method

Document Classification

- Division of a document collection into groups according to their content
    - information retrieval, topic detection and content tracking
- Natural Language Processing (NLP) is used to characterize the documents
    - Stemming, Stopwords
    - Bag of words
    - TF/IDF
    - Distances

K-mediods

- A clustering algorithm based on mediods
  - object of a cluster whose average dissimilarity to all the objects in the cluster is minimal
- Algorithm
  - Select *k* of the *n* data points as the medoids
  - While the cost of the configuration decreases:
    - Reassign each point to the cluster defined by the closest medoid
    - In each cluster, recompute the mediod

2.

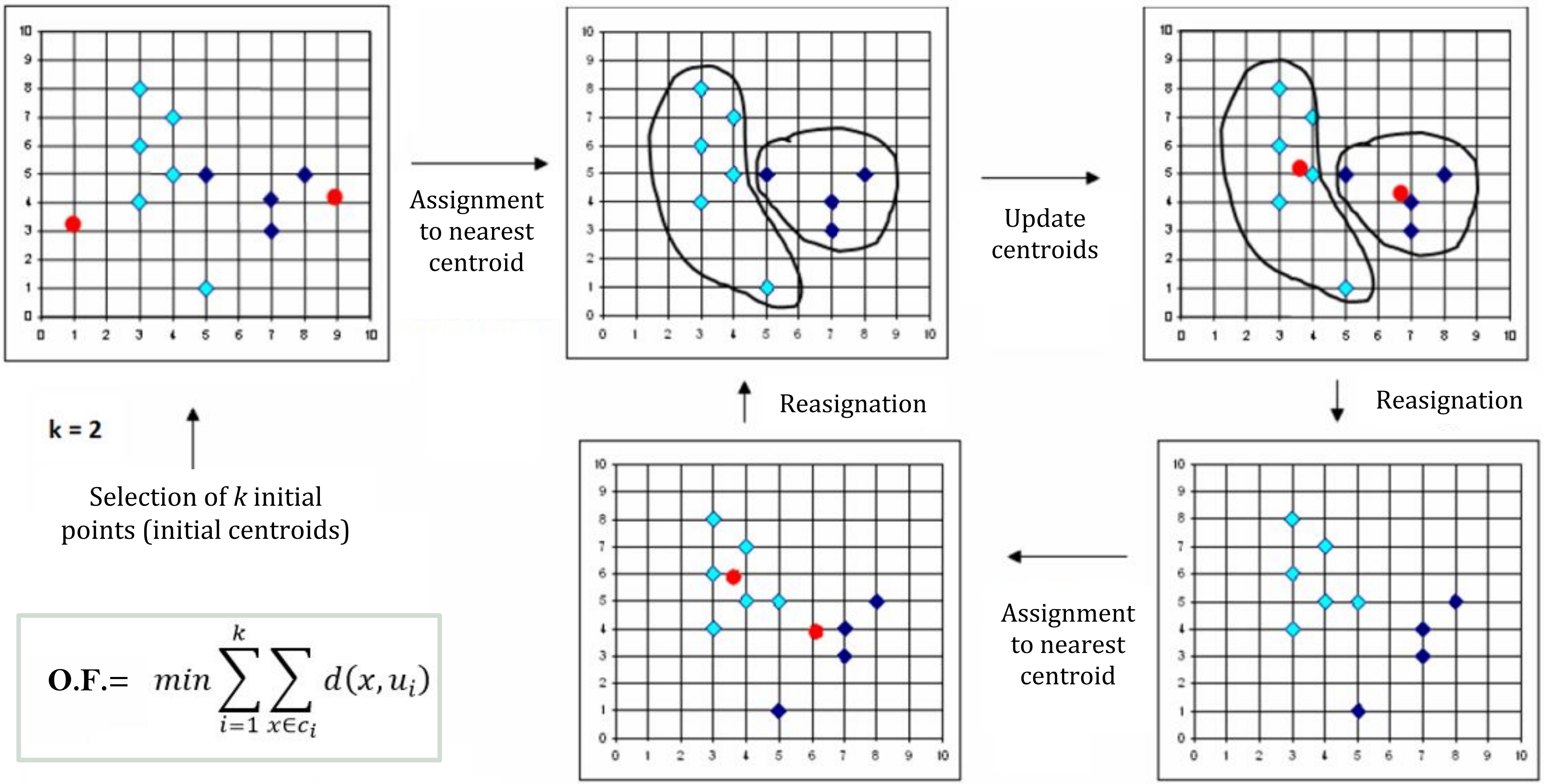
Clustering Algorithms With Size Constraints

In many cases the data or problems, have certain implicit restrictions, which traditional clustering algorithms do not take into account

Clustering with constraints

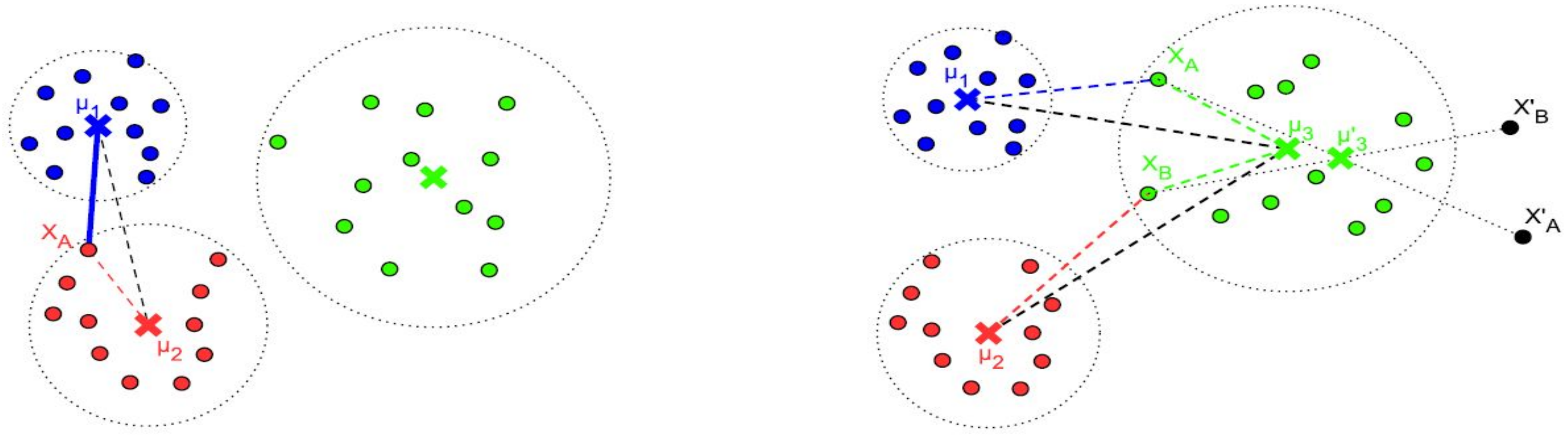
- Class of semi-supervised learning algorithms
  - must-link constraints, cannot-link constraints
  - Works are based on classical partition algorithms for the incorporation of size constraints
- Related to Maximally Diverse Grouping Problem
  - grouping a set of M elements into G mutually disjoint groups in such a way that the diversity among the elements in each group is maximized

K-MedoidsSC



$$\textbf{O.F.=}\quad min\sum_{i=1}^{k}\sum_{x\in c_i} d(x,u_i)$$

K-MedoidsSC (size contraints)



(a) NO points (b) FI and VFI points

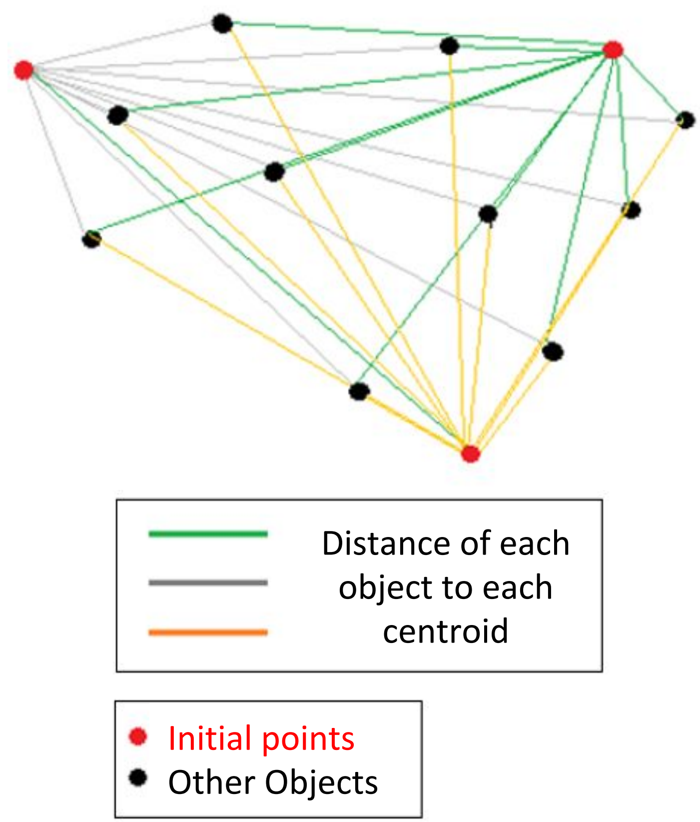
$$J_{KMS} = J_{KM} + \alpha J_A + \beta J_S + \gamma J_L$$

$$J_{KM} = min \sum_{i=1}^{k} \sum_{x \in c_i} d(x, u_i)$$

Penalise the size of the clusters:

- ❑ When cluster size is smaller than expected
- ❑ When the cluster size is larger than expected
- ❑ When the desired cluster size is not achieved

CSCLP



$$A = \begin{array}{c c} & \begin{array}{ccccc} c_1 & c_2 & \dots & \dots & c_k \end{array} \\ \begin{array}{c} D_1 \\ D_2 \\ \cdot \\ \vdots \\ \\ D_n \end{array} & \begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 & 0 \end{pmatrix}_{nxk} \end{array}$$

$$\boxed{\textbf{O.F.} = min\left(\sum_{j=1}^{k}\sum_{i=1}^{n} d_{ij}p_{ij}\right)}$$

Constraints:

Belonged to a single group $\rightarrow$ $$\sum_{j=1}^{k} a_{ij} = 1 \ ; \ \forall i = 1, 2, \dots, n$$

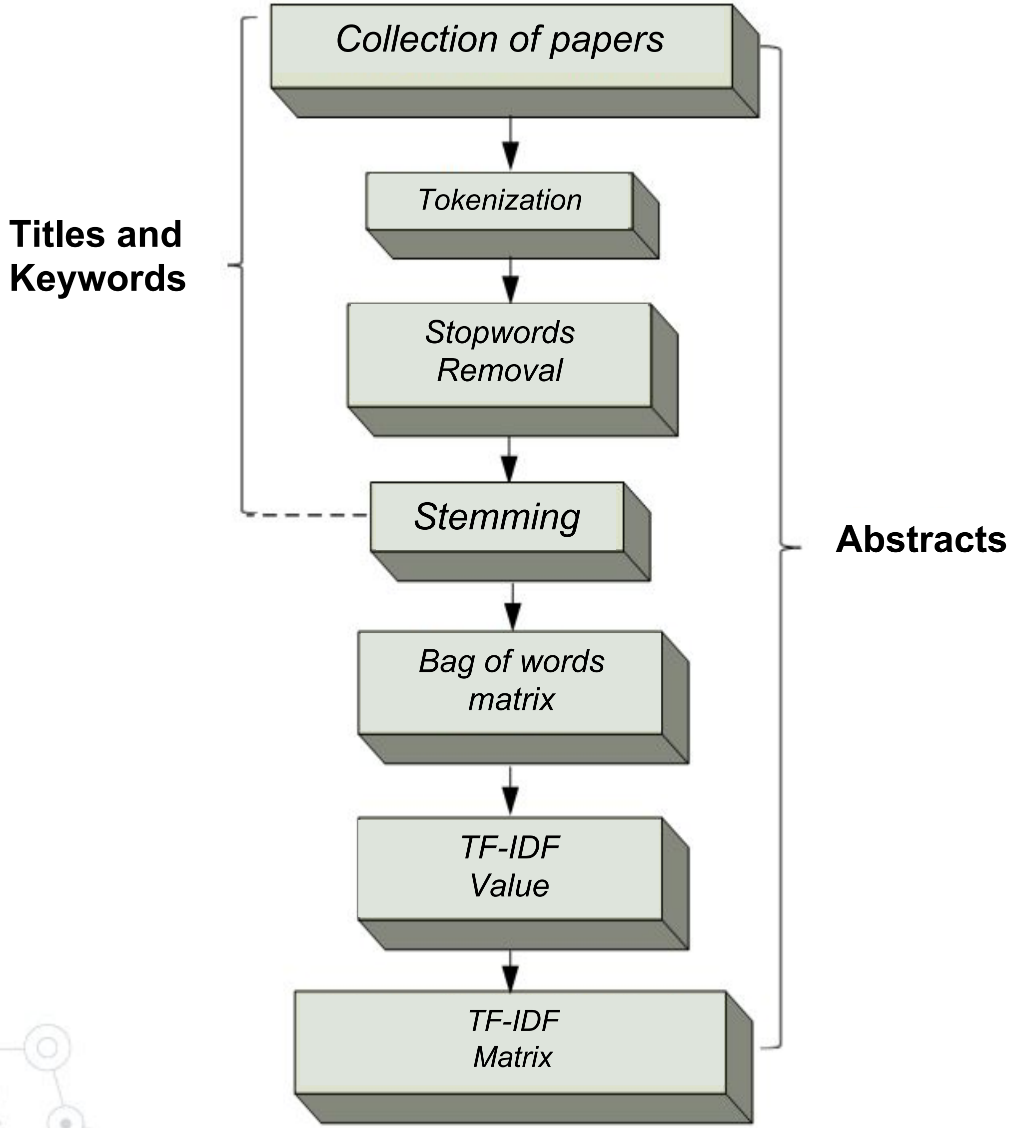
Size $\rightarrow$ $$\sum_{i=1}^{n} a_{ij} = e_j \ ; \ \forall j = 1, 2, \dots, k$$

3.

Methodology

Document clustering has been applied to many fields of study, such as: information retrieval, topic detection and content tracking, all of them are intrinsically related to language

NLP Techniques



Clustering process

Titles Dissimilarity Matrix
(Jaccard coefficient)

+

Keywords Dissimilarity Matrix
(Jaccard coefficient)

+

Abstracts Dissimilarity Matrix
(Cosine distance)

=

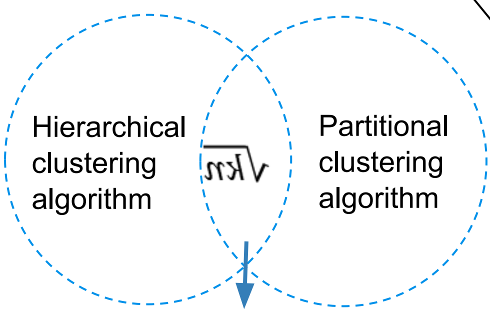
Dissimilarity Matrix

$$S'_{TOT} = \theta S'_{T} + \pi S'_{K} + \omega S'_{A}$$

$$\theta + \pi + \omega = 1\,;\{\forall\, \theta, \pi, \omega\} \in [0,1]$$

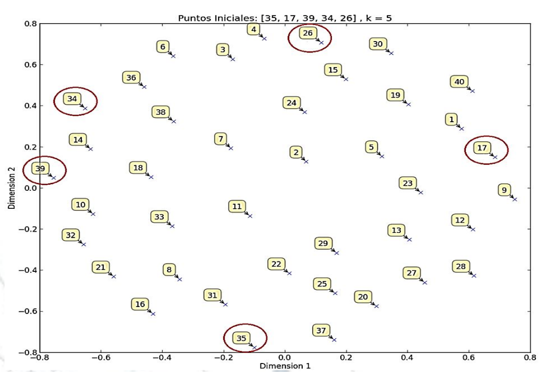
Selection of initial points

Buckshot Technique



Initial centroids

Farthest Neighbour Technique



Apply algorithms
(K-MedoidsSC or SCLP)

Validation

Internal
Compaction of groups

- Silhouette Index

External
Compare clustering with *"best clustering"*

- ARI: Adjusted Rand Index
- NMI: Normalized Mutual Information
- AMI: Adjusted Mutual Information

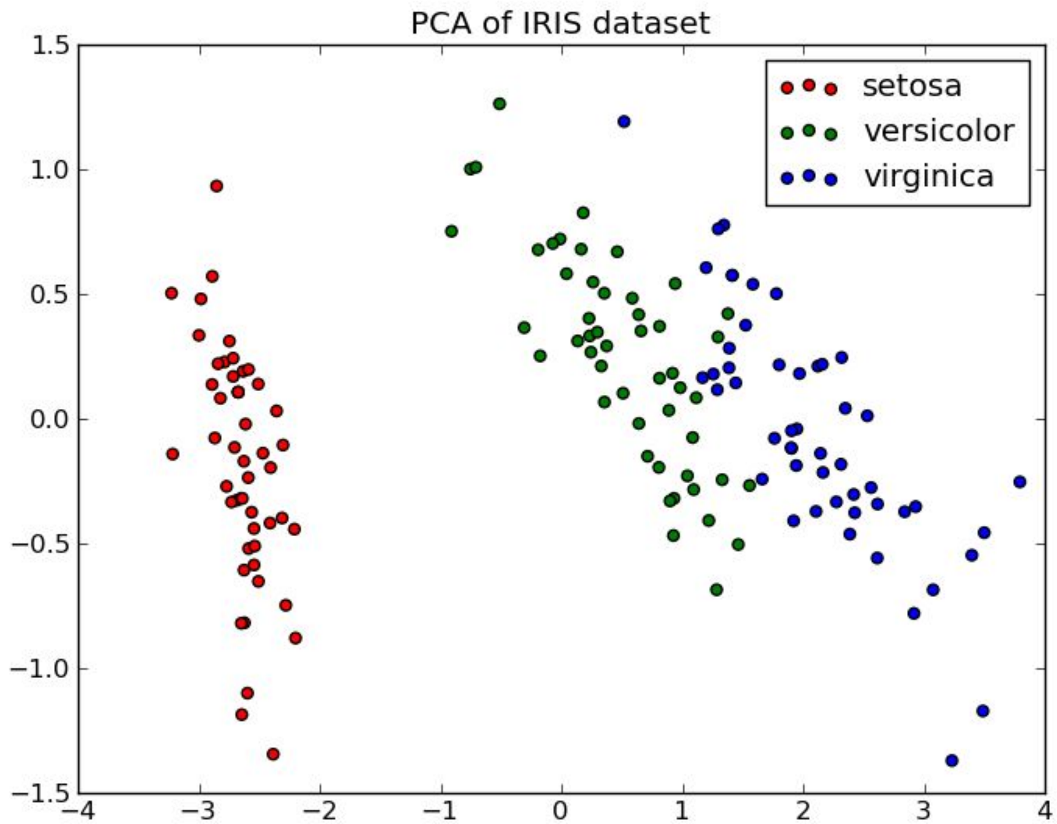
4.

Experiments

We use in the first instance multivariate benchmarking datasets and then we use a documentary dataset

Data

UCI Machine Learning Repository



- IRIS, Wine, seeds

Validation: Benchmarking datasets

Algorithm	Farthest Neighbour Technique			Buckshot Technique		
	Iris	Wine	Seeds	Iris	Wine	Seeds
AHC-FPA*	(50,29,71)	(10,32,136)	(8,42,160)	(50,29,71)	(10,32,136)	(8,42,160)
K-Medoids	(50,55,45)	(60,41,77)	(74,70,66)	(50,45,55)	(88,74,16)	(89,64,57)
CSCLP	(50,50,50)	(59,71,48)	(70,70,70)	(50,50,50)	(59,71,48)	(70,70,70)
K-MedoidsSC	(50,50,50)	(59,71,48)	(70,70,70)	(50,50,50)	(59,71,48)	(70,70,70)
Real Cluster Size	(50,50,50)	(59,71,48)	(70,70,70)	(50,50,50)	(59,71,48)	(70,70,70)
Initial Points IDs	[23,75,119]	[19,118,15]	[23,107,204]	[39,98,113]	[135,80,109]	[48,107,152]

* Does not use any technique to select initial points

Resulting cluster sizes in datasets: Iris, Wine and Seeds, with algorithms: AHC-FPA, K-Medoids, CSCLP and K-MedoidsSC

Validation: Benchmarking datasets

Datasets	Algorithm	Farthest Neighbour Technique				Buckshot Algorithm			
		ARI	AMI	NMI	$S(i)$	ARI	AMI	NMI	$S(i)$
Iris	AHC-FPA*	0.674	0.735	0.760	0.659	0.674	0.735	0.760	0.659
	K-Medoids	0.904	0.897	0.900	0.737	0.904	0.897	0.900	0.737
	CSCLP	0.886	0.861	0.862	0.721	0.886	0.861	0.862	0.733
	K-MedoidsSC	0.818	0.800	0.803	0.717	0.886	0.861	0.862	0.734
Wine	AHC-FPA*	0.059	0.137	0.186	0.728	0.059	0.137	0.186	0.728
	K-Medoids	0.347	0.363	0.373	0.758	0.208	0.196	0.221	0.757
	CSCLP	0.236	0.239	0.247	0.655	0.331	0.371	0.378	0.669
	K-MedoidsSC	0.302	0.297	0.304	0.716	0.347	0.374	0.380	0.699
Seeds	AHC-FPA*	0.223	0.286	0.379	0.659	0.223	0.286	0.379	0.659
	K-Medoids	0.264	0.305	0.330	0.606	0.264	0.305	0.330	0.606
	CSCLP	0.233	0.268	0.275	0.420	0.231	0.249	0.256	0.456
	K-MedoidsSC	0.149	0.179	0.186	0.348	0.162	0.189	0.196	0.276

* Does not use any technique to select initial points

Clustering results in datasets: Iris, Wine and Seeds, with algorithms: AHC-FPA, K-Medoids, CSCLP and K-MedoidsSC

Data

AAAI-13
July 14-18, 2013
IAAI-13
July 16-18, 2013
BELLEVUE, WASHINGTON, USA

UCI
Machine Learning Repository

- Data scraping techniques from its website
- Number of papers: 69, 149, 398

ICMLA 2014 Program Schedule
Marriott Detroit
Renaissance Center, Detroit, Michigan 48243

Day 1: Wednesday, December 3, 2014			
7:30AM - 8:30AM	Light Breakfast		
8:30AM - 8:45AM	Open Remark		
8:45AM -10:25AM	Room:Ambassador I Chair: Fares Hedayati	Room: Brule A Chair: Ali Bou Nassif	Room: Brule B Chair: Moamar Sqyed-mouchaweh
	Session: Information Retrieval I	**Special Session: Machine Learning for Predictive Models I**	**Session: Ensemble Methods**
10:25AM - 10:40AM	Coffee Break		
10:40AM - 12:20PM	Room:Ambassador I Chair: Kaushik Sinha	Room: Brule A Chair: Ali Bou Nassif	Room: Brule B Chair: Bo Luo
	Session: Information Retrieval II	**Speical Session: Machine Learning for Predictive Models II**	**Session: Feature Extraction and Selection**
12:20AM - 1:20AM	Lunch		
1:20PM - 3:00PM	Room:Ambassador I		
	Tutorials: Big Data Industry (Jayashree Ravi)		
3:00PM - 3:20PM	Coffee Break		
3:20PM - 5:00PM	Room:Ambassador I		
	Tutorials: Big Data Industry (Jayashree Ravi)		

Validation: Datasets about papers of scientific conferences

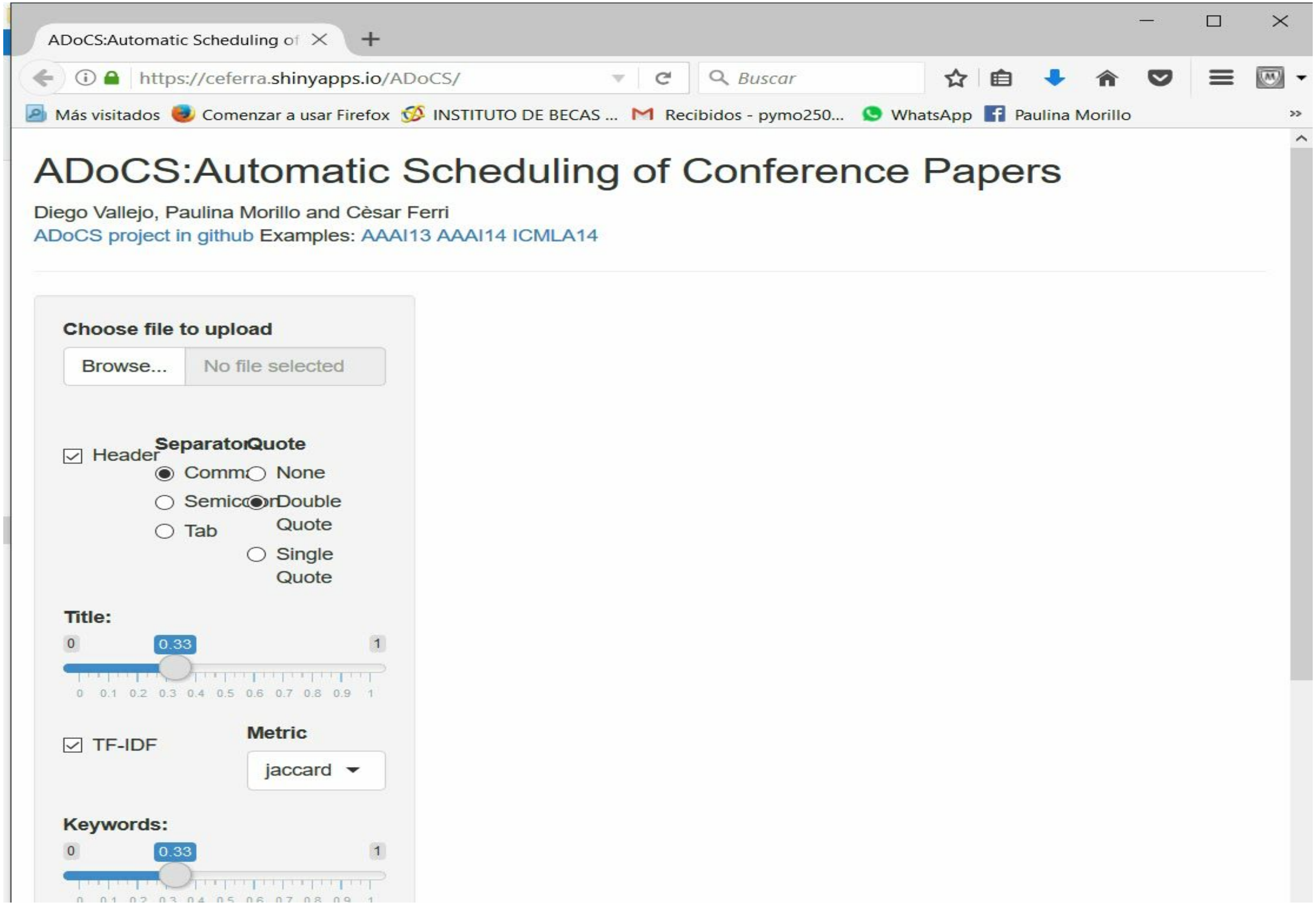
Datasets	Algorithm	Farthest Neighbour Technique k	Cluster Size c_j	ARI	AMI	NMI	$S(i)$
AAAI-13	CSCLP	3	(45,52,53)	0.029	0.030	0.042	0.028
	K-MedoidsSC	3	(45,52,53)	0.010	0.012	0.024	0.023
AAAI-14	CSCLP	11	(10,11,18,19,21,25,30,42,45,57,120)	0.079	0.128	0.185	0.040
	K-MedoidsSC	11	(10,11,18,19,21,25,30,42,45,57,120)	0.025	0.074	0.135	0.035
ICMLA-14	CSCLP	14	(4,5,5,5,5,5,5,5,5,5,5,5,5,5)	0.040	0.074	0.494	0.229
	K-MedoidsSC	14	(4,5,5,5,5,5,5,5,5,5,5,5,5,5)	0.063	0.107	0.512	0.230

Datasets	Algorithm	Buckshot Algorithm k	Cluster Size c_j	ARI	AMI	NMI	$S(i)$
AAAI-13	CSCLP	3	(45,52,53)	0.036	0.037	0.049	0.031
	K-MedoidsSC	3	(45,52,53)	0.018	0.016	0.028	0.030
AAAI-14	CSCLP	11	(10,11,18,19,21,25,30,42,45,57,120)	0.074	0.120	0.178	0.037
	K-MedoidsSC	11	(10,11,18,19,21,25,30,42,45,57,120)	0.048	0.085	0.145	0.036
ICMLA-14	CSCLP	14	(4,5,5,5,5,5,5,5,5,5,5,5,5,5)	0.094	0.146	0.533	0.234
	K-MedoidsSC	14	(4,5,5,5,5,5,5,5,5,5,5,5,5,5)	0.048	0.079	0.497	0.217

Clustering results in datasets: AAAI-13, AAAI-14 and ICMLA-14 with algorithms: CSCLP and K-MedoidsSC and two initial points methods (Farthest Neighbour and Buckshot)

ADoCS: Automatic Scheduling of Conference Papers

A web tool implemented in R


ADoCS:Automatic Scheduling of
https://ceferra.shinyapps.io/ADoCS/
Buscar
Más visitados
Comenzar a usar Firefox
INSTITUTO DE BECAS ...
Recibidos - pymo250...
WhatsApp
Paulina Morillo
ADoCS:Automatic Scheduling of Conference Papers
Diego Vallejo, Paulina Morillo and Cèsar Ferri
ADoCS project in github Examples: AAAI13 AAAI14 ICMLA14
Choose file to upload
Browse...
No file selected
Header
Separator
Quote
Comma
Semicolon
Tab
None
Double Quote
Single Quote
Title:
0
0.33
1
0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1
TF-IDF
Metric
jaccard
Keywords:
0
0.33
1

5.

Conclusions and Future work

Two novel algorithms for semi-supervised clustering are presented, that allow constraints on the sizes of the clusters

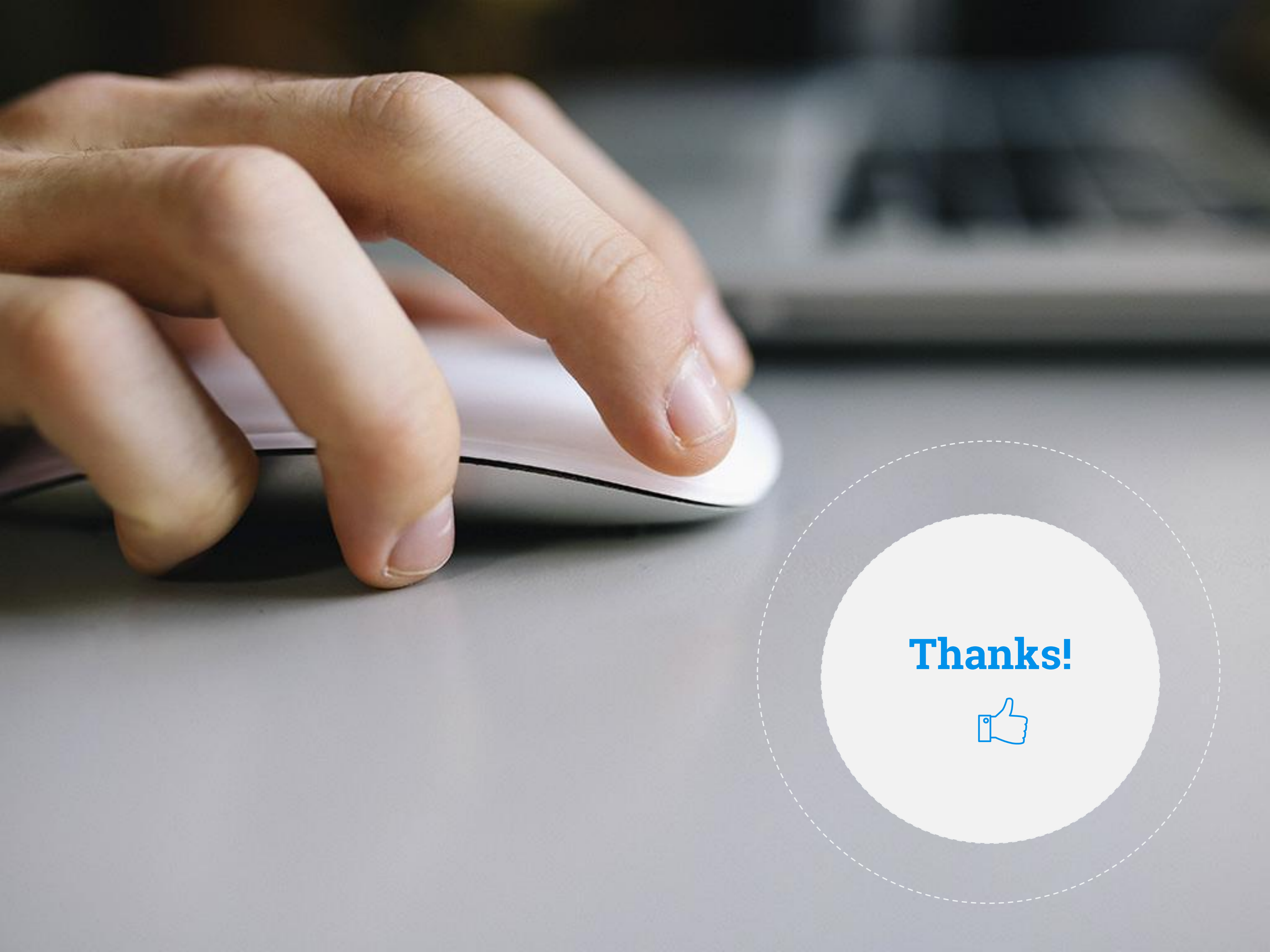
K-MedoidsSC

- Uses functions to penalize breach of desired group size
- It is based on the K-medoid algorithm

CSCLP

- Uses cannot-link constraints
- Uses linear binary integer programming

- New algorithms can solve clustering problems with size constraints.
- Automatic arranging of papers to create an appropriate conference schedule.
- Future work:
  - Developing conceptual clustering methods to find topics to label the created clusters


Thanks!