

FLINT 2001
2001 BISC International Workshop on Fuzzy Logic and the Internet

UC Berkeley, 14-17 August 2001

Dynamic Knowledge Representation for E-Learning Applications

Maria Eduarda Mendes, Lionel Sacks

Department of Electronic and Electrical Engineering
University College London, UK



Supported by **FCT** Fundação para a Ciência e a Tecnologia
MINISTÉRIO DA CIÊNCIA E DA TECNOLOGIA



Agenda

- Internet-based teaching and learning – the CANDLE project
- Adaptive courseware retrieval and navigation
- Knowledge representation and fuzzy clustering
- Metric concepts for document clustering
- Preliminary trials – description and results
- Concluding remarks



Internet-based Teaching and Learning

- Distributed and asynchronous access to learning facilities
- Flexible learning environment:
 - Study program adapted to the learner's background and interests
 - Types of interactions between students (group work, isolated learning)
 - Communication facilities (student/student, student/teacher)
 - Use of a variety of learning materials

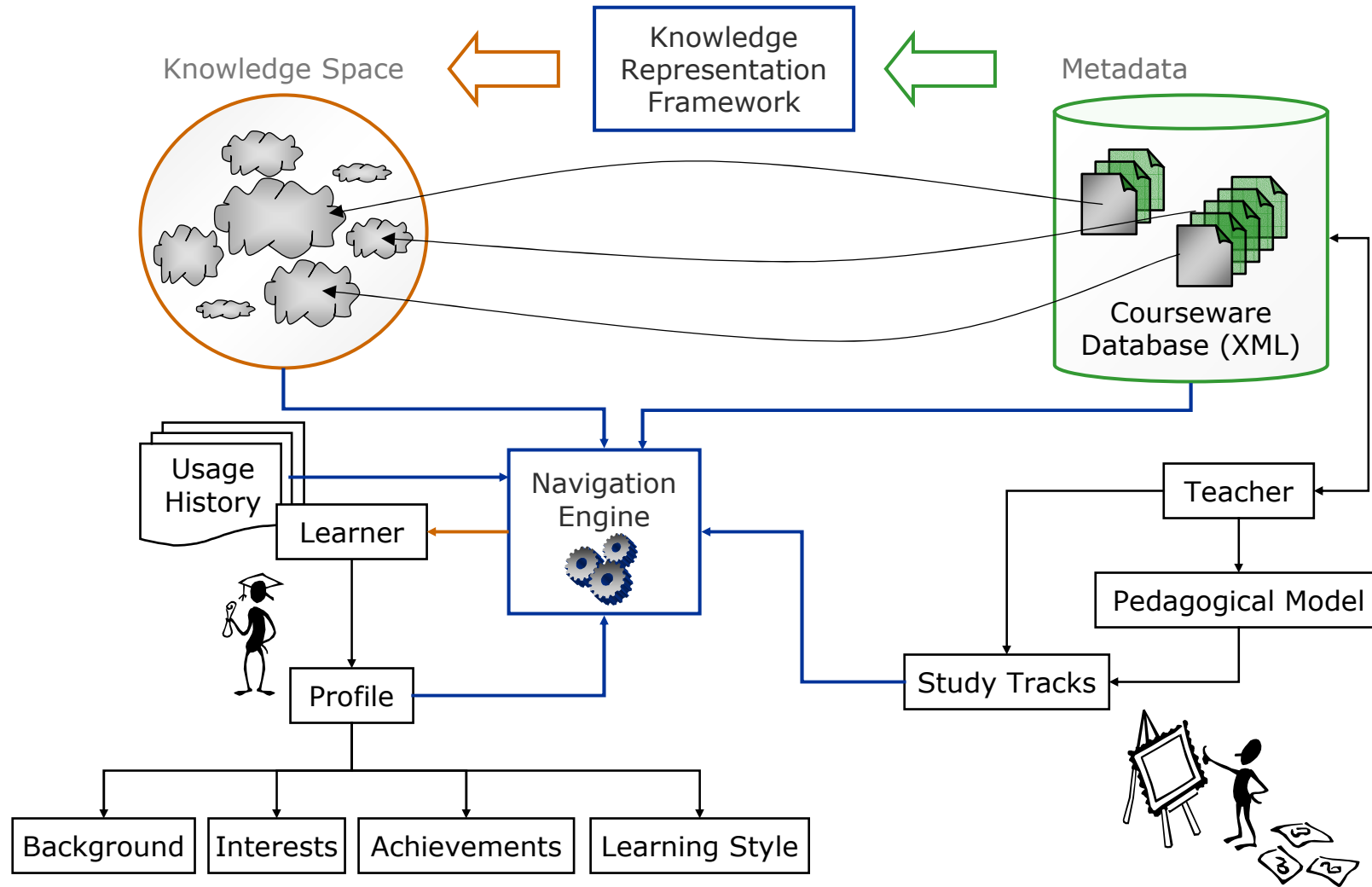


The CANDLE Project

- Collaborative And Networked Distributed Learning Environment
(<http://www.candle.eu.org/>)
- Funded by the European Union IST 5th framework programme over 3 years and 12 partners (academia and industry)
- Main objectives:
 - To use the Internet to improve the quality and reduce the cost of teaching
 - To enable co-operation between universities and industry in creating, sharing and reusing learning material
 - To improve the quality of courseware delivery
 - To increase the flexibility of the learning process
 - ⇒ Support of various *pedagogical frameworks*
 - ⇒ *User-adapted* learning sessions

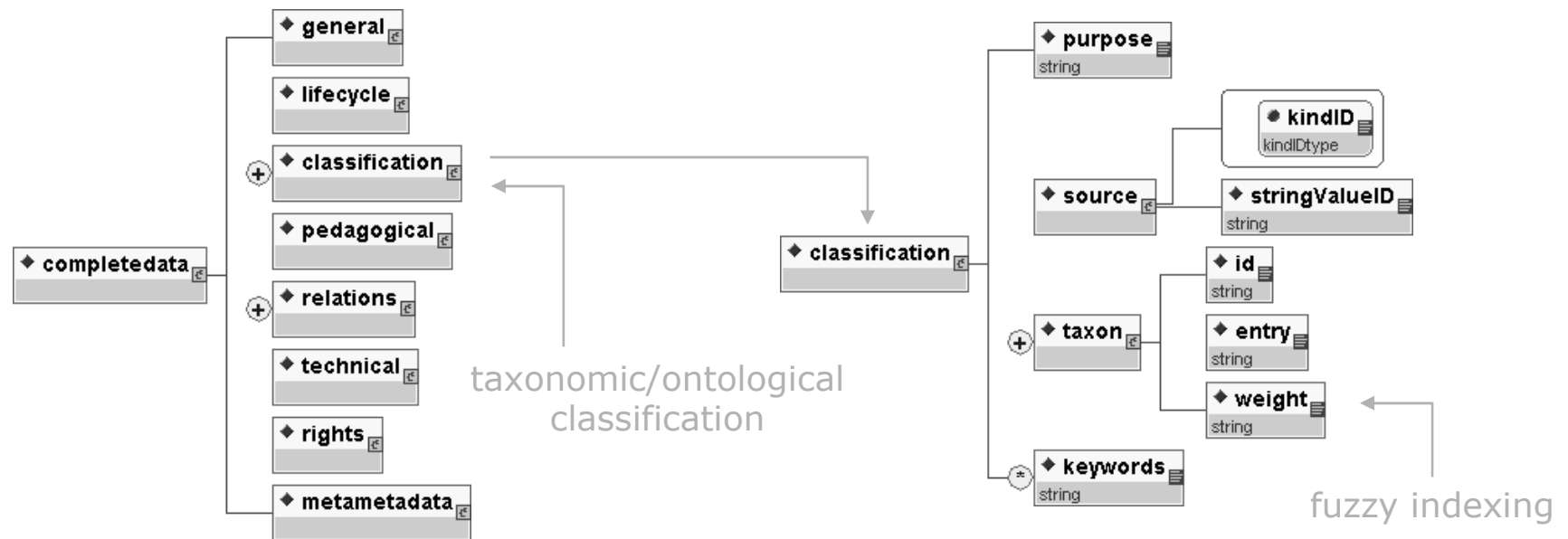


Adaptive Courseware Retrieval and Navigation



Educational Metadata

- IEEE's Learning Objects Metadata standard (LOM) – description of educational information resources
 - IMS project (consortium of US institutions of higher education)
 - ARIADNE project (funded by the European Union)
- CANDLE metadata schema: descriptive and organizational tasks





Knowledge Representation Framework

- Knowledge-based organization of learning material
 - Knowledge space definition
 - Identification of document relationships within the knowledge space
- Ontology of the domain
 - Definition of concepts + relations
 - Inference to find related material (authoring and learning)
 - Limitations:
 - ⇒ who defines and agrees on the “right” ontology?
 - ⇒ how to maintain the ontology as the knowledge fields develop?



Fuzzy Clustering for Knowledge Representation

- Dynamic discovery of knowledge relations between learning material
 - Document clustering
 - ⇒ Input: ontological classification (indexing terms+weights)
 - ⇒ Output: clusters of related documents
- *Why Fuzzy Clustering?*
 - Knowledge space can only be approximated
 - ⇒ "Soft" ontology
 - Author's ontological classification is subjective
 - ⇒ Imprecision



Vector Space Model of Information Retrieval

1. Document indexing: $x_i = [w_{i1} \ w_{i2} \ \dots \ w_{ik}]$

term frequency inverse document frequency (specificity)

2. Term weighting:

$$w_{ij} = \frac{\overbrace{f_{ij}} \cdot \overbrace{\log(N/n_j)}}{\sqrt{\sum_{t=1}^k (f_{it} \cdot \log(N/n_t))^2}} \quad \left. \vphantom{\frac{f_{ij} \cdot \log(N/n_j)}{\sqrt{\sum_{t=1}^k (f_{it} \cdot \log(N/n_t))^2}}} \right\} \text{document length normalization}$$

3. (Dis)similarity function:

$$S(x_\alpha, x_\beta) = \sum_{j=1}^k w_{\alpha j} \cdot w_{\beta j} = x_\alpha \cdot x_\beta^T \quad \Rightarrow 0 \leq S(x_\alpha, x_\beta) \leq 1, \forall \alpha, \beta$$
$$\Rightarrow S(x_\alpha, x_\alpha) = 1, \forall \alpha$$

$$D(x_\alpha, x_\beta) = 1 - S(x_\alpha, x_\beta) = 1 - \sum_{j=1}^k w_{\alpha j} \cdot w_{\beta j}$$



Fuzzy C-Means Clustering Algorithm

- Iterative optimization of an objective function:

$$J_m(U, V) = \sum_{i=1}^N \sum_{\alpha=1}^c u_{\alpha i}^m d_{i\alpha}^2 = \sum_{i=1}^N \sum_{\alpha=1}^c u_{\alpha i}^m \|x_i - v_{\alpha}\|^2$$

- Subject to:

$$u_{\alpha i} \in [0, 1], \quad \forall_{\alpha} \forall_i \quad \sum_{\alpha=1}^c u_{\alpha i} = 1, \quad \forall_i \quad 0 < \sum_{i=1}^N u_{\alpha i} < N, \quad \forall_{\alpha}$$

- Selection of c , m , distance function and convergence threshold ξ
- Fuzzy memberships and cluster centres:

$$u_{\alpha i} = \left[\sum_{\beta=1}^c \left(\frac{d_{i\alpha}^2}{d_{i\beta}^2} \right)^{\frac{1}{m-1}} \right]^{-1} = \left[\sum_{\beta=1}^c \left(\frac{\|x_i - v_{\alpha}\|}{\|x_i - v_{\beta}\|} \right)^{\frac{2}{m-1}} \right]^{-1} \quad v_{\alpha} = \frac{\sum_{i=1}^N u_{\alpha i}^m \cdot x_i}{\sum_{i=1}^N u_{\alpha i}^m}$$



Euclidean Distance vs. Dissimilarity Function

- Example:

	$f(t_1)$	$f(t_2)$	$f(t_3)$	$f(t_4)$	$f(t_5)$
Doc. A	20	0	2	10	0
Doc. B	0	12	0	0	7
Doc. C	47	0	15	6	0

- Euclidean distance:

$$d_{AB} = \sqrt{20^2 + 12^2 + 2^2 + 10^2 + 7^2} = \underline{26.40}$$

$$d_{AC} = \sqrt{27^2 + 13^2 + 4^2} = \underline{30.23}$$

- Dissimilarity:

$$D_{AB} = 1 - \frac{0}{\sqrt{20^2 + 2^2 + 10^2} \cdot \sqrt{12^2 + 7^2}} = \underline{1} \quad 0 \leq D(x_\alpha, x_\beta) \leq 1, \forall \alpha, \beta$$

$$D_{AC} = 1 - \frac{20 \cdot 47 + 2 \cdot 15 + 10 \cdot 6}{\sqrt{20^2 + 2^2 + 10^2} \cdot \sqrt{47^2 + 15^2 + 6^2}} = \underline{0.08}$$



Modified Fuzzy C-Means Algorithm

- Objective function:

$$J_m(U, V) = \sum_{i=1}^N \sum_{\alpha=1}^c u_{\alpha i}^m D_{i\alpha} = \sum_{i=1}^N \sum_{\alpha=1}^c u_{\alpha i}^m \left(1 - \sum_{j=1}^k x_{ij} \cdot v_{\alpha j}\right)$$

- Constraint (Lagrange multiplier):

$$D(v_{\alpha}, v_{\alpha}) = 1 - \sum_{j=1}^k v_{\alpha j} \cdot v_{\alpha j} = 1 - \sum_{j=1}^k v_{\alpha j}^2 = 0, \forall \alpha$$

- Fuzzy memberships and cluster centres:

$$u_{\alpha i} = \left[\sum_{\beta=1}^c \left(\frac{D_{i\alpha}}{D_{i\beta}} \right)^{\frac{1}{m-1}} \right]^{-1} \quad v_{\alpha} = \sum_{i=1}^N u_{\alpha i}^m x_i \cdot \sqrt{\frac{1}{\sum_{j=1}^k \left(\sum_{i=1}^N u_{\alpha i}^m x_{ij} \right)^2}}$$

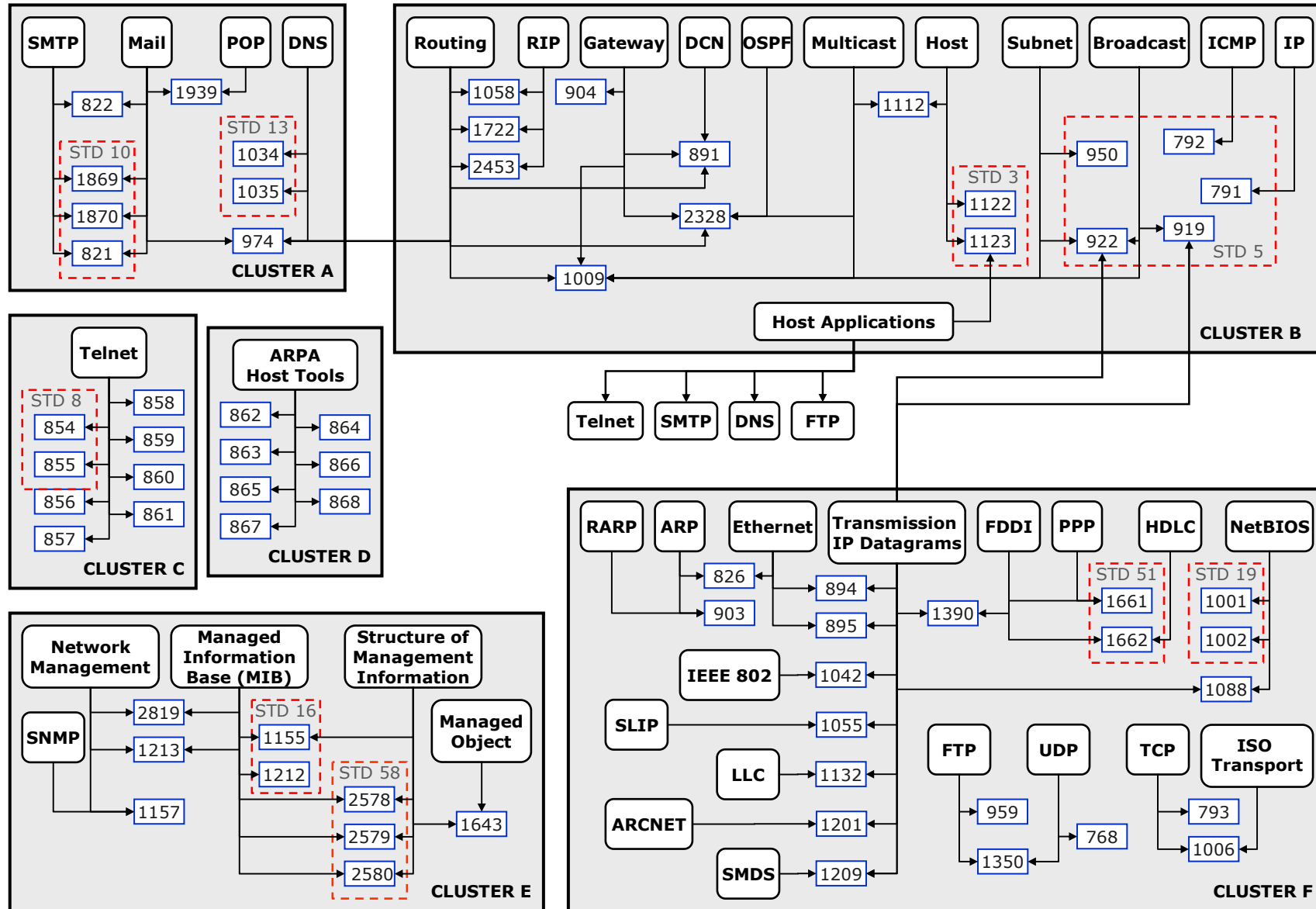


Preliminary Trials

- Analysis of the performance of the fuzzy c-means algorithm
- Comparative analysis of different metric spaces
 - Euclidean distance
 - Dissimilarity function
- Data Set
 - Collection of IETF's RFC standards ($N=67$ text documents)
 - Automatic indexing using terms from an available taxonomy ($k=465$ terms)



Reference Clustering



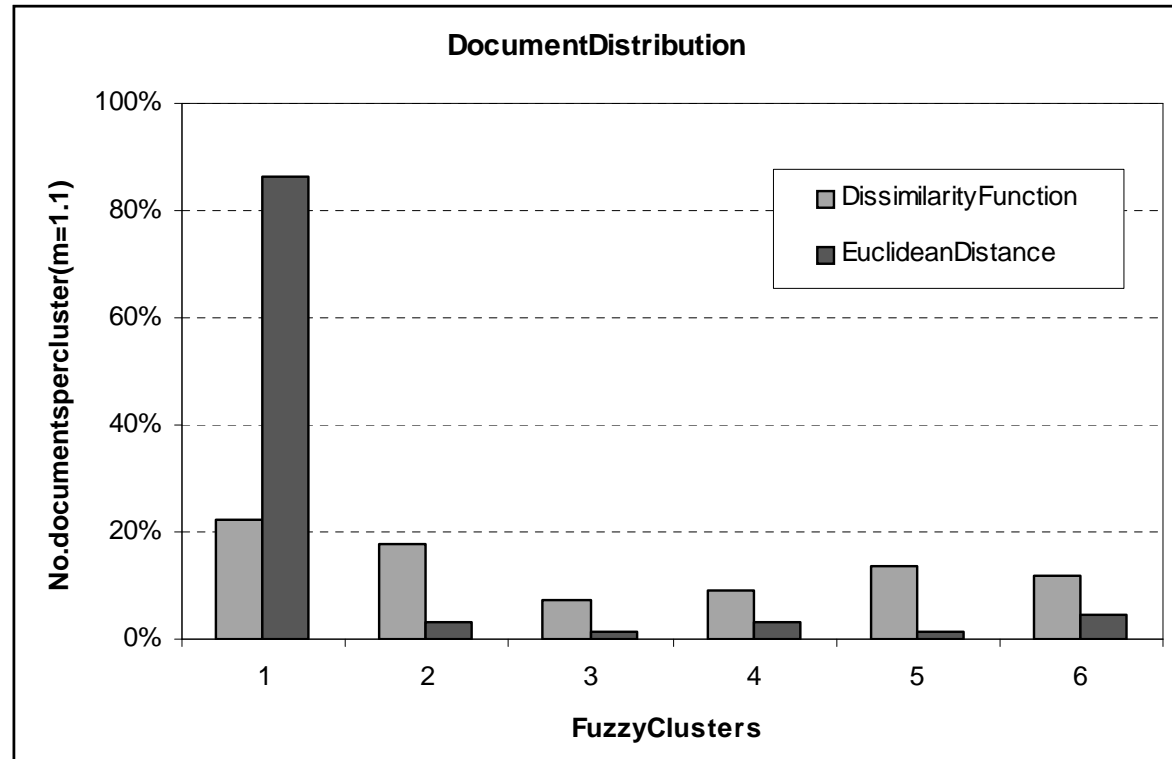


Performance Comparison ($c=6, m=1.1$)

Partition Entropy (PE):

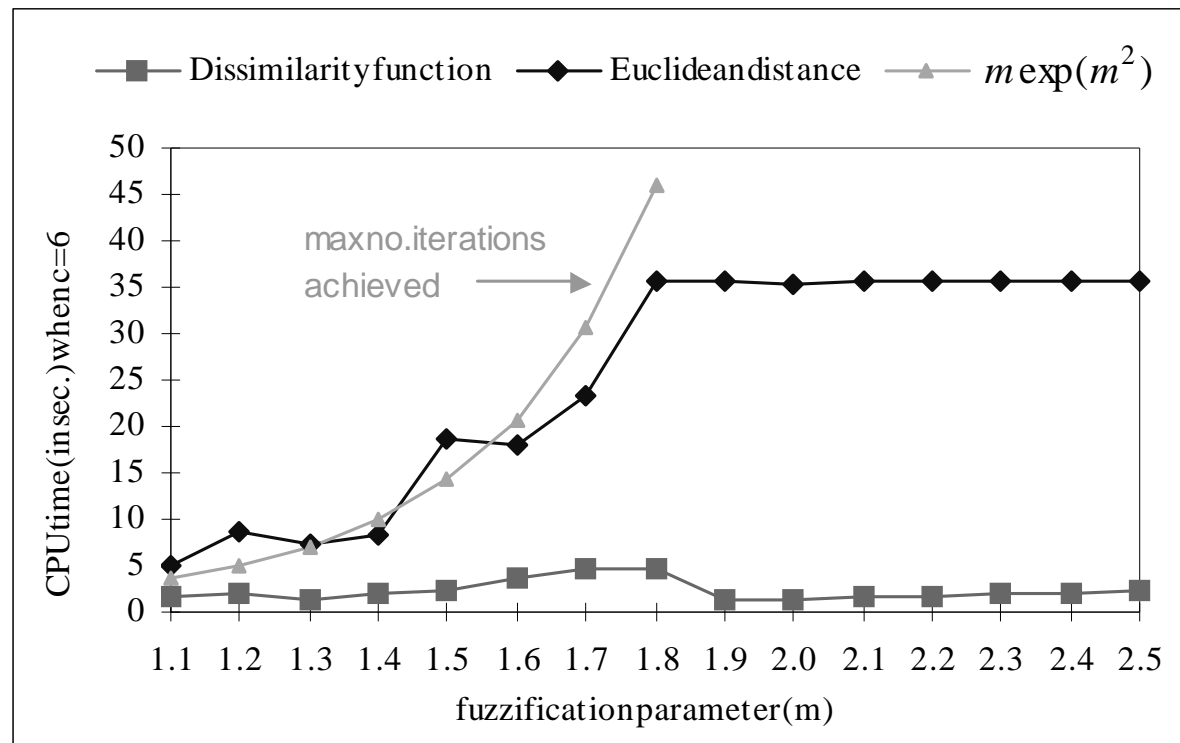
$$PE = -\frac{1}{N} \sum_{i=1}^N \sum_{\alpha=1}^c u_{\alpha i} \log_a(u_{\alpha i})$$

Euclidean distance	0.013
Dissimilarity Function	0.422



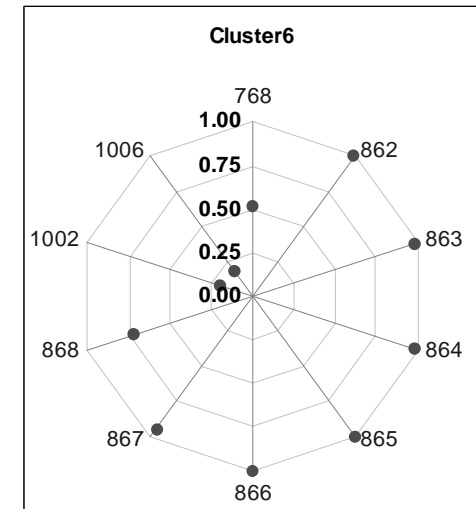
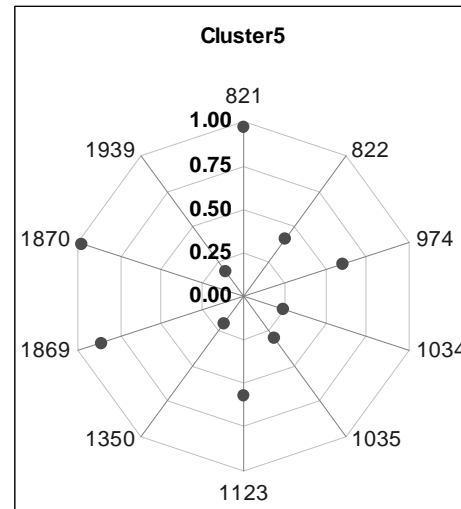
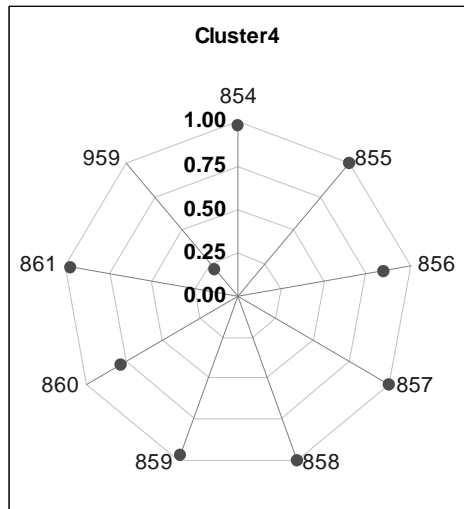
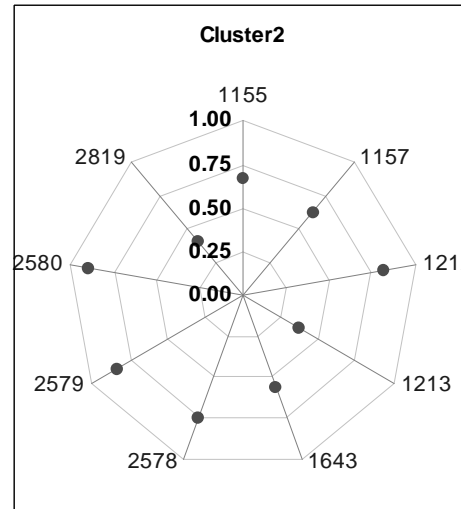
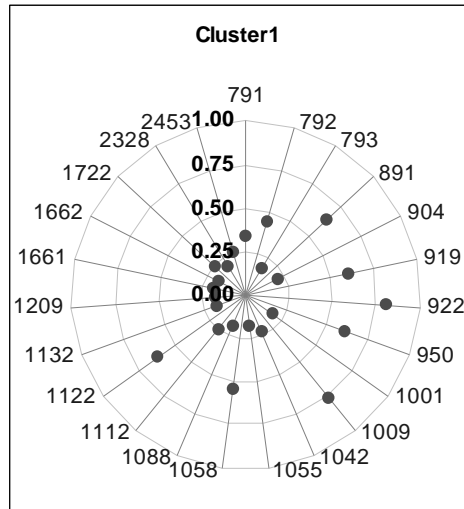
Performance Comparison ($c=6$)

- Execution time for increasing values of m :



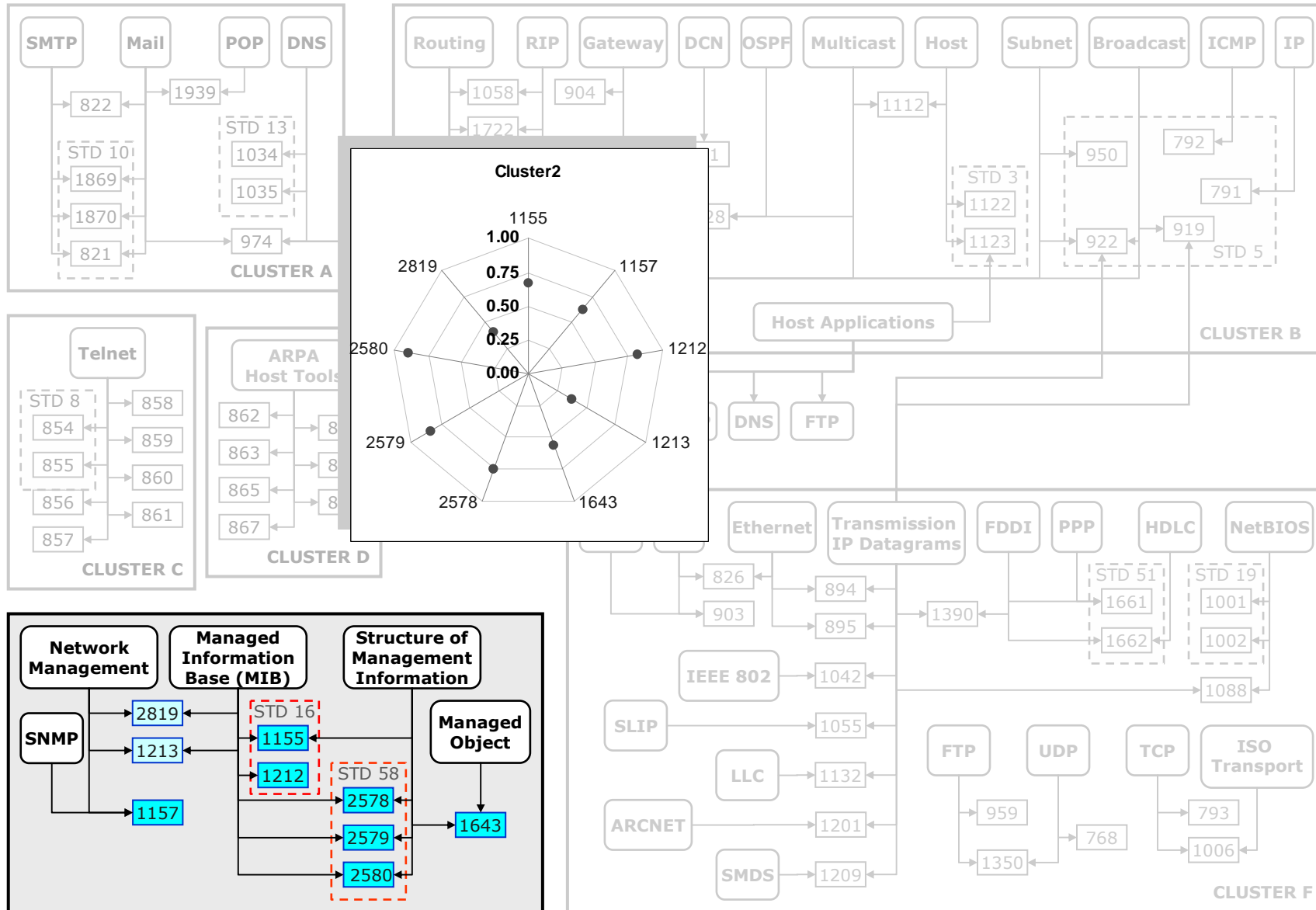


Clustering Results (dissimilarity function, $m=1.5$, $c=6$)



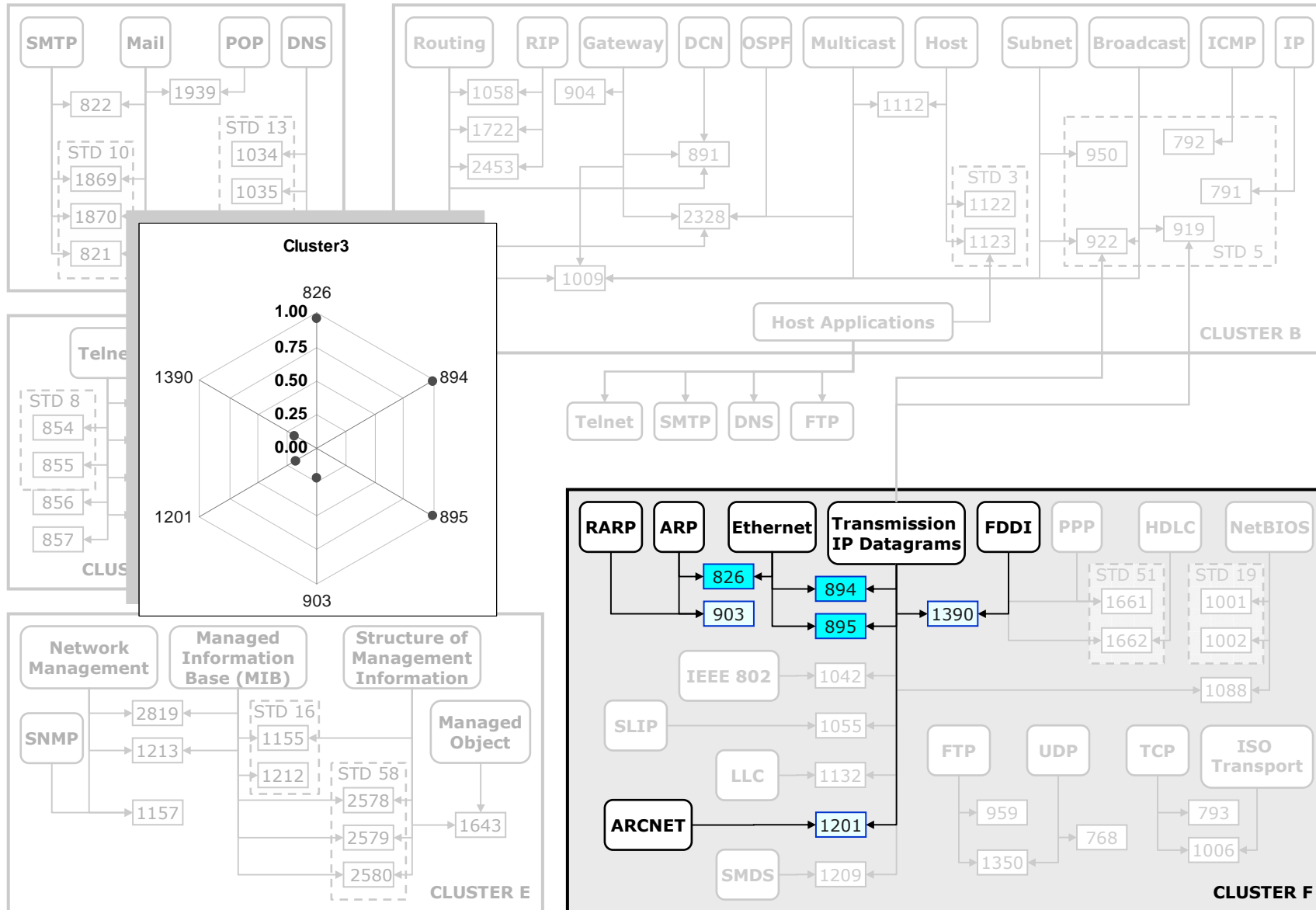


Fuzzy Cluster 2 – Management (SNMP, MIB, SMI)



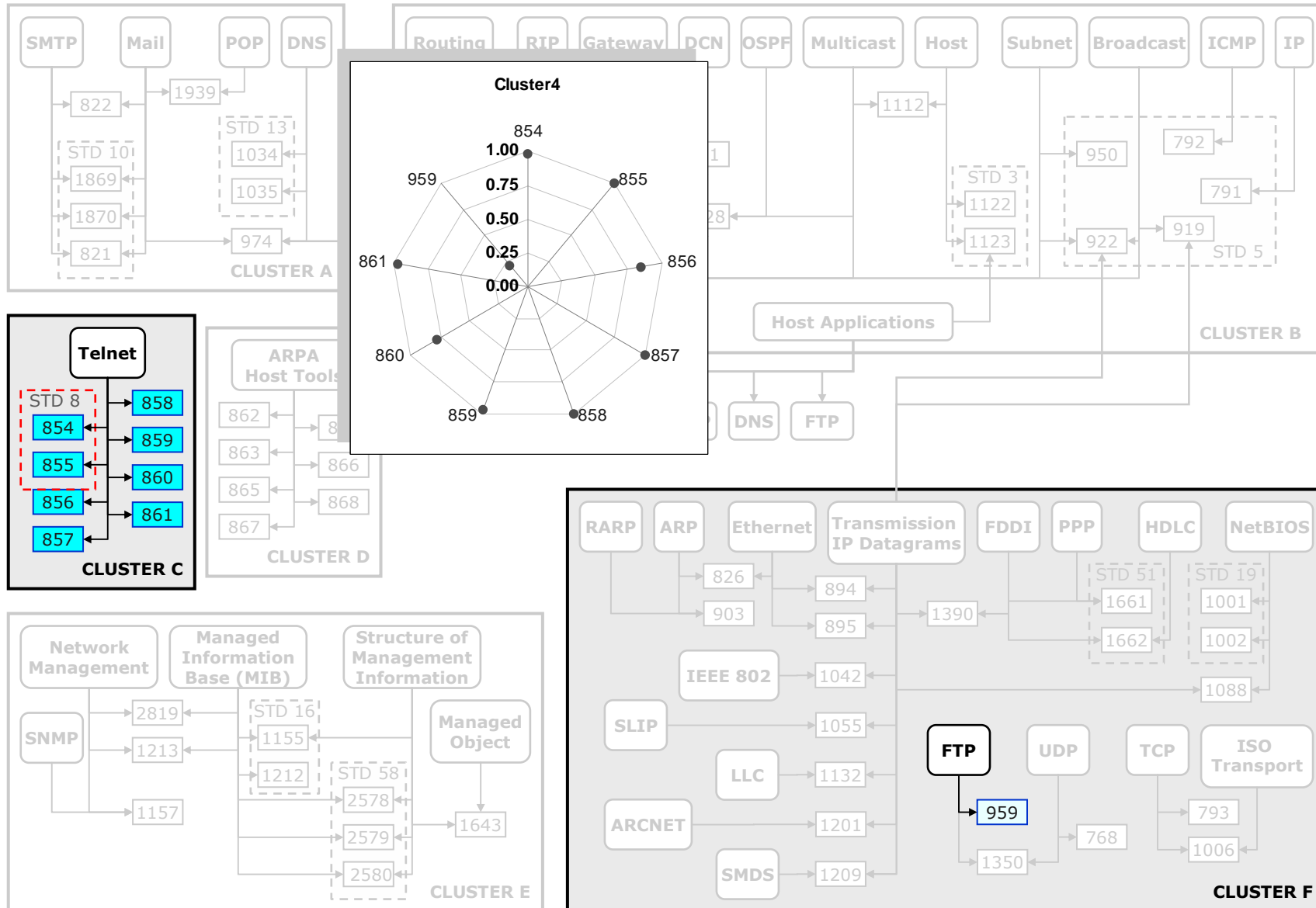


Fuzzy Cluster 3 – Ethernet & Addressing



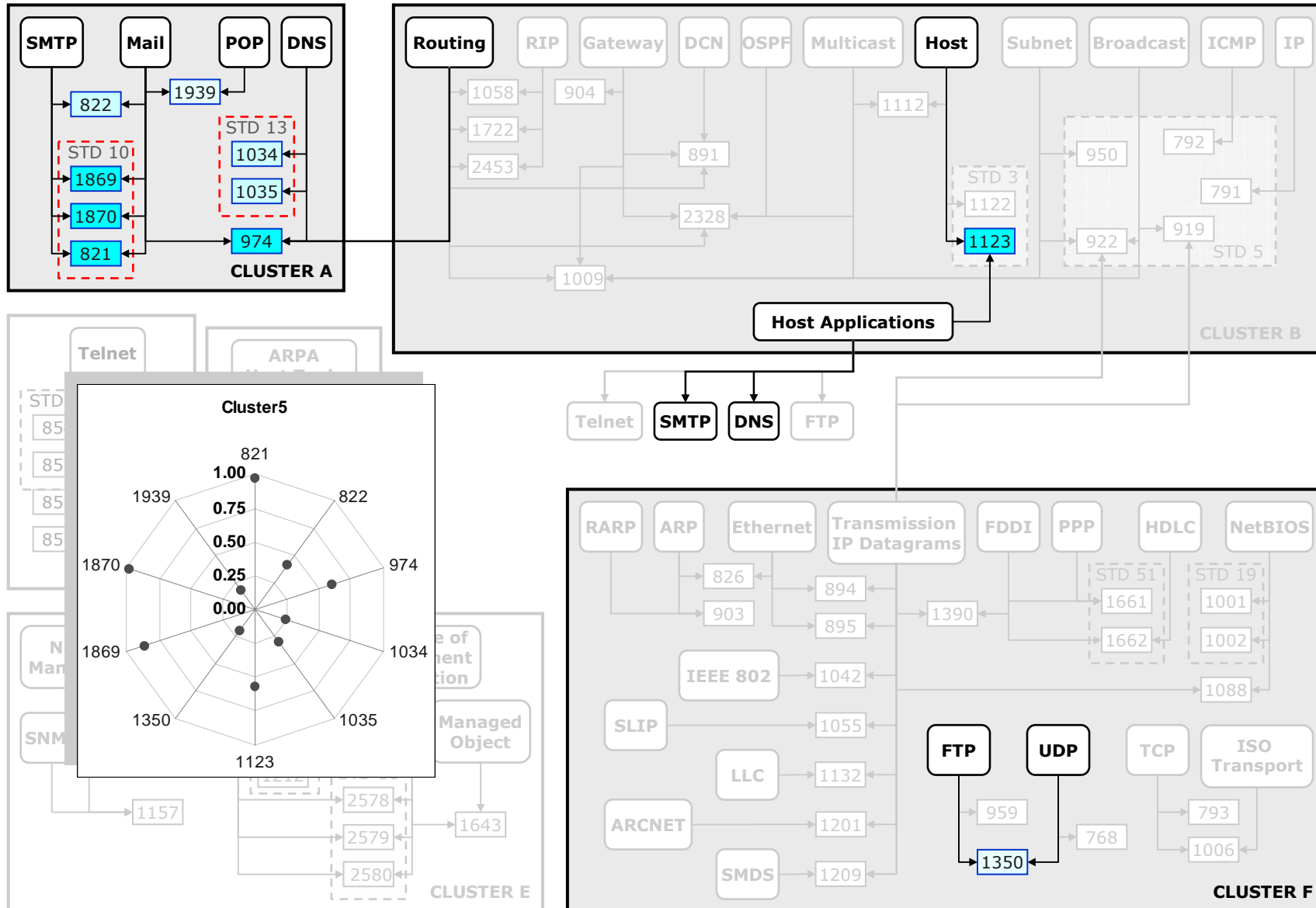


Fuzzy Cluster 4 – Telnet



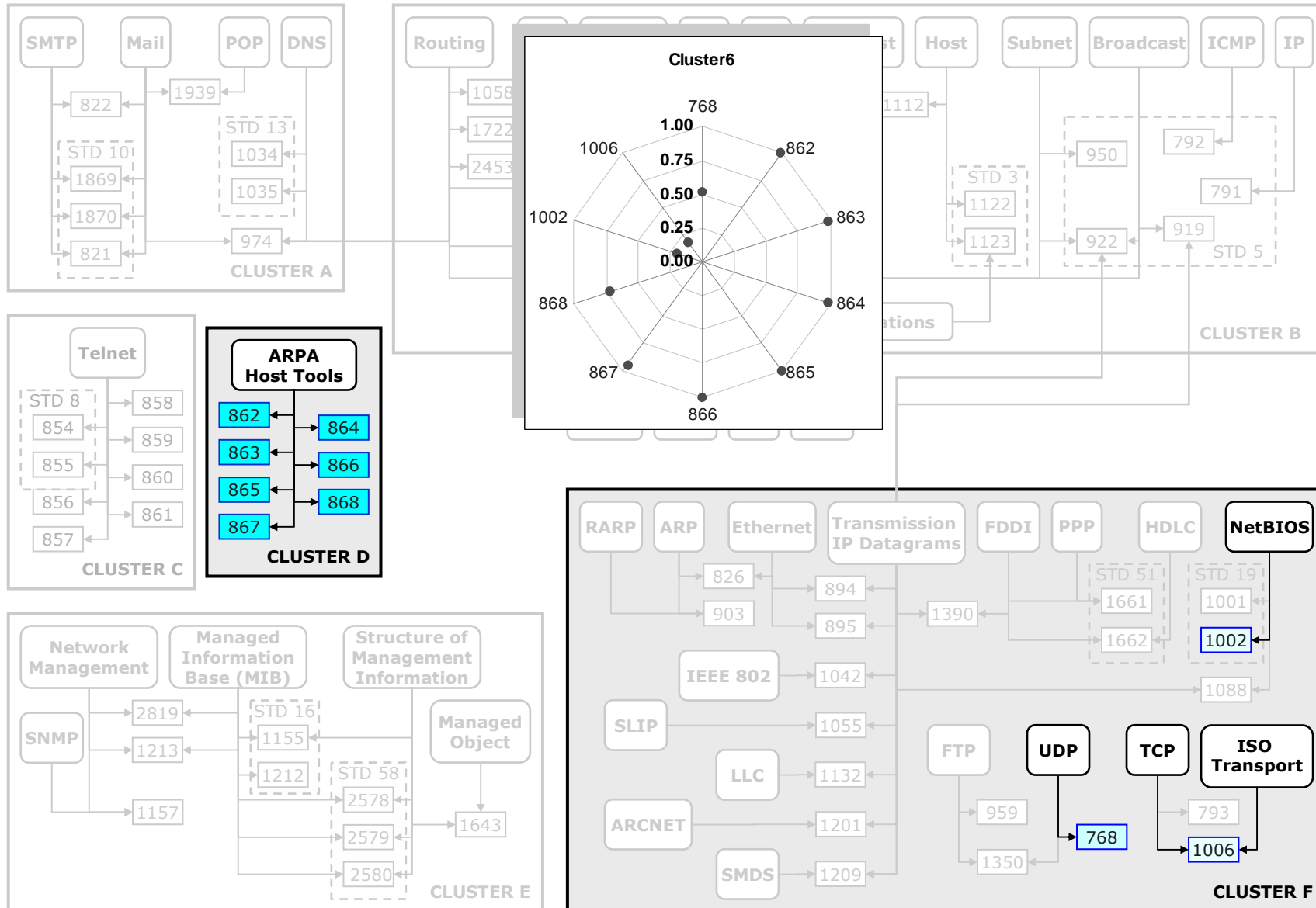


Fuzzy Cluster 5 – Mail & Domain Name System





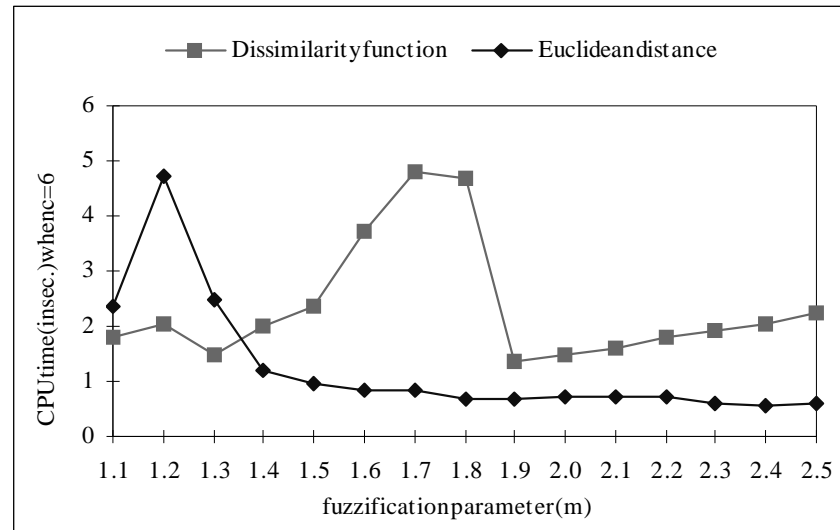
Fuzzy Cluster 6 – Measurement/Debugging Tools & Datagram



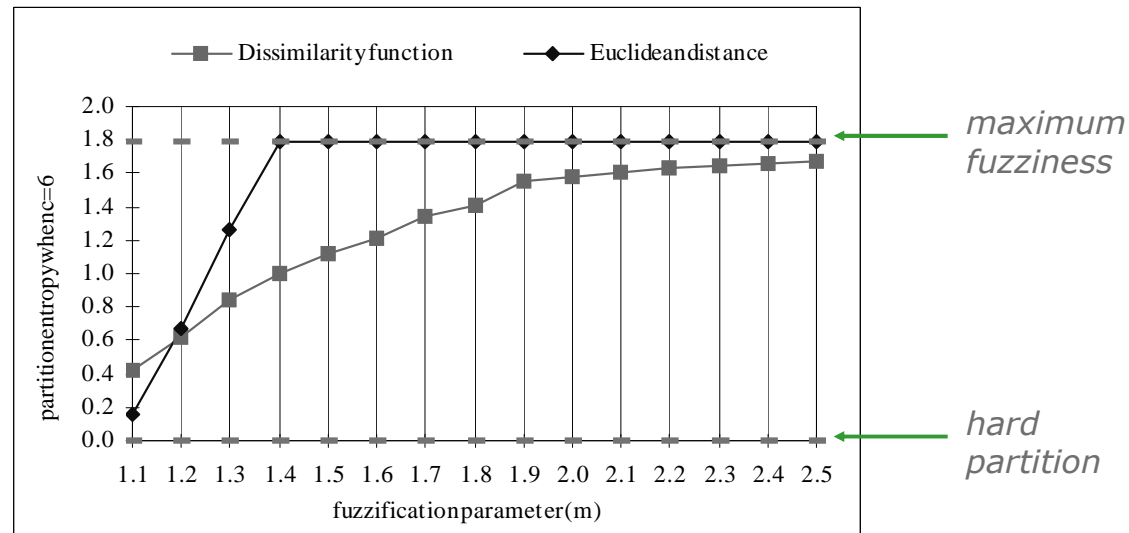


Performance Comparison ($c=6$ clusters)

Execution Time



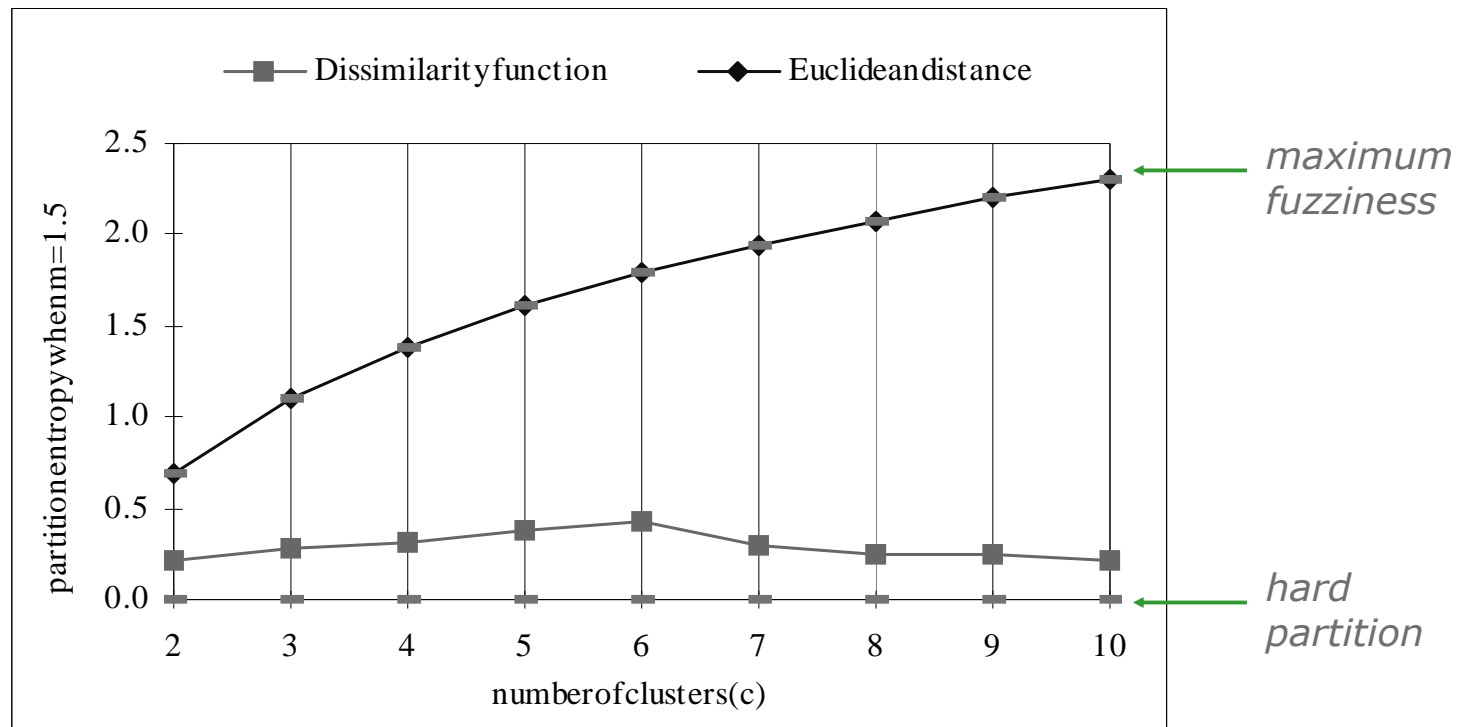
Partition Entropy





Performance Comparison ($m=1.5$)

- Partition Entropy for increasing number of clusters:





Concluding Remarks

- Fuzzy document relations obtained through clustering
- Clusters represent subsets of the knowledge space
- FCM with dissimilarity function behaves better than with the Euclidean distance
- More clusters provide higher refinement of the document relationships
- Further research issues:
 - Incremental update of the fuzzy clusters
 - Cluster refinement through an hierarchy of fuzzy partitions