Faris Alqadah, Raj Bhatnagar
University of Cincinnati

Clustering Multi-Domain Information Networks

1 Information networks are ever-present in modern day applications such as: **bioinformatics**, **social network analysis**, and **recommender systems**. Such networks comprise multiple, interrelated datasets involving several domains.

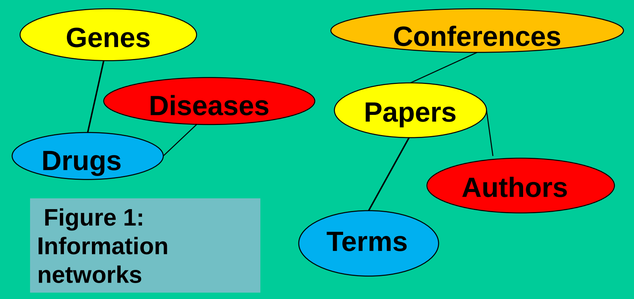

Figure 1: Information networks

2 Most work has focused on real-valued data and hard clusters. Our work focuses on **arbitrarily positioned, overlapping multi-way clusters in binary relations.** Formal Concept Analysis (FCA) is utilized as a theoretical basis for defining and enumerating such clusters.

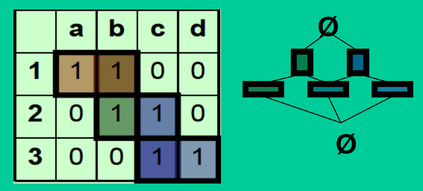

Figure 2: Binary relation with bi-clusters and concept lattice

3 Extended FCA to information networks by defining an information network cluster as matching (or almost) matching concepts across all datasets in the network.

-Orders of magnitude more efficient than single dataset concept enumeration

-Precise clusters but suffer in terms of recall

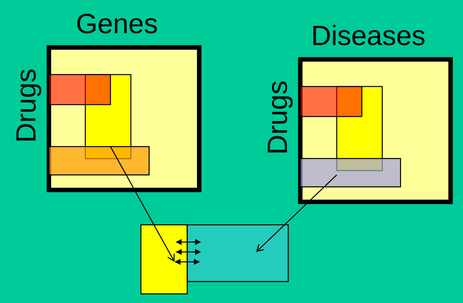

Figure 3: Clusters as matching concepts

4 Relax strict FCA criterion to improve clustering quality in terms of recall. A subspace (G1,…,Gn) is a cluster if the number of 1s encompassed is greater than the expected number of 1s (connectivity requirement).

-Key Fact: Matching concepts always satisfy connectivity requirement

-Key Fact: Neighboring concepts always minimize the number 0s introduced to a subspace.

5 Algorithm

1. Enumerate matching concepts across network

2. Navigate each local concept lattice augmenting clusters until no more augmentation possible

3. Merge all the local clusters and check if connectivity requirement is met.

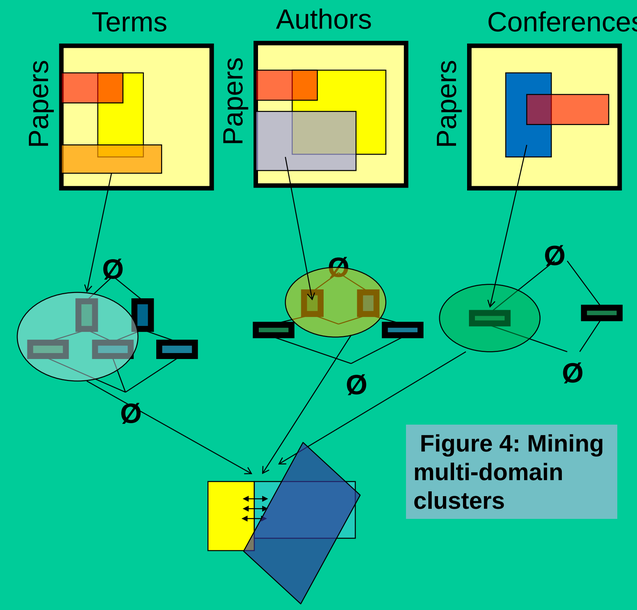

Figure 4: Mining multi-domain clusters

Algorithm	F_1	$F_{0.5}$	F_2
TreeClu	**0.561466**	**0.564308**	**0.558652**
NetClus	0.36124	0.366555	0.356076
MDC	0.508535	0.516275	0.501023
TreeClu	**0.567559**	**0.585322**	**0.550844**
NetClus	0.317618	0.356658	0.286281
MDC	0.416542	0.5222	0.346445
TreeClu	**0.489983**	**0.648295**	**0.393814**
NetClus	0.239803	0.306856	0.1968
MDC	0.364935	0.543009	0.274813
TreeClu	**0.489983**	**0.648295**	**0.393814**
NetClus	0.206759	0.377445	0.142375
MDC	0.233105	0.425454	0.160529

(f) FourAreas

Figure 5: Clustering results

6 Multi-domain clustering still in its infancy. Algorithmic methodologies are still early in development and offer great potential research