A Hybrid Cluster-Lift Method for the Analysis of Research Activities

Boris Mirkin[1] Susana Nascimento[2] Trevor Fenner[1]
Luís Moniz Pereira [2]

[1]Department of Computer Science
Birkbeck University of London

[2]Department of Computer Science and Centre for Artificial Intelligence (CENTRIA)
FCT-Universidade Nova de Lisboa
Portugal

June 24th 2010 / HAIS'10 Conference

Outline

Outline

Outline

Outline

Outline

E-Survey Tool of Scientific Activities (ESSA)

- Selection of up to six topics among the leaf nodes of the ACM-CCS tree
- Assign each topic with a percentage expressing the proportion of the topic in the total of the respondent's research activity for the past four years.

Similarity between Research Topics

Table: Membership values for six ACM-CCS subjects A–F assigned by four individuals.

	i_1	i_2	i_3	i_4
A	0.6			0.2
B	0.4		0.2	0.2
C		0.2	0.4	0.2
D		0.3	0.4	0.2
E		0.5		0.2
F				
n_v	2	3	3	5

– Similarity

$$a(A,B) = \frac{2}{5} \times 0.6 \times 0.4 + \frac{3}{5} \times 0 \times 0 + \frac{3}{5} \times 0 \times 0.2 + \frac{5}{5} \times 0.2 \times 0.2 = 0.136$$

Outline

Additive Fuzzy Clustering Model

Definition

- Given between topics similarity matrix $A = (a_{tt'}), t, t' \in T$
- Assume K fuzzy thematic clusters $(\mathbf{u}_k, \mu_k)$
  - Membership vector $\mathbf{u}_k = (u_{kt})$, s.t. $0 \leq u_{kt} \leq 1$ for all $t \in T$
  - Intensity $\mu_k > 0$
  - unknown K
- Additive Fuzzy Clustering model

$$a_{tt'} = \sum_{k=1}^{K} \mu_k^2 u_{kt} u_{kt'} + e_{tt'}$$

the product $\mu_k^2 u_{kt} u_{kt'}$ expresses the contribution of cluster k to the similarity $a_{tt'}$ bt. topics t and t'.

Outline

Fitting Additive Fuzzy Clustering Model with Spectral Method (1)

- Least-squares fitting with the one-by-one principal component analysis strategy, for finding one cluster at a time.
- Each step minimizes criterion

$$E = \sum_{t,t' \in T} (w_{tt'} - \xi u_t u_{t'})^2 \quad (1)$$

wrt unknown $\xi > 0$ weight and fuzzy membership vector $\mathbf{u} = (u_t)$, given residual similarity value $w_{tt'}$.

Fitting Additive Fuzzy Clustering Model with Spectral Method (2)

- Find ξ by minimizing (1) for arbitrary $\mathbf{u}$

$$minE(\xi, \widehat{\mathbf{u}}) = \sum_{t,t' \in T} (w_{tt'} - \xi u_t u_{t'})^2 \qquad (2)$$

First order optimality condition

$$\xi = \frac{\sum_{t,t' \in T} w_{tt'} u_t u_{t'}}{\sum_{t \in T} u_t^2 \sum_{t' \in T} u_{t'}^2}$$

$$\xi = \frac{\mathbf{u}' W \mathbf{u}}{(\mathbf{u}'\mathbf{u})^2} \qquad (3)$$

which is non-negative if matrix W is semi-positive definite.

Fitting Additive Fuzzy Clustering Model with Spectral Method (3)

- Find $\mathbf{u}$ by minimizing (1) for the derived ξ

$$E(\widehat{\xi}, \mathbf{u}) = \sum_{t,t' \in T} w_{tt'}^2 - \xi^2 \sum_{t \in T} u_t^2 \sum_{t' \in T} u_{t'}^2 = S(W) - \xi^2 \left(\mathbf{u}'\mathbf{u}\right)^2,$$

where $S(W) = \sum_{t,t' \in T} w_{tt'}^2$ is the similarity data scatter.

$$G(u) = \xi^2 \left(\mathbf{u}'\mathbf{u}\right)^2 = \left(\frac{\mathbf{u}'W\mathbf{u}}{\mathbf{u}'\mathbf{u}}\right)^2 \qquad (4)$$

  - Rayleigh quotient squared.

Fitting Additive Fuzzy Clustering Model with Spectral Method (4)

- Similarity data scatter decomposition

$$S(W) = G(\mathbf{u}) + E \tag{5}$$

  - $G(\mathbf{u})$ which is explained by cluster $(\mu, \mathbf{u})$
  - E, which remains unexplained by the cluster.

- Minimizing $E \equiv$ maximizing $G(\mathbf{u})$ or Rayleigh quotient $g(\mathbf{u}) = \sqrt{G(\mathbf{u})}$

$$g(\mathbf{u}) = \xi \mathbf{u}'\mathbf{u} = \frac{\mathbf{u}'W\mathbf{u}}{\mathbf{u}'\mathbf{u}} \tag{6}$$

whose maximum value is the maximum eigenvalue of matrix W, which is reaseach at the corresponding eigenvector.

Fitting Additive Fuzzy Clustering Model with Spectral Method (5)

- Spectral Clustering approach

$$\Lambda(W) = [\lambda, z]$$

  - λ maximum eigenvalue of W
  - z corresponding normed eigenvector for W
- Projection $\mathcal{P}(z)$

$$\mathcal{P}(z) = \begin{cases} 0, & \text{if } z \leq 0; \\ z, & \text{if } 0 < z < 1; \\ 1, & \text{if } z \geq 1. \end{cases}$$

to the set of admissible fuzzy membership vectors.

ADDItive Fuzzy Spectral Clustering Algorithm (ADDI-FS)

---

1 *Set:* $k = 0$, $W = A$, $\epsilon > 0, \tau > 0$;
  $S_0 = Tr(W'W)$;
2 **Repeat**
3 $k = k + 1$;
4 $[\lambda, z] = \Lambda(W)$;
5 **If** $\lambda > 0$
6 $\mathbf{u}_k = \mathcal{P}(z)$ or $\mathbf{u}_k = \mathcal{P}(-z)$ depending on which leads to a larger G_k
7 $\xi_k = \frac{\mathbf{u}_k' W \mathbf{u}_k}{(\mathbf{u}_k' \mathbf{u}_k)^2}$;
8 $G_k = \xi_k^2 (\mathbf{u}_k' \mathbf{u}_k)^2$;
9 $W = W - \xi_k \mathbf{u}_k \mathbf{u}_k'$;
10 $S_k = S_{k-1} - G_k / S_0$;
11 **Else**: Halt
12 **Until** ($\xi_k \leq 0$ or $G_k / S_0 \leq \tau$ or $S_k \leq \epsilon$ or $k == K_{\max}$)

Stop Criteria of Squential Extraction of Fuzzy Clusters

1. The optimal value of ξ (3) for the spectral fuzzy cluster is negative.
2. The relative contribution, G_k/S_0, of a single extracted cluster becomes too low, less than a pre-specified $\tau > 0$ value.
3. The residual data scatter becomes smaller than a pre-specified $\epsilon > 0$ value.
4. A pre-specified number $K_{\max}$ of clusters is reached.

Outline

Parsimonious Lifting Method (1)

- To generalize the contents of a thematic cluster, we lift it to higher ranks of the taxonomy so that if all or almost all children of a node in an upper layer belong to the cluster, then the node itself is taken to represent the cluster at this higher level of the ACM-CCS taxonomy.
- Three types of events:
  - “head subject”, which is a taxonomy node covering (some of) leaves belonging to the cluster, so that the cluster is represented by a set of head subjects.
  - “gaps” are head subject’s children topics that are not included in the cluster.
  - “offshoot” is a taxonomy leaf node that is a head subject (not lifted)

Parsimonious Lifting Method (2)

- The total count of “head subjects”, “gaps” and “offshoots”, each weighted by both the penalties and leaf memberships, is used for scoring the extent of the cluster misfit needed for lifting a grouping of research topics over the classification tree.
- The smaller the score, the more parsimonious the lift and the better the fit.
- Depending on the relative weighting of gaps, offshoots and multiple head subjects, different lifts can minimize the total misfit.

Outline

Analysis of CENTRIA ESSA Survey'09

- CENTRIA ESSA Survey Data

	N. of Contacted Respondents	N. of Participating Respondents	N. of 3rd Layer ACM-CCS Topics Covered
CENTRIA-UNL	23	16	46/318

- ADDI-FS Clusters found after LAPIN

Cluster	Contribution(%)	λ_1	Weight(ξ)	Intensity(μ)
C_1	35.2	46.5	31.04	5.57
C_2	15.2	32.9	20.41	4.52

- ADDI-FS Stop condition: weight, ξ of third cluster becomes negative.

ADDI-FS Cluster 2 at CENTRIA

Cluster 2		**Defuzzification Threshold= 0.1**
Contribution	15.2%	
Eigenvalue	32.90	
Intensity	4.52	
Weight	20.41	
Membership	Code	Topic
0.46756	J.3	LIFE AND MEDICAL SCIENCES (Applications in)
0.40619	I.2.8	Problem Solving, Control Methods, and Search
0.34435	F.2.1	Numerical Algorithms and Problems
0.32681	F.4.1	Mathematical Logic
0.30067	G.1.6	Optimization
0.25967	D.3.3	Language Constructs and Features
0.23748	G.2.2	Graph Theory
0.18722	G.3	PROBABILITY AND STATISTICS
0.17359	B.2.3	Reliability, Testing, and Fault-Tolerance
0.17359	B.7.3	Reliability and Testing
0.17203	I.2.0	General in I.2 ARTIFICIAL INTELLIGENCE
0.1537	G.1.0	General in G.1 NUMERICAL ANALYSIS
0.11827	I.2.3	Deduction and Theorem Proving
0.10195	G.1.7	Ordinary Differential Equations

Parsimonious Lift Mapping of Cluster 2 over the ACM-CCS

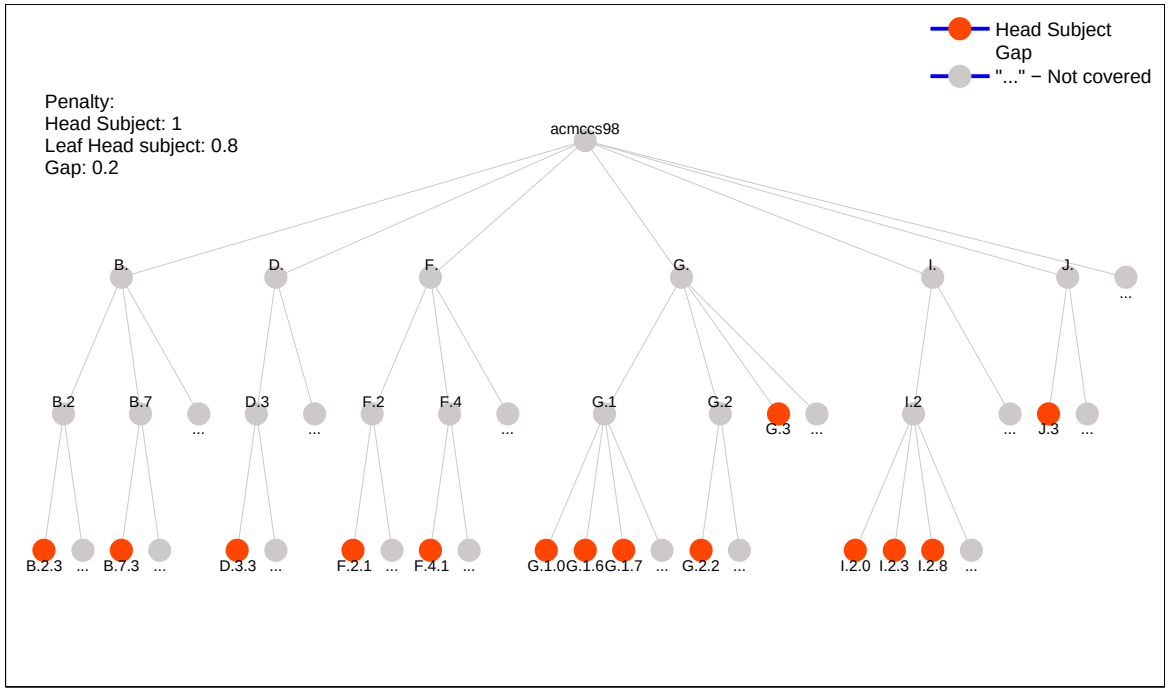
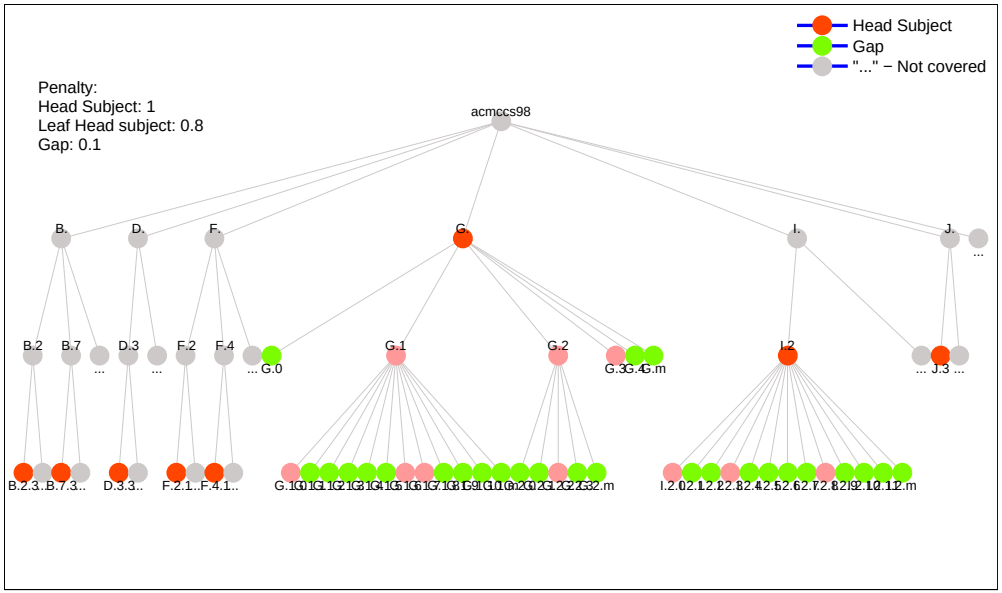


Figure: Mapping of CENTRIA cluster 2 onto the ACM-CCS tree with penalties of the Lifting $h = 1$, $o = 0.8$ and $g = 0.2$.

Parsimonious Representation of CENTRIA Cluster 2

	HEAD SUBJECTS
B.2.3	Reliability, Testing, and Fault-Tolerance
B.7.3	Reliability and Testing
D.3.3	Language Constructs and Features
F.2.1	Numerical Algorithms and Problems
F.4.1	Mathematical Logic
G.1.0	General in G.1 - NUMERICAL ANALYSIS
G.1.6	Optimization
G.1.7	Ordinary Differential Equations
G.2.2	Graph Theory
G.3	PROBABILITY AND STATISTICS
I.2.0	General in I.2 - ARTIFICIAL INTELLIGENCE
I.2.3	Deduction and Theorem Proving
I.2.8	Problem Solving, Control Methods, and Search
J.3	LIFE AND MEDICAL SCIENCES

Parsimonious Lift Mapping of Cluster 2 over the ACM-CCS: higher ranks



Mapping of CENTRIA Cluster 2 on ACM−CCS taxonomy

Figure: By decreasing the gap penalty to $g = 0.1$ leads to a different parsimonious generalization, where the various G's leaf nodes are lifted to higher ranks of the taxonomy and generalized to node *G*.

Parsimonious Representation of CENTRIA Cluster 2: more general

with gap= 0.1

	HEAD SUBJECTS
B.2.3	Reliability, Testing, and Fault-Tolerance
B.7.3	Reliability and Testing
D.3.3	Language Constructs and Features
F.2.1	Numerical Algorithms and Problems
F.4.1	Mathematical Logic
G.	Mathematics of Computing
I.2	ARTIFICIAL INTELLIGENCE
J.3	LIFE AND MEDICAL SCIENCES

Conclusion (1)

- A hybrid method for representing aggregated research activities over a taxonomy. The method constructs fuzzy profiles of the entities constituting the structure under consideration and then generalizes them in two steps:
  - fuzzy clustering research topics according to their thematic similarities, ignoring the topology of the taxonomy; and
  - lifting clusters mapped to the taxonomy to higher ranked categories in the tree.

Conclusion (2)

- These generalization steps cover both sides of the representation process: the empirical – related to the structure under consideration – and the conceptual – related to the taxonomy hierarchy.
- This work is part of the research project *Computational Ontology Profiling of Scientific Research Organization* (COPSRO), main goal of which is to develop a method for representing a Computer Science organization, such as a university department, over the ACM-CCS classification tree.
- In principle, the approach can be extended to other areas of science or engineering, provided that such an area has been systemised in the form of a comprehensive concept tree.